



Kingdom of Saudi Arabia  
Ministry of Higher Education  
Imam Abdulrahman Bin Faisal University  
College of Computer Sciences & Information Technology

## Preemptive Diagnosis of Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia Using Computational Intelligence Techniques

*A project submitted  
in partial fulfillment of the requirements for the degree of  
Bachelor of Science in Artificial Intelligence*

By

Name	ID	Role

Supervised by  
**Dr. Sunday Olusanya Olatunji (Aadam)**  
**Mr. Mohammad Aftab Alam Khan**

Committee Member Names  
**Dr. Mohammed Imran Basheer Ahmed**  
**Ms. Mehwash Farooqui**

10<sup>th</sup> May 2024

## **DECLARATION**

We hereby assert that this project report, titled " Preemptive Diagnosis of Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia Using Computational Intelligence Techniques" is a product of our original research endeavors, with due recognition given to all cited sources and quotations. It is important to note that this report has not been previously presented for consideration for any other academic qualification or distinction, neither at our current institution nor any other university or educational establishment. This project report stands as a significant contribution towards fulfilling the academic requirements for the degree of "Bachelor of Science in Artificial Intelligence" at the Computer Engineering Department, College of Computer Sciences and Information Technology, Imam Abdulrahman Bin Faisal University.

### **Project Team:**

<b>Group Number: 9FA1</b>		
<b>Name</b>	<b>Signature</b>	<b>Date</b>

### **Project Supervisors:**

<b>Name</b>	<b>Signature</b>	<b>Date</b>
<b>Mr. Mohammad Aftab A. Khan</b>		<b>04<sup>th</sup> December 2023</b>
<b>Dr. S.O Olatunji (Aadam)</b>		<b>10<sup>th</sup> October 2023</b>

## **ACKNOWLEDGEMENT**

In the name of the Almighty, the Most Merciful and Compassionate, we begin by expressing our profound gratitude to Allah, the Bestower of blessings, for illuminating our path in this endeavor. We implore Him to continue to bless our work, making it a source of lasting benefit for the community. Our deepest appreciation goes to Dr. Sunday Olatunji, whose unwavering support and astute guidance have been the bedrock of this project, providing us with an invaluable opportunity to expand our knowledge in this field. Additionally, we extend our heartfelt thanks to Mr. Aftab Khan for his invaluable contributions and unwavering commitment to this endeavor.

## ABSTRACT

*The ramifications of chronic illnesses go beyond affecting individuals and impacting societies and economies. Given that chronic diseases are a leading global cause of death and are highly prevalent in Saudi Arabia, it becomes imperative to leverage all available technologies for early detection. In the healthcare industry, Machine Learning (ML) is an emerging method that can effectively diagnose various diseases, including chronic conditions. Previous efforts at diagnosing chronic diseases have mainly concentrated on patients who are already displaying symptoms. Additionally, recent research has demonstrated a reliance on deep learning techniques with Computed Tomography (CT) scan images for disease detection, which presents limitations when applied in community hospitals with limited imaging resources. As a result, the goal of the project is to use ML approaches to detect Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia before symptoms appear using clinical data only. The clinical data used in this project is obtained from online data since the data is typically available. Furthermore, the chosen approach for this study draws inspiration from analogous research endeavors conducted previously. Our method integrates a variety of advanced machine learning approaches, drawing on insights from earlier studies. Promising results were found during the analysis of the identified disorders. Random Forest had an accuracy rate of 91.11% when used to detect and predict the presence of Osteoporosis and Osteopenia using an online dataset. Likewise, using an online dataset for the early detection of epileptic seizures, Logistic Regression achieved an accuracy of 87.67%. Additionally, to diagnose Sickle Cell Anemia, Random Forest model was trained using an online data, and it attained an accuracy of 85.95%. Through meticulous implementation and iterative refinement of these sophisticated techniques, our aim is to significantly enhance the accuracy and effectiveness of chronic disease detection. By harnessing the strengths of these models, we strive to fortify the predictive capabilities necessary for robust and early identification of the chosen chronic diseases, thereby facilitating proactive and targeted intervention strategies.*

## Table of Contents

<b><i>Chapter 1: Introduction</i></b> .....	<b>20</b>
<b>1.1 Introduction</b> .....	20
<b>1.2 Problem Statement</b> .....	22
<b>1.3 Motivation</b> .....	23
<b>1.4 Justification</b> .....	23
<b>1.5 Aims &amp; Objectives</b> .....	23
<b>1.6 Scope / Limitation of the Study</b> .....	24
<b>1.7 Social, Professional, Legal, and Ethical Implication</b> .....	24
<b>1.8 Project Organization</b> .....	25
<b><i>Chapter 2: Background and Review of Literature</i></b> .....	<b>27</b>
<b>2.1 Epileptic Seizure</b> .....	27
<b>2.1.1 Introduction</b> .....	27
<b>2.1.2 Literature Review</b> .....	27
<b>2.2 Osteoporosis</b> .....	31
<b>2.2.1 Introduction</b> .....	31
<b>2.2.2 Literature Review</b> .....	31
<b>2.3 Sickle Cell Anemia</b> .....	36
<b>2.3.1 Introduction</b> .....	36
<b>2.3.2 Literature Review</b> .....	36
<b><i>Chapter 3: Software Project Management Plans (SPMP)</i></b> .....	<b>52</b>
<b>3.1 Project Overview</b> .....	52
<b>3.1.1 Purpose, Scope, and Objectives</b> .....	52
<b>3.1.2 Assumptions, Constraints, and Risks</b> .....	53
<b>3.1.3 Project Deliverables</b> .....	54
<b>3.1.4 Schedule and Budget Summary</b> .....	54
<b>3.1.5 Evolution of the Plan</b> .....	55
<b>3.1.6 Definitions</b> .....	55
<b>3.1.7 Document Structure</b> .....	55
<b>3.2 Project Organization</b> .....	56
<b>3.2.1 External Interfaces</b> .....	56
<b>3.2.2 Internal Structure</b> .....	56
<b>3.2.3 Roles and Responsibilities</b> .....	57

<b>3.3 Managerial Process Plans .....</b>	58
<b>3.3.1 Start-up Plan .....</b>	58
<b>3.3.2 Work Plan .....</b>	59
<b>3.3.3 Project Tracking Plan .....</b>	66
<b>3.3.4 Risk Management Plan.....</b>	67
<b>3.3.5 Project Closeout Plan.....</b>	68
<b>3.4 Technical Process Plans .....</b>	68
<b>3.4.1 Process Model .....</b>	68
<b>3.4.2 Methods, Tools, and Techniques .....</b>	69
<b>3.4.3 Infrastructure .....</b>	69
<b>3.4.4 Product Acceptance.....</b>	70
<b>3.5 Supporting Process Plans .....</b>	70
<b>3.6 Resource Scheduling .....</b>	71
<b><i>Chapter 4: Methodology / Software Requirements Specification (SRS) .....</i></b>	<b>73</b>
<b>4.1 Methodology .....</b>	73
<b>4.1.1 Operational Framework.....</b>	73
<b>4.1.2 The Proposed Individual Computational Intelligence Approaches.....</b>	76
<b>4.1.3 Performance Evaluation of the Proposed Models.....</b>	83
<b>4.1.4 Summery .....</b>	83
<b>4.2 Software Requirements Specification (SRS).....</b>	84
<b>4.2.1 Introduction .....</b>	84
<b>4.2.2 Overall description.....</b>	84
<b>4.2.3 Specific Requirements .....</b>	87
<b>4.2.4 Performance Requirements.....</b>	128
<b>4.2.5 Design Constraints .....</b>	129
<b>4.2.6 Software System Attributes .....</b>	129
<b><i>Chapter 5: Software Design Specification (SDS) .....</i></b>	<b>131</b>
<b>5.1 System Overview .....</b>	131
<b>5.1.1 System Functionality.....</b>	131
<b>5.2 Design consideration .....</b>	132
<b>5.2.1 Assumptions and Dependencies .....</b>	132
<b>5.2.2 General Constraints.....</b>	133
<b>5.3 User Interface Design.....</b>	134
<b>5.3.1 Overview of User Interface .....</b>	134
<b>5.3.2 Interface Design Rules .....</b>	135

<b>5.3.3 Screen Images .....</b>	136
<b>5.3.4 Screen Objects and Actions .....</b>	159
<b>5.3.5 Other Interface .....</b>	162
<b>5.4 System Architecture .....</b>	168
<b>5.4.1 Architectural Design .....</b>	169
<b>5.4.2 Subsystem Architecture .....</b>	169
<b>5.5 Data Design .....</b>	172
<b>5.5.1 Data Description.....</b>	172
<b>5.5.2 Data Dictionary .....</b>	173
<b>5.5.3 Database Description .....</b>	174
<b>5.6 Component Design .....</b>	175
<b>5.6.1 Login Function .....</b>	176
<b>5.6.2 Account Functions.....</b>	176
<b>5.6.3 Admin Functions .....</b>	177
<b>5.6.4 Diagnostic Functions.....</b>	179
<b>5.7 Detailed System Design.....</b>	179
<b>5.7.1 Classification, Definition, and Responsibilities .....</b>	180
<b>5.7.2 Constraints and Composition .....</b>	181
<b>5.7.3 Uses/Interactions .....</b>	182
<b>5.7.4 Processing.....</b>	187
<b>5.7.5 Interface/Exports.....</b>	200
<b>5.7.6 Detailed Subsystem Design.....</b>	200
<b>5.8 Other Design Features .....</b>	201
<b>5.9 Detailed System Design.....</b>	201
<b>Chapter 6: Implementation Process .....</b>	<b>203</b>
<b>6.1 Changes from the proposal phase and justifications .....</b>	203
<b>6.2 Empirical Study on Osteoporosis Dataset.....</b>	203
<b>6.2.1 Data Description.....</b>	203
<b>6.2.2 Statistical Analysis of the Dataset .....</b>	204
<b>6.2.3 Experimental Setup.....</b>	206
<b>6.2.4 Description of proposed Techniques.....</b>	207
<b>6.2.5 Optimization strategy .....</b>	208
<b>6.2.6 Results and discussion.....</b>	209
<b>6.3 Empirical Study on Epileptic Seizures Disease Dataset .....</b>	213
<b>6.3.1 Data Description.....</b>	213

<b>6.3.2 Statistical Analysis of the Dataset .....</b>	215
<b>6.3.3 Experimental Setup.....</b>	218
<b>6.3.4 Description of the Proposed Techniques .....</b>	219
<b>6.3.5 Optimization strategy .....</b>	220
<b>6.3.6 Results and discussion.....</b>	221
<b>6.4 Empirical Study on Sickle Cell Anemia Disease Dataset.....</b>	227
<b>    6.4.1 Data Description.....</b>	227
<b>    6.4.2 Statistical Analysis of the Dataset .....</b>	228
<b>    6.4.3 Experimental Setup.....</b>	228
<b>    6.4.4 Description of the Proposed Techniques .....</b>	230
<b>    6.4.5 Optimization strategy .....</b>	231
<b>    6.4.6 Results and Discussion .....</b>	233
<b>6.5 Website Implementation .....</b>	240
<b>    6.5.1 Login.....</b>	240
<b>    6.5.2 Forgot Password Page.....</b>	241
<b>    6.5.3 Registration Page .....</b>	242
<b>    6.5.4 Email Confirmation Token.....</b>	242
<b>    6.5.5 Profile Page.....</b>	245
<b>    6.5.6 Diagnose Interface.....</b>	248
<b>    6.5.7 Diagnose Result .....</b>	251
<b>    6.5.8 View History .....</b>	251
<b>    6.5.9 Manage Users .....</b>	252
<b>    6.5.10 Rebuild Diagnostic Model .....</b>	253
<b>Chapter 7: Software Testing Plan.....</b>	255
<b>    7.1 Test Items .....</b>	255
<b>    7.2 Features to be Tested.....</b>	255
<b>    7.3 Approach .....</b>	255
<b>        7.3.1 Data and Database Integrity Testing .....</b>	255
<b>        7.3.2 Component Testing .....</b>	256
<b>        7.3.3 Integration Testing .....</b>	295
<b>        7.3.4 User Interface Testing .....</b>	299
<b>        7.3.5 Interface Testing .....</b>	299
<b>        7.3.6 Validation and Verification Testing .....</b>	299
<b>        7.3.7 Security Testing .....</b>	300

<b>7.3.8 Performance Testing .....</b>	300
<b>7.3.9 Constraints.....</b>	300
<b>7.3.10 Beta Testing.....</b>	300
<b>7.3.11 Acceptance Testing .....</b>	300
<b>7.3.12 Test report (ABET* Mandatory) .....</b>	301
<b>7.4 Pass/Fail Testing .....</b>	301
<b>7.5 Testing Process.....</b>	301
<b>7.5.1 Test Deliverables.....</b>	301
<b>7.5.2 Testing Tasks.....</b>	301
<b>7.5.3 Responsibilities .....</b>	301
<b>7.5.4 Resources .....</b>	302
<b>7.5.5 Schedule .....</b>	302
<b>7.5.6 Environmental Requirements .....</b>	302
<b>7.6 Comparison of design choice: (ABET requirement).....</b>	303
<b>7.6.1 Osteoporosis disease.....</b>	303
<b>7.6.2 Epileptic seizure disease .....</b>	304
<b>7.6.3 Sickle cell anemia disease .....</b>	305
<b><i>Chapter 8: Conclusion .....</i></b>	<b>307</b>
<b>8.1 introduction .....</b>	307
<b>8.2 Finding and Contributions .....</b>	307
<b>8.3 Entrepreneurship Impact.....</b>	307
<b>8.4 Issue Faced .....</b>	308
<b>8.5 Lessons Learned .....</b>	308
<b>8.6 Client requirements.....</b>	309
<b>8.6 Recommendations for future works .....</b>	309
<b>References.....</b>	<b>310</b>
<b>APPENDIX .....</b>	<b>321</b>
<b>Appendix A: Response to comments .....</b>	321
<b>Appendix B: plagiarism check .....</b>	323
<b>Appendix C.1: Bi-weekly reports 1.....</b>	324
<b>Appendix C.2: Bi-weekly reports 2.....</b>	325
<b>Appendix C.3: Bi-weekly reports 3.....</b>	326
<b>Appendix C.4: Bi-weekly reports 4.....</b>	327
<b>Appendix D.1: Osteoporosis Journal Paper .....</b>	328
<b>Appendix D.2: Osteoporosis Conference Paper .....</b>	354

<b>Appendix E.1: Epileptic Seizure Journal Paper.....</b>	361
<b>Appendix E.2: Epileptic Seizure Conference Paper .....</b>	391
<b>Appendix F.1: Sickle Cell Anemia Journal Paper .....</b>	401
<b>Appendix F.1: Sickle Cell Anemia Conference Paper.....</b>	426
<b>Appendix G: Final Presentation .....</b>	436

## Table of Tables

<b>Table 1 List of Abbreviations.....</b>	20
<b>Table 2 Epileptic Seizure Literature Review Summary .....</b>	43
<b>Table 3 Osteoporosis Literature Review Summary .....</b>	47
<b>Table 4 Sickle Cell Anemia Literature Review Summary .....</b>	51
<b>Table 5 Project Deliverables .....</b>	54
<b>Table 6 The project Schedule .....</b>	54
<b>Table 7 Terms and Definitions .....</b>	55
<b>Table 8 Roles and Responsibilities .....</b>	57
<b>Table 9 Human recourses .....</b>	59
<b>Table 10 Skills needed for each Role.....</b>	59
<b>Table 11 Work Breakdown Structure .....</b>	62
<b>Table 12 Schedule Allocation.....</b>	65
<b>Table 13 Resource Allocation .....</b>	66
<b>Table 14 Project Metrics .....</b>	67
<b>Table 15 Risk Management Plan.....</b>	68
<b>Table 16 The Waterfall Phases .....</b>	69
<b>Table 17 Used Tools .....</b>	69
<b>Table 18 The Essential Requirements .....</b>	70
<b>Table 19 Supporting Process Plans .....</b>	71
<b>Table 20 resource scheduling .....</b>	72
<b>Table 21 User Characteristics.....</b>	87
<b>Table 22 The Main Fields of the Login Interface.....</b>	88
<b>Table 23 The Main Fields of the Create Account Interface .....</b>	88
<b>Table 24 The Main Fields of the Profile Interface .....</b>	89
<b>Table 25 The Main Fields of the Change Email Interface .....</b>	89
<b>Table 26 The Main Fields of the Change Password Interface.....</b>	89
<b>Table 27 The Main Fields of the Manage Users Tab in the Admin Interface .....</b>	90
<b>Table 28 The Main Fields of the Rebuild and Update Model Tab in the Admin Interface ..</b>	91
<b>Table 29 The Main Fields of the Medical Specialist Interface .....</b>	91
<b>Table 30 The Main Fields of the Laboratory Specialist Interface.....</b>	92
<b>Table 31 The Main Fields of the Registered User Interface .....</b>	93
<b>Table 32 The Main Fields to Diagnose Diabetes Mellitus .....</b>	94
<b>Table 33 The Main Fields to Diagnose Chronic Kidney Disease .....</b>	94
<b>Table 34 The Main Fields to Diagnose Coronary Heart Disease .....</b>	95
<b>Table 35 The Main Fields to Diagnose Asthma Disease.....</b>	95
<b>Table 36 The Main Fields to Diagnose Thyroid Cancer .....</b>	96

Table 37 The Main Fields to Diagnose Schizophrenia .....	96
Table 38 The Main Fields to Diagnose Glaucoma .....	97
Table 39 The Main Fields to Diagnose Alzheimer's Disease .....	97
Table 40 The Main Fields to Diagnose Lung Cancer.....	98
Table 41 The Main Fields to Diagnose Rheumatoid Arthritis .....	99
Table 42 The Main Fields to Diagnose Hypothyroidism .....	99
Table 43 The Main Fields to Diagnose Prostate Cancer .....	99
Table 44 The Main Fields to Diagnose Cervical Cancer .....	100
Table 45 The Main Fields to Diagnose Multiple Sclerosis .....	100
Table 46 The Main Fields to Diagnose Liver Cirrhosis .....	101
Table 47 The Main Fields to Diagnose Chronic Obstructive Pulmonary Disease.....	101
Table 48 The Main Fields to Diagnose Parkinson's Disease .....	102
Table 49 The Main Fields to Diagnose Hepatitis C .....	103
Table 50 The Main Fields to Diagnose Depression .....	104
Table 51 The Main Fields to Epileptic Seizure .....	105
Table 52 The Main Fields to Diagnose Osteoporosis .....	106
Table 53 The Main Fields to Diagnose Sickle Cell Anemia .....	107
Table 54 The Main Fields in the Diagnosis Interface .....	107
Table 55 The Main Field of The Result Interface .....	108
Table 56 Login Functionality .....	109
Table 57 Retrieve Password Functionality.....	110
Table 58 Update Profile Functionality .....	111
Table 59 View Diagnosis History Functionality .....	112
Table 60 Diagnose Functionality .....	121
Table 61 Print Result Functionality.....	122
Table 62 Rebuild Diagnosis Model for Admin Functionality .....	122
Table 63 Replace Diagnosis Model Functionality .....	123
Table 64 Add User Functionality .....	124
Table 65 Remove User Functionality .....	125
Table 66 View Diagnosis History for Medical Specialists.....	125
Table 67 View Storing Data and Performing an Overall Diagnosis .....	127
Table 68 Create Account Functionality .....	127
Table 69 Off-Record Diagnosis Functionality .....	128
Table 70 Screen Object and Actions .....	162
Table 71 The Database Entities and Fields .....	173
Table 72 Data Dictionary .....	174
Table 73 Component, Classification, Definition and Responsibility .....	181
Table 74 Component, Limitation Pre-condition, and post-condition.....	182
Table 75 Description of "Login" Component .....	188
Table 76 Description of "Retrieve Password" Component.....	188
Table 77 Description of "Update Profile" Component .....	189
Table 78 Description of "View Diagnosis History" Component .....	189
Table 79 Description of "Diagnose" Component.....	198
Table 80 Description of "Print Result" Component.....	198
Table 81 Description of "Rebuild Diagnosis Model" Component.....	198

Table 82 Description of “Update Diagnosis Model” Component.....	198
Table 83 Description of “Add User” Component .....	199
Table 84 Description of “Remove User” Component.....	199
Table 85 Description of “Create Account” Component.....	200
Table 86 Description of “Overall Diagnosis” Component.....	200
Table 87 Requirement Tractability Matrix .....	202
Table 88 Features’ description.....	204
Table 89 Statistical analysis of numerical features .....	204
Table 90 the optimal hyperparameter for each classifier in the original data .....	208
Table 91 the optimal hyperparameter for each classifier in over-sampled data with all the features selected .....	209
Table 92 The results of the proposed models before and after sampling was applied.....	209
Table 93 Testing Accuracy for Different Feature Subsets.....	211
Table 94 Confusion Matrix for (a) RF, (b) GBoost, (c) SVM, (d) KNN, and (e) XGBoost .....	212
Table 95 Features’ description.....	214
Table 96 Statistical analysis of numerical features .....	215
Table 97 the best hyperparameter selected on the original data.....	221
Table 98 The results of the proposed models before and after sampling was applied.....	222
Table 99 result with feature selection.....	223
Table 100 Confusion Matrix of (a) RandomForest, (b) SVM, (c) KNN, (d) Gboost, (e) XGBoost, (f) LoR, (g) DT, (h) AdaBoost, .....	225
Table 101 Features’ description.....	227
Table 102 Statistical analysis for numerical attribute .....	228
Table 103 Best hyperparameter with original dataset .....	232
Table 104 Best hyperparameters with original dataset and features selection.....	232
Table 105 5 Best hyperparameters with oversampled dataset .....	233
Table 106 The results of the proposed models before and after sampling was applied....	234
Table 107 Results with feature selection.....	236
Table 108 Confusion Matrix of (a) RandomForest, (b) SVM, (c) KNN, (d) Gboost, (e) XGBoost, (f) LoR, (g) DT, (h) AdaBoost, (i) GaussianNB and (j) SVM_Gaussian .....	238
Table 109 Data and database integrity test cases .....	256
Table 110 Test Cases of ‘Login’ feature.....	257
Table 111 Test Cases of ‘Reset Password’ feature .....	258
Table 112 Test Cases of ‘Create Account’ feature.....	259
Table 113 Test cases of 'Update Profile' feature.....	260
Table 114 Test cases of 'Change Email' feature .....	261
Table 115 Test cases of 'Change Password' feature.....	262
Table 116 Add User Test Case.....	262
Table 117 Delete User Test Case.....	263
Table 118 Rebuild Model Test Case.....	263
Table 119 Update Model Test Case.....	264
Table 120 Medical Specialist’s View of Diagnosis History Test Case.....	264
Table 121 Registered User’s View of Diagnosis History Test Case .....	264
Table 122 Laboratory Specialist’s Diagnosis Test Case.....	265

Table 123 Diagnose Diabetes Mellitus Disease Test Case.....	266
Table 124 Diagnose Chronic Kidney Disease Test Case .....	266
Table 125 Diagnose Coronary Heart Disease Test Case .....	268
Table 126 Diagnose Asthma Disease Test Case.....	269
Table 127 Diagnose Thyroid Cancer Disease Test Case .....	270
Table 128 Diagnose Schizophrenia Disease Test Case .....	271
Table 129 Diagnose Glaucoma Disease Test Case.....	272
Table 130 Diagnose Alzheimer's Disease Test Case.....	273
Table 131 Diagnose Lung Cancer Disease Test Case .....	275
Table 132 Diagnose Rheumatoid Arthritis Disease Test Case .....	277
Table 133 Diagnose Hypothyroidism Disease Test Case.....	278
Table 134 Diagnose Prostate Cancer Disease Test Case .....	278
Table 135 Diagnose Cervical Cancer Disease Test Case .....	279
Table 136 Diagnose Multiple Sclerosis Disease Test Case.....	281
Table 137 Diagnose Liver Cirrhosis Disease Test Case.....	282
Table 138 Diagnose Chronic Obstructive Pulmonary Disease Test Case .....	283
Table 139 Diagnose Parkinson's Disease Test Case .....	285
Table 140 Diagnose Hepatitis C Disease Test Case .....	287
Table 141 Diagnose Depression Disease Test Case .....	289
Table 142 Diagnose Epileptic Seizure Disease Test Case.....	292
Table 143 Diagnose Osteoporosis Disease Test Case .....	294
Table 144 Table 143 Diagnose Sickle Cell Anemia Disease Test Case .....	295
Table 145 View Diagnostic Result .....	295
Table 146 Integration Test Case of 'Admin Home' Interface .....	296
Table 147 Integration Test Case of 'Medical Specialist Home' Interface.....	297
Table 148 Integration Test Case of 'Laboratory Specialist Home' Interface .....	297
Table 149 Integration Test Case of 'Registered User Home' Interface .....	298
Table 150 Integration Test Case of 'Guest Home' Interface .....	298
Table 151 Testing Tasks .....	301
Table 152 Software Testing Resources .....	302
Table 153 Testing Tasks Schedule.....	302

## Table of Figures

Figure 1 Project Phases Block Diagram .....	24
Figure 2 External Interface.....	56
Figure 3 Internal Structure .....	57
Figure 4 The Waterfall Model .....	68
Figure 5 Operational Framework .....	76
Figure 6 Majority Voting in Random Forest [72].....	77
Figure 7 Hyperplane that separates two classes with support vectors points [81].....	78
Figure 8 Mapping nonlinearly separable to a higher dimension using kernel function [85]	79
Figure 9 The system use-case diagram .....	86
Figure 10 The Activity Diagram of the Login Functionality .....	109
Figure 11 The Activity Diagram of the Retrieve Password Functionality .....	110
Figure 12 The Activity Diagram of the Update Profile Functionality .....	111
Figure 13 The Activity Diagram of the View History Functionality .....	112
Figure 14 The Activity Diagram of the Diagnose Functionality.....	121
Figure 15 The Activity Diagram of the Print Result Functionality.....	122
Figure 16 The Activity Diagram of the Rebuild Diagnosis Model Functionality.....	123
Figure 17 The Activity Diagram of the Replace Diagnosis Model Functionality .....	123
Figure 18 The Activity Diagram of the Add User Functionality .....	124
Figure 19 The Activity Diagram of the Remove User Functionality.....	125
Figure 20 The Activity Diagram of View Diagnosis History Functionality .....	126
Figure 21 The Activity Diagram of Storing Data and Performing an Overall Diagnosis Functionality .....	127
Figure 22 The Activity Diagram of the Create Account Functionality .....	128
Figure 23 The Activity Diagram of Off-Record Diagnosis Functionality .....	128
Figure 24 Login Interface.....	137
Figure 25 Retrieve Password Interface .....	137
<i>Figure 26 Create Account Interface .....</i>	138
<i>Figure 27 Admin Profile Tab .....</i>	138
<i>Figure 28 Change Email Interface .....</i>	139
Figure 29 Change Password Interface .....	139
<i>Figure 30 Manage Users Tab .....</i>	140
<i>Figure 31 Add User Tab .....</i>	140
<i>Figure 32 Remove User Tab .....</i>	141

<i>Figure 33 Rebuild Model Tab .....</i>	142
<i>Figure 34 Medical Specialist Profile Interface.....</i>	142
<i>Figure 35 Diagnosis History Tab .....</i>	143
<i>Figure 36 Diagnose Diabetes Mellitus Tab .....</i>	143
<i>Figure 37 Diagnose Chronic Kidney Disease Tab .....</i>	144
<i>Figure 38 Diagnose Coronary Heart Disease Tab .....</i>	144
<i>Figure 39 Diagnose Asthma Tab .....</i>	145
<i>Figure 40 Diagnose Thyroid Cancer Tab .....</i>	145
<i>Figure 41 Diagnose Schizophrenia Tab .....</i>	146
<i>Figure 42 Diagnose Glaucoma Tab.....</i>	146
<i>Figure 43 Diagnose Alzheimer's Disease Tab.....</i>	147
<i>Figure 44 Diagnose Lung Cancer Tab .....</i>	147
<i>Figure 45 Diagnose Rheumatoid Arthritis Tab .....</i>	148
<i>Figure 46 Diagnose Hypothyroidism Tab.....</i>	148
<i>Figure 47 Diagnose Prostate Cancer Tab.....</i>	149
<i>Figure 48 Diagnose Cervical Cancer Tab.....</i>	149
<i>Figure 49 Diagnose Multiple Sclerosis Tab .....</i>	150
<i>Figure 50 Diagnose Liver Cirrhosis Tab.....</i>	150
<i>Figure 51 Diagnose Chronic Obstructive Pulmonary Disease Tab .....</i>	151
<i>Figure 52 Diagnose Parkinson's Disease Tab .....</i>	151
<i>Figure 53 Diagnose Hepatitis C Tab .....</i>	152
<i>Figure 54 Diagnose Depression Tab .....</i>	152
<i>Figure 55 Diagnose Epileptic Seizure Tab .....</i>	153
<i>Figure 56 Diagnose Osteoporosis Tab .....</i>	153
<i>Figure 57 Diagnose Sickle Cell Anemia Tab .....</i>	154
<i>Figure 58 Diagnosis Result Interface .....</i>	154
<i>Figure 59 Laboratory Specialist Profile Interface .....</i>	155
<i>Figure 60 Laboratory Test Tab.....</i>	155
<i>Figure 61 Registered User Profile Interface .....</i>	156
<i>Figure 62 Registered Users' Diagnosis History Tab .....</i>	157
<i>Figure 63 Invalid Login Error message .....</i>	163
<i>Figure 64 Third Invalid Login Error Message .....</i>	163
<i>Figure 65 Password Expired Error Message .....</i>	163
<i>Figure 66 Non-existing Email Error Message .....</i>	163
<i>Figure 67 Third Invalid Email Error Message .....</i>	164
<i>Figure 68 Invalid Email Error Message.....</i>	164
<i>Figure 69 Invalid Confirmation Email Error Message .....</i>	164
<i>Figure 70 Invalid Password Error Message.....</i>	164
<i>Figure 71 Invalid Confirmation Password Error Message .....</i>	165
<i>Figure 72 Invalid Old Password Error Message.....</i>	165
<i>Figure 73 Incomplete Information Error Message .....</i>	165
<i>Figure 74 Invalid Dataset Error Message .....</i>	165
<i>Figure 75 Invalid MRN Error Message .....</i>	165
<i>Figure 76 Invalid Name Error Message .....</i>	166
<i>Figure 77 Save Changes Confirmation Message.....</i>	166

<i>Figure 78 Update Model Confirmation Message .....</i>	166
<i>Figure 79 Remove User Confirmation Message .....</i>	166
<i>Figure 80 Deleting an Account Confirmation Message .....</i>	167
<i>Figure 81 Retrieve Password Information Message.....</i>	167
<i>Figure 82 Update Model Information Message.....</i>	167
<i>Figure 83 User Added Information Message.....</i>	167
<i>Figure 84 User Deleted Information Message.....</i>	168
<i>Figure 85 Deleting an Account Information Message.....</i>	168
<i>Figure 86 Architectural Design Diagram.....</i>	169
<i>Figure 87 General View of the System.....</i>	170
<i>Figure 88 All users' subsystems .....</i>	170
<i>Figure 89 Admin Subsystem.....</i>	170
<i>Figure 90 Medical Specialist Subsystem.....</i>	171
<i>Figure 91 Laboratory Specialist Subsystem .....</i>	171
<i>Figure 92 Registered User Subsystem .....</i>	172
<i>Figure 93 Guest User Subsystem .....</i>	172
<i>Figure 94 Entity Relationship Diagram .....</i>	175
<i>Figure 95 ERD Mapping.....</i>	175
<i>Figure 96 The “Login” sequence diagram .....</i>	183
<i>Figure 97 The “Retrieve Password” sequence diagram. ....</i>	183
<i>Figure 98 The “Change Email” sequence diagram .....</i>	184
<i>Figure 99 The “Change Password” sequence diagram .....</i>	184
<i>Figure 100 Change Personal Information.....</i>	184
<i>Figure 101 The “View Diagnosis History” sequence diagram .....</i>	185
<i>Figure 102 The “Diagnose” sequence diagram. ....</i>	185
<i>Figure 103 The “Print Result” sequence diagram .....</i>	185
<i>Figure 104 The “Rebuild Diagnosis Model” sequence diagram .....</i>	186
<i>Figure 105 The “Update Diagnosis Model” sequence diagram .....</i>	186
<i>Figure 106 The “Add User” sequence diagram .....</i>	186
<i>Figure 107 The “Remove User” sequence diagram.....</i>	187
<i>Figure 108 The “Create Account” sequence diagram .....</i>	187
<i>Figure 109 The “Overall Diagnosis” sequence diagram .....</i>	187
<i>Figure 110 description of the categorical features .....</i>	205
<i>Figure 111 The proposed framework for the pre-emptive diagnosis of Osteoporosis and osteopenia .....</i>	206
<i>Figure 112 Number of samples of each class before and after applying SMOTETomek..</i>	210
<i>Figure 113 (a) RF, (b) GradientBoosting, (c) SVM, (d) KNN, and (e) XGBoost ROC Curve .....</i>	213
<i>Figure 114 distribution of categorical features .....</i>	218
<i>Figure 115 The proposed framework for the pre-emptive diagnosis of Epileptic Seizures</i>	219
<i>Figure 116 Number of samples of each class of the target feature before and after applying SMOTETomek. ....</i>	222
<i>Figure 117 ROC curves of the eight models applied. ....</i>	226
<i>Figure 118 The distribution of gender attribute .....</i>	228
<i>Figure 119 The proposed framework for the pre-emptive diagnosis of SCA .....</i>	229

Figure 120 Number of samples of each class of the target feature before and after applying SMOTETomek.....	235
Figure 121 Feature selection process .....	235
Figure 122 ROC curves of the ten models applied. ....	240
Figure 123 Login' Interface .....	241
Figure 124 Forgot Password' Interface .....	241
Figure 125 Confirmation Email .....	242
Figure 126 Registration Form' Interface.....	242
Figure 127 Confirmation email after the user creates an account.....	243
Figure 128 Confirmation email after a user changes their email address. ....	243
Figure 129 The email is confirmed successfully.....	244
Figure 130 URL for an email has been already confirmed.....	244
Figure 131 URL has been expired.....	245
Figure 132 Admin's Profile.....	245
Figure 133 Medical Specialist's Profile .....	246
Figure 134 Laboratory Specialist's Profile .....	246
Figure 135 Registered User's Profile .....	246
Figure 136 Change Email .....	247
Figure 137 Change Password.....	248
Figure 138 Change Password Confirmation Email.....	248
Figure 139 Diagnosis form .....	249
Figure 140 'Medical Specialist Diagnosis' Interface .....	250
Figure 141 'Laboratory Specialist Diagnosis' Interface .....	250
Figure 142 Result Interface.....	251
Figure 143 Medical Specialist's View of Diagnosis History' interface .....	251
Figure 144 Registered User's View of Diagnosis History' Interface .....	252
Figure 145 Manage Users Interface .....	252
Figure 146 Add User .....	253
Figure 147 Add User Email .....	253
Figure 148 Rebuild Model .....	253
Figure 149 ROC-AUC for the RF classifier .....	304
Figure 150 a graph shows AUC for LR classifier. ....	305
Figure 151 a graph shows AUC of RF classifier.....	306

## LIST OF ABBREVIATIONS

<b>AdaBoost</b>	Adaptive Boosting
<b>ANN</b>	Artificial Neural Network
<b>ApEn</b>	Approximate Entropy
<b>AUC</b>	Area Under the Curve
<b>AUROC</b>	Area Under the Receiver Operating Characteristic
<b>AutoML</b>	Automated machine learning
<b>BMD</b>	Bone Mineral Density
<b>BMI</b>	Body Mass Index
<b>C&amp;R</b>	Classification and Regression tree
<b>Catboost</b>	Categorical Boosting
<b>CNN</b>	Convolution Neural Network
<b>COA</b>	Coyote optimization algorithm
<b>CT</b>	Computed Tomography
<b>DCNNs</b>	Deep Convolutional Neural Networks
<b>DCSAE-ESDC</b>	Deep canonical sparse autoencoder-based epileptic seizure detection and classification
<b>DNN</b>	Deep Neural Network
<b>DT</b>	Decision Tree
<b>DWT</b>	Discrete Wavelet Transform
<b>DXA</b>	Dual-Energy X-Ray Absorptiometry- Dual-energy x-ray absorption
<b>EC</b>	Ensemble Classifier
<b>ECOC</b>	Error-Correcting Output Codes
<b>EEG</b>	Electroencephalogram
<b>EMRs</b>	Electronic Medical Records
<b>ET</b>	Extra Trees
<b>ETFS</b>	Extra-Tree Feature Selection
<b>ELM</b>	Extreme Learning Machine
<b>FFNN</b>	Feed-Forward Neural Network
<b>FFT</b>	Fast Fourier transform

<b>FN</b>	False Negative
<b>FNC</b>	Functional Network Connectivity
<b>FP</b>	False Positive
<b>FR</b>	Face recognition
<b>FuzzyEn</b>	Fuzzy Entropy
<b>GB</b>	Gradient boosting
<b>GBM</b>	Gradient Boosting
<b>GBT</b>	Gradient Boosting Trees
<b>Gboost</b>	Gradient Boosting
<b>HEXA</b>	Health Examinee
<b>ICUs</b>	Intensive Care Units
<b>iEEG</b>	intracranial electroencephalography
<b>KHA</b>	Krill herd algorithm
<b>KNN</b>	K-Nearest Neighbors
<b>lightGBM</b>	light gradient-boosting machine
<b>ResNet</b>	Residual Network
<b>LoR</b>	Logistic Regression
<b>ROC</b>	Receiver Operating Characteristic
<b>LSTM</b>	Long Short-Term Memory
<b>ML</b>	Machine Learning
<b>MLP</b>	Multilayer Perceptron
<b>ORAI</b>	Osteoporosis Risk Assessment Instrument
<b>OSIRIS</b>	Osteoporosis Index Of Risk
<b>OST</b>	Osteoporosis Self-Assessment Tool
<b>PCA</b>	Principal Component Analysis
<b>PKP</b>	Percutaneous Kyphoplasty
<b>RBC</b>	Red Blood Cells
<b>RF</b>	Random Forest
<b>RNN</b>	Recurrent neural network
<b>ROI</b>	Region Of Interest
<b>SampEn</b>	Sample Entropy
<b>SCA</b>	Sickle Cell Anemia
<b>SCD</b>	Sickle Cell Diseases
<b>SCORE</b>	Simple Computed Osteoporosis Risk Estimation
<b>SFFS</b>	Sequential Forward Floating Selection
<b>SMOTE</b>	Synthetic Minority Oversampling Technique
<b>STD</b>	Standard Deviation
<b>SVM</b>	Support Vector Machine
<b>UCI</b>	University of California Irvine
<b>UV-Vis</b>	Ultraviolet-visible

<b>VGG</b>	Visual Geometry Group
<b>XGboost</b>	Extreme Gradient Boosting
<b>XAI</b>	Explainable Artificial Intelligence
<b>Xception</b>	Extreme Inception

*Table 1 List of Abbreviations*

## Chapter 1: Introduction

This chapter serves as a concise introduction to the project, offering a comprehensive overview of its key components, including a discussion on the scope and limitations of the identified problem. Following this, a delineation of the underlying causes and their justifications is presented, along with a clear outline of the goals that must be attained. The chapter culminates with an in-depth examination of the project's organizational structure and methodology.

### 1.1 Introduction

Chronic diseases result from a combination of genetic, environmental, and lifestyle factors, often lasting a long time. Chronic diseases pose a significant challenge globally, with over three-quarters of the 31.4 million chronic diseases related deaths occurring in regions facing economic challenges. These diseases affect people of all ages and regions, with 17 million chronic diseases related deaths happening before age 70. Alarmingly, a vast majority of these early deaths occur in areas with limited economic resources. Risk factors like unhealthy diets, lack of exercise, exposure to tobacco smoke, and pollution make children, adults, and the elderly vulnerable. These diseases are on the rise due to factors like rapid urbanization, globalization, and an aging population [1]. They're closely tied to poverty, hindering efforts to reduce it, as chronic diseases bring hefty healthcare costs. Costly treatments and income loss push millions into poverty annually [2]. To combat chronic diseases, focus on reducing modifiable risk factors and monitor progress. A unified approach across various sectors, including health, finance, education, and more, is crucial. Investing in chronic disease management, including early detection and timely treatment, is essential and economically wise. By using the latest technologies, the aim of this project is to provide supportive care for individuals' health by implementing a pre-emptive diagnostic system using Machine Learning (ML) technologies.

In Saudi Arabia, chronic diseases are prevalent, with a variety of enduring health conditions such as osteoporosis and epileptic seizures. According to the Kingdom of Saudi Arabia's

National Plan for Osteoporosis Prevention and Management, the prevalence of osteoporosis and its precursor, osteopenia, is 37.8% in men and 28.2% in women over the age of 50. With the projected demographic shift in Saudi Arabia, where the 50+ age group is expected to rise steeply over the next few decades, and life expectancy on the rise, the Kingdom anticipates a surge in osteoporosis cases [3]. Moreover, according to Rumayyan et al. [4], in Saudi Arabia, epilepsy exerts a notable impact, affecting approximately 6.54 individuals per 1000 in the population. Sickle cell disease presents a significant health concern in Saudi Arabia, with a prevalence of over 45,100 cases per 1,000,000 among adults, as reported by Zuair et al. [5] Among children and adolescents in the Saudi population, an estimated 2400 per 1,000,000 are affected by this inherited blood disorder. This high prevalence underscores the importance of robust screening, early intervention, and comprehensive healthcare services for individuals living with sickle cell disease in the country. Consequently, this project focuses on the pre-emptive diagnosis of Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia.

Epilepsy is a neurological disorder that causes repeated seizures caused by a momentary disruption in brain electrical activity. Typically, the brain emits small, organized electrical signals that travel through the intricate network of nerve cells to various parts of the body via neurotransmitters. However, in epilepsy, this delicate balance of electrical rhythms is disturbed, leading to the occurrence of repeated seizures [6]. These seizures manifest as abrupt and synchronized surges of electrical energy, potentially resulting in transient alterations in consciousness, movement, or sensory perception.

Additionally, osteoporosis is a condition that arises when there's a decline in bone mineral density and mass, or when the composition and strength of bones undergo changes [7]. This weakening of bones can significantly elevate the risk of fractures. Often referred to as a "silent" ailment, osteoporosis typically progresses without noticeable symptoms, and individuals may only become aware of it when they experience a bone break. This disease is a primary cause of fractures in older men and postmenopausal women, frequently occurring in bones like the hip, spine vertebrae, and wrist. It's important to note that osteoporosis doesn't discriminate based on race or ethnicity, affecting individuals from all backgrounds [8]. While the risk of developing osteoporosis increases with age, it can manifest at any stage of life.

Furthermore, sickle cell disease (SCD) is a prevalent hematological condition, impacting millions globally. It arises from a genetic mutation, causing the formation of abnormal hemoglobin [5]. This can lead to persistent symptoms that have the potential to affect various organs within the body. Additionally, this can result in episodes of severe pain, known as sickle cell crises, as well as an increased risk of infections and various complications. While advancements in medical treatment have improved the management of SCD, it remains a chronic condition that requires ongoing care and support.

Over the past decade, the transformative influence of ML in the realm of healthcare has been undeniable, particularly in the realm of personalized medicine. ML models have demonstrated remarkable capabilities when dealing with complex and voluminous datasets, paving the way for tailored healthcare approaches. By harnessing cutting-edge ML techniques, researchers have unlocked a realm of potential within clinical studies of these conditions, offering exciting prospects for precision medicine in both research and practical healthcare settings [9]. Inspired by the potential of ML to revolutionize early disease identification [10], we are compelled to expand our work into previously uncharted territory. Specifically, we are embarking on investigations into the domains of Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia.

While numerous research efforts have made strides in forecasting and identifying these three medical conditions, a majority of them relied on specialized imaging technology, which might not be accessible in all healthcare facilities, thus constraining the widespread utilization of these predictive models. Additionally, previous investigations predominantly concentrated on disease diagnosis once symptoms had manifested in the patient. As a response to these limitations, this project aims to address the issue by employing readily available clinical data to train and evaluate a variety of ML algorithms for the proactive diagnosis of these chronic diseases.

## 1.2 Problem Statement

Chronic diseases are considered life-threatening for human lives. In addition to causing disease, it may cause permanent disability, whether mentally or physically. People who have a chronic disease require constant medical attention and may need to take medications for the rest of their lives. To diagnose these diseases, clinicians use laboratory test results and other methods. These types of diagnostics do not always lead to a correct diagnosis because of the similarities between chronic diseases and other kinds of diseases. In Saudi Arabia, while chronic diseases such as Osteoporosis and Sickle Cell Anemia are prevalent, the relatively lower incidence of Epilepsy does not diminish its gravity. Epilepsy stands as one of the most common neurological conditions globally, affecting over 50 million people worldwide. Alarmingly, Saudi Arabia reports a rate of 6.54 out of 1000 citizens suffering from Epilepsy, necessitating immediate attention [11]. Epilepsy patients could be treated with surgery or medications. Nevertheless, those types of treatments may not provide the desired result if the seizures have already accrued. Then the epileptic seizure cannot be controlled. And here comes the importance of predicting these seizures before they happen to prevent any other consequences, such as major depression [12]. Moreover, Osteoporosis is a widely common condition that causes bone weakness and susceptibility to fractures. According to the Saudi Ministry of Health, women are more susceptible to being infected by osteoporosis, especially after menopause. Bone fractures, disability, and even death are some of the complications of the disease [13]. Osteoporosis could be diagnosed using bone density measurement, or DXA. However, it is not widely available because of its cost [14]. Furthermore, sickle cell anemia is a serious blood disease that is transmitted genetically and affects children from the time they are born. Sickle Cell Anemia causes severe and chronic pain and damages some body organs, in addition to other threatening complications. Thus,

early detection of this disease could prevent and control any complications using available treatment methods [15]. Notably, the eastern region of Saudi Arabia bears the highest burden of Sickle Cell Anemia compared to other regions, as indicated by a recent study [16].

As discussed earlier, the spreading of these diseases could be suppressed using pre-emptive diagnostic techniques. Clinical data has increased massively; with the appropriate use of these data, we can reduce all these risks of late detection of chronic diseases. Machine learning algorithms can be employed to detect this disease in its early stages to avoid any further complications. Therefore, a better quality of life is possible for people who have the potential to suffer from chronic diseases.

### **1.3 Motivation**

1. Helping Saudi Arabia's Vision 2030 goal of transforming the health sector.
2. Reducing the healthcare burden by diagnosing and addressing chronic disease issues before they become serious, thus contributing to the sustainability and efficiency of the healthcare industry.
3. Contributing to population health by proactively reducing the incidence of specific chronic diseases through early diagnosis.
4. Assisting medical practitioners in the early detection of chronic diseases, allowing them to deliver timely and tailored interventions.

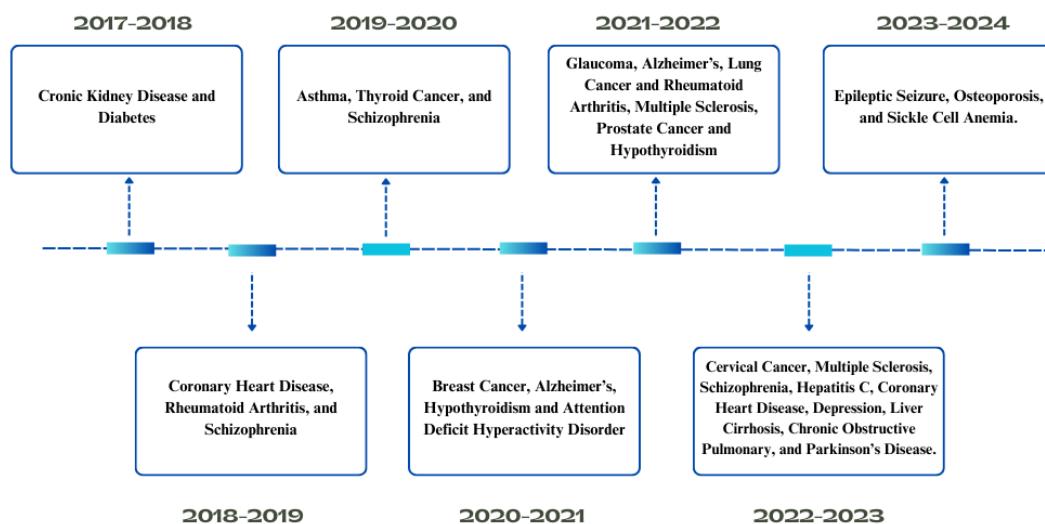
### **1.4 Justification**

Given that our population is expanding, and more people are becoming ill, especially from chronic diseases that endanger the health of our country, the Kingdom's Vision 2030 calls for the consideration of a digital transformation program in the health sector. Saudi Arabia must therefore approach the problem of chronic diseases from several digital angles, and we think that ML methods can help us realize our nation's goals. As a result, we have decided to use ML to enhance supportive care to pre-emptively diagnose three serious chronic diseases in this project phase: Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia. This will improve the quality of life for the population.

### **1.5 Aims & Objectives**

The project aims to improve the Saudi Arabian population's access to healthcare by integrating ML techniques in the early preemptive detection of selected chronic diseases before symptoms manifest, allowing for timely actions and improving overall health outcomes. Previous phases of this project have already completed implementing pre-emptive diagnosis models for Diabetes, Coronary Heart Disease, Chronic Kidney Disease, Schizophrenia, Asthma, Breast Cancer, Rheumatoid Arthritis, Thyroid Cancer, Alzheimer's, Hypothyroidism, Attention Deficit Hyperactivity Disorder, Glaucoma, Lung Cancer, Prostate Cancer, Cervical Cancer, MS, Schizophrenia, Hepatitis C, Depression, Liver Cirrhosis, Chronic Obstructive Pulmonary, and Parkinson's Disease.

In this seventh stage, we prepare to cover Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia. The project seeks to improve the performance of previously established models and achieve the best accuracy using fewer features while applying ML models using Saudi datasets.



*Figure 1 Project Phases Block Diagram*

## 1.6 Scope / Limitation of the Study

This project focuses on the pre-emptive diagnosis of Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia in Saudi Arabia. Due to time restrictions, we initially planned to focus the project on three chronic conditions. However, if time allows, the work's scope could be expanded to include more ailments. Since the project's present phase is restricted to Saudi Arabian society, the datasets will be gathered from Saudi Arabian hospitals. However, if we are unable to acquire local data, we will use online databases to achieve our goals. Simple clinical data are the only data types allowed for this project since they are less expensive, more accessible, and simpler to analyze than imaging data. Thus, there will be less need for expensive imaging technology to be purchased, and local hospitals without imaging equipment will be able to detect the mentioned diseases earlier.

## 1.7 Social, Professional, Legal, and Ethical Implication

Our team came up with this idea entirely on its own. The project's purpose is to enhance Saudi Arabian citizens' quality of life and health. It also supports hospitals in offering better service to their patients, reducing costs, and enhancing their competitiveness in the Early Detection and Prevention and Delivering Specialized Care categories of the Ada'a Health Award. Furthermore, it aligns with the digital health transformation currently supported by the Saudi Arabian Health Ministry and the Saudi 2030 Vision by providing cutting-edge technology that supports healthcare professionals to preemptively diagnose chronic diseases resulting in improving the overall health outcomes of the population.

Our project is designed to comply with privacy laws and regulations to protect the confidentiality of patients while being user-friendly and easy to navigate without requiring any technical knowledge from the users. Additionally, the data used to train the models are ethically approved and do not contain any personal information. With the innovative features and functionalities of our project, we strongly believe that it can make a significant impact on the healthcare industry in Saudi Arabia, ultimately benefiting the entire population.

## 1.8 Project Organization

Project organization refers to the structure that defines how a project is planned, executed, managed, and monitored. The project management process in this report is divided into six chapters. The key points that succinctly explain each phase are listed below:

### ❖ Chapter 1: Introduction

The introduction section of project organization provides an initial overview and context for the entire project. It outlines the purpose, objectives, and scope of the project organization's efforts. This section serves as a foundation for understanding the project's goals and the strategies employed to achieve them.

### ❖ Chapter 2: Background and Review of Literature

This section includes various prior research for the specified diseases that use ML approaches to categorize the disorders. Each literature review covers the type of dataset analyzed, the methodologies employed, and the evaluation metrics obtained.

### ❖ Chapter 3: Software Project Management Plans (SPMP)

It serves as the roadmap and governing framework for the entire project lifecycle. The SPMP outlines the project's objectives, scope, schedule, resource allocation, risk management, and quality assurance measures.

### ❖ Chapter 4: Software Requirement Specification (SRS)

SRS serves as a fundamental document outlining the functional and non-functional requirements of the system. This comprehensive guide details system behavior, constraints, and user interactions, providing a clear roadmap for development.

### ❖ Chapter 5: Software Design Specification (SDS)

It is a pivotal component of project organization, detailing the architectural and technical design of the system outlined in the SRS. It encompasses a high-level overview of system components, their interactions, and the data flow within the system.

### ❖ Chapter 6: Implementation

The implementation phase in project organization marks the transition from planning to execution. This critical stage involves empirical research for each chronic condition, as well as the datasets used and suggested ML approaches. Following that, it uses online and Saudi datasets in the tests and then presents the outcomes to better elucidate them. It also describes the website implementation in depth, including a full explanation of each interface.

### ❖ Chapter 7: Software Testing Plan (STP)

STP is a critical document within the project organization outlining the approach and strategies for testing the developed software, which comprises of test methodologies to evaluate the project's soundness and functionality.

#### ❖ Chapter 8: Conclusion

The conclusion section of the project organization documentation serves as a summary and final assessment of the project's organizational framework. It encapsulates the key strategies and methodologies employed to ensure efficient project execution.

## **Chapter 2: Background and Review of Literature**

This chapter involves a review of all previous literature conducted to predict and diagnose each disease using Artificial Intelligence techniques. Each literature review includes a summary of the methodology used, the type of dataset used, the assessment criteria, and the results that were achieved.

### **2.1 Epileptic Seizure**

#### **2.1.1 Introduction**

Epilepsy is a prevalent neurological disorder affecting millions worldwide, with around 0.654% of Saudis impacted [17]. It is characterized by recurrent, unprovoked seizures, posing physical and societal challenges. While approximately 70% of individuals with epilepsy can be effectively treated, the impact of the disorder extends beyond the affected individuals, affecting families and the healthcare system. This brief introduction highlights the significance of epilepsy in Saudi Arabia and the global context, emphasizing the need for awareness, diagnosis, and improved treatment options to enhance the quality of life for those affected.

#### **2.1.2 Literature Review**

This section discussed the research papers that were concerned about diagnosing Epileptic seizure using different AI techniques.

Chen et al. [18] developed an automated method for detecting and classifying epileptic seizures using electroencephalogram (EEG) signals. They employed a feature fusion and selection approach to extract mixed features such as Fuzzy Entropy (FuzzyEn), Sample Entropy (SampEn), Approximate Entropy (ApEn), and Standard Deviation (STD) from subbands obtained through Discrete Wavelet Transform (DWT) decomposition. Feature selection was performed using the RF algorithm, and the classification of epilepsy EEG signals was accomplished using CNN. The suggested approach was tested on benchmark datasets involving the EEG datasets from Bonn and the New Delhi datasets. The results showed high accuracy and performance, with the model achieving 99.9% accuracy for the Bonn datasets and 100% classification accuracy for the New Delhi datasets.

A study conducted by Ode et al. [19] addresses the critical need for early detection of focal epileptic seizures by leveraging alterations in heart rate variability (HRV) derived from electrocardiogram (ECG) data, which can be indicative of autonomic nervous system (ANS) disturbances occurring between 15- and 20-minutes preceding seizure onset. The major goal is to create a machine learning algorithm capable of real-time monitoring of R-R interval (RRI) data for seizure prediction using SA-AE algorithm. The application of the developed technique to clinical data indicates its effectiveness across a majority of patients but has a false positive (FP) rate of 0.85. Moreover, the dataset collection involved 39 individuals with

focal epilepsy who were hospitalized to Tokyo Medical and Dental University's (TMDU) Medical Hospital. Additionally, the SA-AE model had a sensitivity of 74%, a precision of 0.35, a false positive rate of 0.85 times/h, and an (AUC) of 0.97. Notably, 29 patients achieved 100% sensitivity, with eight of them experiencing no false positives.

A paper introduced by Nazari et al. [20] proposes a novel approach for epileptic seizure prediction, focusing on patients with late-onset seizures, where recording preictal signals proves challenging. The suggested method employs a convolutional neural network (CNN) and leverages few-shot learning, requiring minimal data for training. The method is evaluated on EEG data from three cases from the CHB-MIT dataset. The evaluation focuses on a 10-minute SPH, which is a seizure prediction horizon, and a 20-minute SOP, which is a seizure occurrence period, yielding an average sensitivity of 95.70% and a false prediction rate (FPR) of 0.057/h. While these results are promising, it is important to acknowledge limitations, including the small sample size and the need for validation across a more diverse patient population. Further research should focus on expanding the dataset and validating the approach across diverse patient populations to establish its robustness and applicability in clinical settings.

A study by Tawfik et al. [21], This paper aims to explore the use of deep learning and ML for epileptic seizure detection, comparing their effectiveness and trying to enhance existing detection methods. The EEG dataset from Bonn University is available on the UCI website which includes 400 individuals where 200 have epileptic and 200 do not. Epilepsy in individuals can be identified by analyzing EEG signals using various ML methods like SVM, LoR, KNN, DT, RF, XGboost, Catboost, and others. The EEG dataset performed very well with the CNN algorithm, achieving 99.2% accuracy, 99.3% specificity, and 98.7% sensitivity. The hybrid DNN (CNN with LSTM) achieved 98.7% accuracy.

The goal of the study proposed by Hilal et al. [22] was to create an intelligent model employing electroencephalogram (EEG) signals to identify and categorize epileptic seizures. The authors present a deep canonical sparse autoencoder-based epileptic seizure detection and classification (DCSAE-ESDC) model that includes two primary processes: feature selection and classification. utilizing coyote optimization algorithm (COA) for feature selection and krill herd algorithm (KHA) for tuning the model's parameters. A benchmark dataset for the detection of epileptic seizures from the UCI repository was used for the investigation. The analytical findings demonstrate that the DCSAE-ESDC approach surpasses existing techniques in binary and multi- classification, with an accuracy rating of 98.67% and 98.73%, respectively.

A study written by Jemal et al. [23], the author's aim of the study is to examine a comprehensible Deep-learning (DL) classifier to predict seizures by the EEG data. The publicly available dataset includes a recording of continuous multi-channel scalp EEG for 940 hours from 23 patients between the ages of 1.5 to 19 years at Boston Children's Hospital. The architecture of the DNN employs the algorithm Filter Bank Common Spatial Pattern

(FBCSP) for EEG data analysis. It includes temporal and depth-wise convolutions, batch normalization, and feature extraction for long continuous EEG data analysis. They reached an accuracy of 90.9% and compared to other studies they achieved the highest sensitivity which is 96.1%.

Ouichka et. al. [24] have conducted a study regarding the automatic epileptic seizure's prediction. The study focuses on improving the accuracy of epileptic seizure prediction and early detection using intracranial electroencephalogram (iEEG) datasets. Five deep learning models were presented for automated seizure prediction, together with Convolutional Neural Networks (CNNs) and transfer learning with ResNet50. The 3-CNN and 4-CNN models yielded the highest accuracy, according to the experimental data, achieving an impressive 95%. The dataset utilized in this study originates from the American Epilepsy Society for Seizure Prediction. This dataset encompasses intracranial EEG signals (iEEG) obtained from a diverse sample pool, consisting of five (5) dogs and two human patients.

A study authored by Usman et. al. [25], the main focus was to develop an epileptic seizure prediction method utilizing EEG monitoring, with a specific focus on accurately predicting the preictal state preceding seizure onset. The methodology comprises three key steps: firstly, employing empirical model decomposition for noise reduction and utilizing Generative Adversarial Networks to address class imbalance by generating preictal samples during EEG signal preprocessing. Secondly, automated feature extraction is performed using a three-layer CNN. Lastly, Long Short-Term Memory units are employed in order to distinguish between preictal and interictal states. The CHBMIT dataset, which contains scalp EEG signals from 22 subjects, is utilized. The proposed method achieves high sensitivity (93%) and specificity (92.5%) with an average anticipation time of 32 minutes for predicting seizure onset.

A study written by Almustafa [26] aimed to find the best classification algorithm for an epileptic seizure dataset to determine whether a seizure was present or not by using various classification techniques, as well as to assess the behavior of the algorithm when parameters are changed. The dataset used in this study consisted of 11,500 samples, each with 178 features, and was categorized into five classes based on different conditions during the recording of EEG signals such as open eyes, closed eyes, and various stages of epileptic seizures. The ML algorithms applied were KNN, Naïve Bayes, RF, DT, LoR, J48, and Stochastic Gradient Descent. As a result, the RF classifier outperformed all others with an accuracy of 97.08%.

A study by Alqaseer et al. [27], the authors present the critical need for accurate and timely detection of epileptic seizures, using an algorithm that leverages Discrete Cosine Transformation (DCT) type II to transform EEG signals into the frequency-domain, extracting energy features from 16 sub-bands. Data is segmented into non-overlapping 1-second frames, and based on Euclidean distance, the K-Nearest Neighbor (KNN) model is utilized to recognize those frames that are either ictal (seizure) or interictal (non-seizure).

Testing on 21 patients from the CHB-MIT dataset yields an impressive average F1-score of 93.12, with a low False-Positive Rate (FPR) average of 0.07.

A study by Savadkoohi et al. [28], the authors employed machine learning methods to detect brain electrical activities and patterns from an epileptic Electroencephalogram (EEG) that indicates an epileptic seizure is imminent. The dataset used a signal record from 5 healthy and 5 epileptic patient volunteers. To select the required feature from the dataset T-test and SFFS were applied. Additionally, the SVM and KNN algorithms are employed to differentiate between preprocessed EEG signals, with SVM exhibiting a slight performance advantage over KNN. SVM accuracy result is 100%, while KNN accuracy is 99.5%.

A study by Usman et al. [12], proposed a deep learning technique to predict epileptic seizures. Convolution neural networks (CNN) were applied as a feature extraction method, while Support vector machines (SVM) were utilized as the classification approach to differentiate between interictal and preictal states. The model is employed on a dataset of 24 subjects of scalp EEG. The result of the study was successful and accomplished a sensitivity of 92.7% and specificity 90.8%. Nonetheless, this model only tested for epileptic patients, it would improve the model to test it on non-epileptic patients that who experience these seizures.

LIU et al. [29] developed a research paper that aims to develop a prediction model for epileptic seizures using multi-view convolutional neural networks (CNNs). By leveraging the power of CNNs in processing spatial and temporal features. The researchers performed experiments on kaggle dataset to enhance the accuracy as well as the reliability of seizure prediction. This paper has been achieved with several key steps. Firstly, the researchers preprocessed the EEG signals, applying techniques such as data augmentation, PCA and FFT. Subsequently, they divided the preprocessed data into multiple views to capture different aspects of the EEG signal, possibly focusing on different EEG channels or spectral representations. Next, a multi-view CNN architecture was developed and trained on the dataset to predict epileptic seizures. Furthermore, on two subjects from the CHB-MIT scalp EEG dataset, the suggested model attained (AUC) values of 0.82 and 0.89.

A study written by Usman et. al. [30], the study introduces a robust machine learning model for predicting epileptic seizures, focusing on effective EEG signal preprocessing and feature extraction. The approach involves converting multiple EEG channels into a surrogate signal and applying empirical mode decomposition (EMD) to enhance signal quality. Various features, including approximate entropy, entropy, spectral, Hjorth parameters, and statistical moments, are extracted, providing valuable discriminative information. For distinguishing between preictal and interictal states, the SVM classifier is employed. Moreover, the results on the CHBMIT dataset, which contains scalp EEG signals from 22 subjects, demonstrate a notably improved true positive rate of 92.23% for detecting the preictal state, surpassing conventional methods.

Kornek et al. [31] proposed a research paper aiming to provide a solution for predicting epileptic seizures, which would enhance patient care and allow for timely intervention, using deep learning and big data techniques. The researchers used iEEG data from ten patients obtained from a seizure advisory system. The researchers first trained a deep learning classifier is used to discriminate between preictal (seizure-prone) and interictal (non-seizure) data. They next evaluated the classifier's performance using held-out iEEG data from all patients, comparing it to a random predictor. Additionally, the prediction system was fine-tuned to prioritize either sensitivity (the ability to detect seizures) or time in warning. The researchers proved that the prediction system may be deployed on an ultra-low-power neuromorphic chip for autonomous operation in a wearable device. The results revealed that the prediction system outperformed a random predictor by 42% for all patients, with a mean sensitivity of 69% and a mean time in warning of 27%.

After reviewing the studies above, it is noticeable that all of them used electroencephalogram (EEG). This data need testing and sensors, which contradicts the idea of this project. The project aims to use clinical data that requires the minimum number of examinations. Some of the common machine and deep learning algorithms used are CNN, KNN, and DT. Table 2 provides a summary of the results in these studies.

## 2.2 Osteoporosis

### 2.2.1 Introduction

Osteoporosis is a popular condition that causes a reduction in bone density, which exposes bones to fractures more quickly. In Saudi Arabia, 74% of women over 70 years old are osteoporotic, and 21% of men over 50 years old have Osteoporosis [32]. Because Osteoporosis may cause disability and death, society needs to give more consideration to it. To prevent any undesirable consequences.

### 2.2.2 Literature Review

Several studies were conducted for the aim of diagnosing osteoporosis using ML and deep learning techniques. Here in this section we will illustrate some of them and what are the results of these studies.

Albuquerque et al. [33] proposed a study that aims to develop an automated system for the early detection of osteoporosis. The researchers utilized a dataset consisting of bone mineral density measurements from many patients, reaching 505. Approximately 21.8% of individuals are healthy, while the majority, approximately 78.2%, have low bone mineral density and are affected by osteoporosis. The dataset was collected through DXA scans, which are considered the gold standard for detecting osteoporosis. The methodology involved training a machine learning model on these DXA scan measurements to classify individuals as either osteoporotic or non-osteoporotic. A 5-fold cross-validation is performed, and 20% of the dataset is used for testing. By incorporating electromagnetic wave analysis, Random Forest is the best model, with a sensitivity of 0.853, a specificity of 0.879, and an F1 score of 0.859. The findings emphasize age, BMI, and Osseus' signal attenuation

as the most critical criteria in the categorization of osteoporosis. The effectiveness of the RF model, combined with Osseus measurements, supports early osteoporosis screening, leading to a reduction in costs and improving patients' quality of life.

A study by Wu et al. [34], The aim of their study is to create predictive models using ML algorithms to serve as screening mechanisms for identifying osteoporosis type 2 diabetes (T2DM) patients. The data was gathered from 433 participants and employed nine categorical ML algorithms to choose features that are based on clinical and demographic variables. They evaluated the models using a range of metrics, optimized them through 5-fold cross-validation, and assessed feature importance using shapley additive explanations (SHAP). Furthermore, to identify distinct subgroups, they employed latent class analysis (LCA) within the dataset. ML algorithms had average precision scores ranging from 0.444 to 1.000. The final model, XGBoost, achieved AUROC values of 0.940 in training, 0.772 in 5-fold cross-validation, and 0.872 in testing. The study holds some limitations, such as the cross-sectional nature of this study, which introduces inherent limitations in conducting predictive analysis, and the study's sample size was inadequate to assess the model.

A study written by Wu and Park [35], This study aimed to build a model to predict the osteoporosis risk using ML in adults over 40 in the Ansan/Anseong cohort and to analyze its association with fractures in the HEXA cohort. In the Ansan/Anseong cohort, 8,842 participants had 109 factors, including demographics, measurements, genetics, nutrition, and lifestyle, manually chosen and integrated into the ML algorithm. The data was randomly split into 80% training with 7,074 samples and 20% testing with 1,768 samples. Using 109 standardized variables, they select LoR, SVM, XGBoost, DT, RF, KNN, and DNN for predicting metabolic status models to improve ROC area, accuracy, and K-fold performance in testing. XGBoost, DNN, and RF produced a highly accurate prediction model with an AUC of 0.86 in the ROC using 10, 15, and 20 features. The top two models for accuracy were XGBoost and RF with 15 features, achieving 0.902 and 0.903, respectively. The study has some limitations; for example, the cross-sectional design measured biomarkers once, but the large sample size helps BMD measured with a peripheral densitometer, though less precise than DEXA, remain suitable for clinical osteoporosis risk assessment.

Inui et al. [36] aimed in their study to develop an effective osteoporosis screening technique that did not require DXA scans. The authors employed a machine learning (ML) approach, utilizing patient body mass index (BMI), age, and blood test data as input features. The study included medical data from 2541 women who visited an osteoporosis clinic. Moreover, to identify and predict low BMD in patients, multiple ML models such as LR, DT, RF, Gradient Boosting Trees (GBT), and Light Gradient Boosting Machine were trained. Furthermore, among the trained models, lightGBM outperformed the other models with the highest accuracy of 83.4% and AUC of 96.1%, indicating its superior predictive performance for screening osteoporosis. The study has some limitations, including its focus on an outpatient clinic in Japan, its use of only female records, and the absence of imaging data like X-rays or CT scans, which could lead to different conclusions.

A study by Ma et al. [37], The goal of their study is to determine whether various alternative ML models could give a prediction that is better than LoR models and to choose the best model. A retrospective analysis was performed on 529 patients who had PKP at their institution between 6/2017 and 6/2020. They split the dataset into 75% for training and 25% for testing sets, and then built the ML models after 10 cross-validations. subsequently, all models were assessed using the testing set, and their performance was evaluated by measuring the AUC for each model. Except for DT, ML algorithms showed better performance than LoR, and among them, RF demonstrated the highest predictive capability. This study showed some limitations, such as the fact that retrospective studies can introduce selection and subjective bias, and the single-center study sample size remains relatively small.

Fasihi et al. [38] conducted a study that aims to explore the use of artificial intelligence in diagnosing osteoporosis based on risk factors derived from clinical data for both women and men. Additionally, the authors propose sports protocols that can potentially mitigate the negative impacts of osteoporosis. The dataset was obtained from three hospitals in Tehran, Iran, specifically from the scanning center for DXA. It's included clinical information from a total of 1224 individuals, both men and women. They used various machine learning models, such as RF, KNN, SVM, DT, AB, GB, ET, and MLP analysis, to predict osteoporosis and suggest suitable exercise programs for treatment. The dataset was separated into training (80%) and test (20%) sets, with the results evaluated on the test set. Based on the AUROC curve, the FR algorithm yielded the most accurate predictions for men, achieving an AUROC of 0.91. For women, the GB algorithm performed best with an AUROC of 0.95. In terms of exercise recommendation, the RF algorithm performed best in detecting exercise for healthy individuals as well as those with osteopenia and osteoporosis, with AUROCs of 0.96 and 0.99 in women and men, respectively.

A study by Dzierzak and Omietek [39], The objective of this study was to determine whether DCNNs could be used to develop a reliable approach for detecting osteoporosis based on CT scans of the spine. The study's research dataset contains CT scans of the L1 spongy tissue from 100 participants, of whom 50 have osteoporosis and 50 are healthy. This study used six DCNN architectures that are pre-trained (MobileNetV2, Xception, InceptionResNetV2, VGG16, VGG19, and ResNet50) with various topological depths. The VGG16 model, which has the lowest topological depths, showed the highest results with 95% accuracy, 96% true positive rate, and 94% true negative rate. The study limitation was having a small dataset with 400 images, and to overcome this, they used pre-trained DCNN models.

A study by Kwon et al. [40], the researchers aimed to develop a pre-screening method for early diagnosis of osteoporosis in postmenopausal Korean women using machine learning algorithms. The Korea National Health and Nutrition Examination Surveys were utilized to collect the data, which included 1431 postmenopausal women between the ages of 40 and 69. Feature selection techniques were used to identify 20 relevant features affecting

osteoporosis. Moreover, three machine learning algorithms, AdaBoost, Gradient Boosting (GBM), and RF, were trained on three different models: A, checkup features, B, survey features, and C, both checkup and survey features. The evaluation process was conducted based on accuracy and the AUROC. Finally, the results of the evaluation showed that Model C had the highest accuracy rates for AdaBoost, GBM, and RF with scores of 84.9%, 82.9%, and 83.2%, respectively, and AUROC scores of 92.1%, 90.8%, and 91.9%, respectively. This study shows some limitations, such as limited data from cross-sectional observational studies and focus on women aged 40–69, which makes generalizing its findings to all Korean populations difficult.

A study authored by Jang et al. [41], they proposed the OsPor-screen model that diagnoses osteoporosis from Chest Radiographs. The model was tested in internal and external datasets. The internal dataset is 13,026 chest radiographs and DXA, while the external dataset is 1089 chest radiographs, both of which were provided by Asan Medical Center. The model was trained using supervised learning techniques. The internal dataset test result was 82.40% accurate. On the other hand, the external dataset has an accuracy of 77.69%. The model was tested with a dataset that was collected from one center, and the performance of the model could be improved using electronic medical records (EMRs) clinical data. These limiting factors prevent the model from achieving better performance.

A research study was carried out by Jang et al. [14], the research paper introduces a deep neural network model (DNN) that predicts osteoporosis via simple hip radiography. The model was established to be a screening tool instead of using dual-energy X-ray absorptiometry (DXA). DXA is a tool to diagnose osteoporosis, but because of its high cost, it cannot be used widely. Consequently, the model proposed as a solution for this problem. The model used a dataset of a DXA for 1001 females that are over fifty-five years old. The accuracy that the model achieved was 81.2%. Yet, there was a limitation in this study, which was only including females in the study.

A study by Ou Yang et al. [42], the authors proposed a machine learning model that can predict the presence of osteoporosis in adults over fifty years old. The study aims to use the proposed model as a screening tool for osteoporosis in clinical practice. This will enable patients to be more alert regarding osteoporosis and prevent any serious complications due to it. The authors also desired to compare the results of this model with traditional methods that were used for predicting the disease. The dataset that was used in this study was from people who registered at the medical center in Taiwan. The dataset contains 5982 data points in total; 3053 of them are for men and 2929 are for women. Men were separated from women, and each had different results in accuracy. The algorithms used in the model are SVM, KNN, ANN, RF, and logistic regression (LoR). To identify the performance-best model Comparing the model's performance was done using AUROC. The model achieves 0.840, 0.821, 0.837, 0.843, and 0.827 for men and 0.807, 0.767, 0.781, 0.811, and 0.772 for women for SVM, KNN, ANN, RF, and LoF, respectively. Some of these algorithms show better performance than traditional methods. Thus, patients can gain an advantage by using the model to predict osteoporosis in the future.

The study achieved by Anam et al. [43], which aims to review the application of ML techniques in predicting osteoporosis for trabecular bone. In this review, the authors sought to analyze various datasets that have been used in studies related to osteoporosis prediction for trabecular bone using ML. The methodology provided a comprehensive summary of the ML techniques employed in osteoporosis prediction for trabecular bone. They reviewed several algorithms, such as SVM, RF, neural networks, and deep learning models. The authors emphasized the importance of feature selection and extraction techniques to enhance the accuracy of the predictions. Furthermore, they discussed the importance of cross-validation techniques to validate the performance of the models. The result reported notable achievements in predicting osteoporosis for trabecular bone using ML techniques. They highlighted that the developed models exhibited high accuracy rates when tested on the selected datasets. The review identified that certain algorithms, such as ensemble models and deep learning architectures, yielded superior results in comparison to traditional ML approaches.

The research paper by Yamamoto et al. [44] introduces a method to diagnose osteoporosis and classify it from hip radiographs. The authors used a deep learning algorithm which is CNN. Moreover, the authors wanted to test if adding clinical data to the image-based dataset they used would enhance the performance of the diagnosis and increase its accuracy. ResNet18, ResNet34, GoogleNet, EfficientNet b3, and EfficientNet b4 are the CNN models used to classify hip radiographs into osteoporosis and non-osteoporosis. Each model was assessed first on an image-based dataset containing 1131 hip radiograph images and then with the same dataset in addition to clinical data. GoogleNet and EfficientNet b3 had the highest accuracy of 0.8407 for the dataset without clinical data. With the clinical data added, there was an improvement in the accuracy for each model; the highest was EfficientNet b3 with 0.8850.

The aim of this research paper, which was achieved by Yu et al. [45], is to establish and apply an ANN model to aid in the diagnosis of osteoporosis using a preprocessed dataset of patients diagnosed with osteoporosis, which consists of X-ray images, clinical symptoms, and demographic information, as well as bone mineral density (BMD) measurements. The researchers applied the MLP-based ANN model to the diagnostic system of osteoporosis. The MLP architecture consists of an input layer, hidden layers, and an output layer where the final diagnostic prediction is provided. The ANN model was trained using a backpropagation algorithm, and the hyperparameters were optimized to achieve the highest possible accuracy. The ANN model achieved a high level of accuracy in diagnosing osteoporosis (90%). The researchers validated the model using a separate test set, further confirming its effectiveness in accurately classifying osteoporosis cases.

A study by Yoo et al. [46], this paper aimed to develop and validate machine learning models to accurately identify the risk of osteoporosis in postmenopausal women. The models were developed using medical records from postmenopausal Korean women acquired as part of

the Korea National Health and Nutrition Examination Surveys. Furthermore, prediction models were built using machine learning methods such as SVM, RF, LoR, and ANN. The efficacy of these machine learning models was compared to four traditional clinical decision tools: osteoporosis index of risk (OSIRIS), osteoporosis self-assessment tool (OST), simple computed osteoporosis risk estimation (SCORE), and osteoporosis risk assessment instrument (ORAI). Finally, compared to other approaches, SVM exhibited a much superior AUC of the receiver operating characteristic. SVM had an AUC of 82.7% and an accuracy of 76.7% for predicting osteoporosis risk. This study has some limitations due to its cross-sectional survey and difficulties in addressing drug effects as well as development for Korean women. It also has an imbalanced class distribution. Further research with large, diverse samples is required to validate the findings.

Several algorithms were utilized in the research paper, such as ANN, XGboost, and SVM. Most of these paper used an image-based dataset to diagnose osteoporosis. Specifically using DXA test that needs X-ray, which might not be accessible every time. The project goal is to overcome this cons by using only clinical data. A summarization of these papers is provided in table 3.

## 2.3 Sickle Cell Anemia

### 2.3.1 Introduction

One of the most frequent hematological illnesses is sickle cell disease (SCD). It is caused by a single gene mutation that results in the synthesis of defective hemoglobin, which can eventually lead to chronic manifestations that affect practically every organ in the body. The frequency of sickle cell anemia is high in Saudi Arabia. It affects more than 45,100 adults per 1,000,000 people, and an estimated 2400 children and adolescents per 1,000,000 are diagnosed with sickle cell disease. Given these numbers, it is critical to focus treatment and attention on this illness. [5] To treat sickle cell anemia, a collaborative effort of healthcare practitioners and researchers is needed to improve early diagnosis, treatments, awareness, and support systems, eventually lessening the effect, increasing the quality of life, and reducing the healthcare burden in Saudi Arabia.

### 2.3.2 Literature Review

Below are some research papers that employed AI to build models capable of diagnosing sickle cell anemia.

The research paper that was produced by Darrin et al. [47] aimed to explore the potential of convolutional and recurrent neural networks in classifying red cell dynamics, specifically in the context of sickle cell disease. To achieve this, the researchers utilized a carefully collected dataset of red cell images, capturing different morphological variations associated with the disease. The methodology involves extracting features from sequential red cell images using CNNs and then leveraging RNNs to capture temporal dependencies within the data. The researchers trained their model on a dataset consisting of red cell images from healthy

individuals and individuals with sickle cell disease. The results demonstrated the effectiveness of their approach, with the model achieving high accuracy in classifying red blood cells between the two groups with an accuracy of 97% and an F1-score of 0.94.

A study by Ayoade et al. [48] aimed to enhance sickle cell disease (SCD) prediction accuracy by comparing individual Machine Learning (ML) algorithms and their ensemble models, targeting elongated erythrocyte shape identification. Moreover, three ML algorithms—MLR, XGBoost, and RF—were evaluated, and ensemble models were created. Performance metrics including accuracy, sensitivity (ROC-AUC), and F1 score were employed to assess the models' efficacy. Additionally, the analysis was conducted using Python programming, and medical datasets were employed. While individual algorithms achieved good accuracies (MLR = 87%, XGBoost = 90%, RF = 93%), the hybrid models RF-MLR and RF-XGBoost outperformed with accuracies of 92% and an impressive 99%, respectively.

The research study achieved by Nguyen et al. [49] was aimed at investigating the association of interleukin-6 (IL6) and interleukin-8 (IL8) with Sickle Cell Anemia (SCA) patients and exploring the possibility of predicting their presence using ANN based on hemoglobin alleles and other hematological variables. The dataset for this study was collected using a cross-sectional study that involved 74 healthy individuals and 60 sickle anemia patients. The researchers utilized a deep learning model to analyze the data, train the models using supervised learning techniques, and evaluate their performance using various evaluation metrics. They discovered a non-linear association between hemoglobin alleles and IL6 and IL8 production in SCA patients. The model achieved an impressive accuracy of 90.9% and an r-squared value of 0.88, demonstrating its potential to aid in the development of specific treatments and diagnostics for those suffering from Sickle Cell Anemia and associated immune complications.

A paper written by Saputra et al. [50], the authors propose the development of an automated prediction model to assist doctors in distinguishing between four types of anemia. This is essential because diagnosing anemia is challenging due to its wide range of symptoms and diverse forms. This study involved 165 females and 25 males aged 15 to 41 who had been diagnosed with various types of anemia. This study model was created by the ELM algorithm. Afterward, its performance was assessed using a confusion matrix with a dataset of 190 samples to represent the four types. The model achieved high results with an accuracy rate of 99.21%, sensitivity of 98.44%, precision of 99.30%, and an F1 score of 98.84%.

A paper by Vohra et al. [51], the authors considered the diagnosis problem as a classification task with three classes: mild, moderate, and severe, in contrast to previous studies, which represented it as a binary classification problem. Patient data from the Eureka diagnostic center in Lucknow, India, available on the Mendeley Data Repository, was used with 400 samples. They applied six ML algorithms for multi-class classification on the dataset,

benchmarking their performance using both 10-fold cross-validation and hold-out methods. The results of the MLP network model showed the highest performance, achieving an accuracy of 99.35% on the SMOTE dataset.

Patgiri and Ganguly [52] conducted a comprehensive exploration of automatic disease diagnosis through image processing techniques, with a specific focus on sickle cell anemia (SCA) detection from microscopic blood images. In this study, the authors proposed an innovative approach for SCA detection, employing a segmentation method that combines local adaptive thresholding and an active contour-based algorithm. Furthermore, supervised classifiers, including SVM and ANN, were utilized to identify sickle cells based on the geometric features of RBCs. Their findings revealed that the SVM classifier outperformed the ANN, achieving an impressive accuracy rate of 99.2%, while the ANN, trained with resilient backpropagation and featuring ten hidden neurons, demonstrated promising accuracy at 99%. Additionally, the review also delved into various segmentation methods used in medical image analysis, encompassing thresholding, edge-based, region-based, and clustering-based techniques.

A paper written by Shahzad et al. [53], the authors aimed to predict the severity of anemia by extracting essential morphological features to identify anemic pictures using a 3-tier deep convolutional fused network (3-TierDCFNet). The dataset includes 11,500 pictures with about 750,000 red blood cell elements. Out of these pictures, 5,750 are considered normal, and the other 5,750 show signs of anemia. The model consists of 2 modules: Module-I classifies pictures as Anemic or healthy, and Module-II detects Mild or Chronic anemia. Validation reduces inappropriate feature selection after each module's training. Evaluation metrics include specificity, recall, F1-Score, and accuracy. The Tier-III model achieved the highest result with 89.29% accuracy, 95.96% recall, 95.87% F1-Score, and 96.34% specificity.

This study was conducted by Srivastave et al. [54] and aimed to diagnose sickle cell anemia using the AutoML approach on UV-VIS absorbance spectroscopy data. The researchers utilized a dataset consisting of UV-VIS absorbance spectra collected from blood samples of individuals with and without sickle cell anemia. The dataset was collected from a diverse set of patients representing different age groups and ethnicities. The methodology involved the use of AutoML, a machine learning technique that automates the process of model selection and hyperparameter tuning. The dataset was divided into 70% for training and 30% for testing. The researchers trained various machine learning models on the dataset and evaluated their performance using cross-validation. This approach accurately identifies the presence of sickle hemoglobin with a sensitivity of 100% and a specificity of 93.84%. This study demonstrates the potential of AutoML in the field of medical diagnostics, offering an effective and efficient approach for diagnosing sickle cell anemia based on UV-VIS absorbance spectroscopy data.

A study was done by Yeruva et al. [55]. The paper introduced a deep neural network used to identify red blood cells and classify them into three categories: normal, sickle cell, and thalassemia. To enhance the accuracy of distinguishing sickle cells from other types of cells, the authors used an approach called Multi-Layer Perceptron (MLP) that provides more efficiency than the rest of the algorithms that have been used in the study. In addition to other machine learning approaches such as SVM, RF, KNN, Logistic Regression, and DT, the dataset used was received from the Thalassemia and Sickle Cell Society (TSCS), which has 1387 records. The results of the study confirm that using the MLP approach will enhance the accuracy of recognizing sickle cells. MLP has an accuracy of 99%, which is the best, followed by RF with 97%.

A paper written by Abdulkarim, H.A. et al. [56], the focus was to construct a deep-learning AlexNet model for red blood cell classification in sickle cell anemia (SCA) patients. The researchers used a dataset of over 9,000 single red blood cell (RBC) images from 130 SCA patients, with 750 cells in each class. Furthermore, the algorithm was developed in two stages: the automation of RBC extraction to identify the region of interest (ROI) in the blood smear image and the use of a deep-learning AlexNet model for classifying and predicting the presence of abnormalities in SCA patients. Finally, the results demonstrate that the proposed framework achieved a classification prediction accuracy of 95.92% for identifying abnormalities in the RBCs of SCA patients using a deep-learning AlexNet model.

Research conducted by Alzubaidi et al. [57] aims to develop a classification model for red blood cells to distinguish sickle cells from other kinds of cells. Besides, the authors wanted to highlight the lack of red blood cells in people infected with sickle anemia. The model used lightweight deep learning with transfer learning techniques to resolve the problem of lacking data. In addition to data augmentation to expand the quantity of data to be trained, the model was trained and tested with three different datasets and different scenarios. The final results confirm that the model has an accuracy of 99.54%. Additionally, adding an SVM classifier to the model enhanced the performance slightly, achieving an accuracy of 99.98%.

A study by Mohammed et al. [58], the researchers investigated the potential of ML techniques in predicting the appearance of organ failure in adult SCD patients admitted to intensive care units (ICUs). A dataset comprising continuous physiological data was collected from 134 adult subjects at Methodist Le Bonheur Hospital in Memphis, Tennessee. Moreover, four machine learning algorithms were used to build classification models: multi-layer perceptron (MLP), RF, LoR, and SVM. Finally, the RF model accurately predicted organ failure up to six hours before its onset, with an accuracy of 94.57%, sensitivity of 90.24%, and specificity of 98.9%.

The paper by Alzubaidi et al. [59] discussed using a deep convolutional neural network to categorize red blood cells (RBC) into three categories: normal cells, sickle cells, and other cells that have another kind of disease. The proposed model overcomes the problem of classifying RBC using traditional methods. It needs more time and effort because of the complexity of RBC shapes. To train and test the model, 340 images of RBC were used as a

dataset for this model, collected from Wadsworth Center data. With the help of error-correcting output codes (ECOC), the model's accuracy is 92.06%. The result shows the effectiveness of this model in classifying the RBC, so less time is needed to diagnose sickle anemia.

A study authored by Xu et al. [60], the authors present a comprehensive overview of their research on sickle cell disease (SCD) and the development of an automated RBC pattern classification framework. The authors' framework is divided into three phases: the automatic RBC extraction method, the RBC patch-size normalization method, and deep convolutional neural network (CNN) classification. Furthermore, the study includes experiments on microscope image datasets from eight SCD individuals from two different hospitals. Additionally, the authors highlight the importance of shape factor quantification and discuss the potential for clinical applications in SCD management.

The purpose of Alkrimi et al. [61] research is to develop an automated method for classifying red blood cells (RBCs) as normal or abnormal using Support Vector Machine (SVM) classification. The authors highlight the significance of medical imaging in the diagnosis of blood disorders as well as the function of RBC shape in clinical diagnosis. Furthermore, image processing techniques such as segmentation and mean filtering are used in this study to extract geometric, texture, and color properties from RBC images using photo imaging microscopy. To distinguish between normal and abnormal RBCs, the SVM is used as the classifier. Moreover, the dataset utilized comprises 1000 images of RBCs obtained from the Department of Hematology at Serdang Hospital in Malaysia. Finally, the experimental findings show that the proposed classifier algorithm achieves high accuracy rates, with an accuracy of 99.9%.

The reviews revealed that most studies used image-based dataset of blood cells to differentiate between normal and sickle cells. Several classifiers were used for this purpose including XGboost, RF, and ANN. This project contributes to the previous paper by using only clinical data. Table 4 summarize the research and their finding.

N	Author/s	Year	Title	Technique/s	Results
[18]	Wenna Chen, Yixing Wang, Yuhao Ren, Hongwei Jiang, Ganqin Du, Jincan Zhang, and Jinghua Li	2023	An automated detection of epileptic seizures EEG using CNN classifier based on feature fusion with high accuracy.	Feature fusion and selection approach using CNN and RF algorithms with Fuzzy Entropy (FuzzyEn), Sample Entropy (SampEn), Approximate Entropy (ApEn), and Standard Deviation (STD).	The suggested approach tested on two different datasets achieving a high accuracy and performance, with the model achieving 99.9% accuracy for the Bonn datasets and 100% accuracy for the New Delhi datasets.
[19]	Rikumo Ode, Koichi Fujiwara, Miho Miyajima, Toshikata Yamakawa, Manabu Kano, et al.	2022	Development of an epileptic seizure prediction algorithm using R-R intervals with self-attentive autoencoder.	Self-attentive autoencoder (SA-AE), Anomaly detection framework for raw RRI data.	sensitivity, precision, FP rate, and Area Under Curve (AUC) were 74%, 0.35, 0.85 times/h, and 0.97, respectively.
[20]	Jamal Nazari, Ali Motie Nasrabadi, Mohammad Bagher Menhaj, and Somayeh Raiesdana	2022	Epilepsy seizure prediction with few-shot learning method.	CNN, few-shot learning method to enhance accuracy.	Sensitivities of 98.52% with low false prediction rates (FPRs) of 0.045/h
[21]	Mohammed Tawfik, Ezzaldden Mahyoub, Zeyad A. T. Ahmed, Nasser M. Al-Zidi and Sunil Nimbhore	2022	Classification of epileptic seizure using machine learning and deep learning based on electroencephalography (EEG)	Machine learning: SVM, Decision Tree, Random Forest, Quadratic Discriminant Analysis, KNN, XGboost, Catboost. Deep learning: MPL classifier, ANN, LSTM, CNN, CNN-LSTM.	The EEG dataset performed very well with the CNN algorithm, achieving 99.2% accuracy, 99.3% specificity, and 98.7% sensitivity. The hybrid deep neural network (CNN with LSTM) achieved 98.7% accuracy.
[22]	Anwer Mustafa Hilal, Amani Abdulrahman Albraikan, Sami Dhahbi, Mohamed K. Nour, Abdullah Mohamed,	2022	Intelligent Epileptic Seizure Detection and Classification Model Using Optimal Deep Canonical Sparse Autoencoder.	DCSAE-ESDC algorithm utilizing COA for feature selection and KHA for tuning the model's parameters.	The results of the analysis demonstrate that the DCSAE-ESDC technique outperforms existing techniques with a maximum accuracy of 98.67% and

	Abdelwahed Motwakel, Abu Sarwar Zamani, and Mohammed Rizwanullah				98.73% in binary and multi-classification, respectively.
[23]	Imene Jemal, Neila Mezghani, Lina Abou-Abbas and Amar Mitiche	2022	An Interpretable Deep Learning Classifier for Epileptic Seizure Prediction Using EEG Data	Filter Bank Common Spatial Pattern (FBCSP).	They reached an accuracy of 90.9% and compared to other studies they achieved the highest sensitivity which is 96.1%.
[24]	Omaima Ouichka, Amira Echtioui, Habib Hamam.	2022	Deep Learning Models for Predicting Epileptic Seizures Using iEEG Signals.	CNN model, 2-CNN, 3-CNN, 4-CNN, and transfer learning with ResNet50.	3-CNN and 4-CNN gave the best results. They both achieve an accuracy value of 95%.
[25]	Syed Muhammad Usman, Shehzad Khalid, Zafar Bashir.	2021	Epileptic seizure prediction using scalp electroencephalogram signals.	preprocessing with EMD and Generative Adversarial Networks, feature extraction with CNNs, and classification using Long Short Term Memory units.	The proposed method achieved 93% sensitivity, 92.5% specificity, and an average anticipation time of 32 minutes, outperforming recent seizure prediction methods.
[26]	Khaled Mohamad Almustafa	2020	Classification of epileptic seizure dataset using different machine learning algorithms	KNN, Naïve Bayes, RF, DT, LR, J48, and Stochastic Gradient Descent.	RF classifier outperformed all others with an accuracy of 97.08%.
[27]	Mahmood Alqaseer, Ammar Ibrahim Shihab, and keel Abdulkareem Farhan.	2020	Epileptic Seizures Detection Using DCT-II and KNN Classifier in Long-Term EEG Signals.	Automatic channel selection and K-Nearest Neighbour (KNN) model.	F1-score of 93.12 and a False-Positive Rate (FPR) average of 0.07.
[28]	Marzieh Savadkoohi, Timothy Oladduni and Lara Thompson	2020	A Machine Learning Approach to Epileptic Seizure Prediction using Electroencephalogram (EEG) Signal.	Feature selection by T-test and Sequential Forward Floating Selection (SFFS), classifying the signals using SVM and KNN	SVM classifier achieved the highest accuracy of 100%, while KNN classifier got 99.5%

[12]	Syed Muhammad Usman, Shehzad Khalid and Muhammad Haseeb Aslam	2020	Epileptic Seizures Prediction Using Deep Learning Techniques	Feature extraction using CNN and classifying with SVM	Sensitivity of 92.7% and specificity 90.8%.
[29]	Chien-Liang Liu, (Member, IEEE), Bin Xiao, Wen-Hoar Hsiao , and Yincent S.Tseng	2019	Epileptic Seizure Prediction with Multi-View Convolutional Neural Networks	Data augmentation, PCA, FFT, feature extraction, auto regression, correlation entropy, Multi-view CNN, Lasso GLM, Random Forest, Bagged SVM, SVM, CNN, Class weighted SVM and Linear SVM.	Average area under the curve (AUCs) of 0.82 and 0.89 on two subjects of the CHB-MIT scalp EEG dataset.
[30]	Syed Muhammad Usman,Muhammad Usman ,and Simon Fong	2017	Epileptic Seizures Prediction Using Machine Learning Methods.	Preprocessing with EMD and feature extraction, including entropy, Hjorth parameters, spectral moments, and statistical moments. SVM is used for classification.	Average sensitivity 92.23% and specificity 93.38%.
[31]	Isabell Kiral-Kornek, Subhrajit Roy, Ewan Nurse, Benjamin Mashford, Philippa Karoly, Thomas Carroll, Daniel Payne, Susmita Saha, Steven Baldassano, Terence O'Brien, David Grayden, Mark Cook, Dean Freestone, Stefan Harrer.	2017	Epileptic Seizure Prediction Using Big Data and Deep Learning: Toward a Mobile System	Deep learning classifier, Deep neural network and benchmarking.	mean sensitivity of 69% and a mean time in warning of 27%, surpassing the performance of a random predictor by 42% for all patients.

Table 2 Epileptic Seizure Literature Review Summary

N	Author/s	Year	Title	Technique/s	Results
[33]	Gabriela A. Albuquerque, Dionísio D.A. Carvalho, Agnaldo S. Cruz <sup>1,2</sup> , João P. Q. Santos, Guilherme M. Machado, Ignácio S. Gendriz, Felipe R. S. Fernandes, Ingridy M. P. Barbalho, Marquiony M. Santos, César A. D. Teixeira, Jorge M. O. Henrique, Paulo Gil, Adrião D. D. Neto, Antonio L. P. S. Campos, Josivan G. Lima, Jailton C. Paiva, Antonio H. F. Morais, Thaisa Santos Lima & Ricardo A. M. Valentim.	2023	Osteoporosis screening using machine learning and electromagnetic waves	Random forest, XG boost, Gradient boosting, Extra trees, Hist grad boosting, Bagging, Ada boost, Gaussian process, XG boost RF, Decision tree, Linear discriminant, Logistic regression, K neighbors, Linear SVC, Quadratic discriminant, SVC, Stochastic gradient desc, Extra tree, Gaussian naive bayes, 5-fold cross-validation, confusion matrix and Classification reports.	The best model, Random Forest, achieves a sensitivity of 0.853, specificity of 0.879, and an F1 score of 0.859.
[34]	Xuelun Wu, Furui Zhai, Ailing Chang, Jing Wei, Yanan Guo and Jincheng Zhang	2023	Development of Machine Learning Models for Predicting Osteoporosis in Patients with Type 2 Diabetes Mellitus—A Preliminary Study	XGBoost, LR, LightGBM, RF, AdaBoost, GNB, MLP, SVM and KNN	Machine learning algorithms had AP scores from 0.444 to 1.000. The final model, XGBoost, achieved AUROC values of 0.940 in training, 0.772 in 5-fold cross-validation, and 0.872 in testing.
[35]	Xuangao Wu and Sunmin Park	2023	A Prediction Model for Osteoporosis Risk Using a Machine-Learning Approach	LR, XGBoost, DT, KNN, SVM, RF and DNN	XGBoost, deep neural network, and random forest produced a highly accurate prediction model with an AUC of 0.86 in the ROC using 10,

			and Its Validation in a Large Cohort		15, and 20 features. The top two models for accuracy were XGBoost and random forest with 15 features, achieving 0.902 and 0.903, respectively.
[36]	Atsuyuki Inui, Hanako Nishimoto, Yutaka Mifune, Tomoya Yoshikawa, Issei Shinohara, Takahiro Furukawa, Tatsuo Kato, Shuya Tanaka, Masaya Kusunose, and Ryosuke Kuroda.	2023	Screening for osteoporosis from blood test data in elderly women using a machine learning approach	LR, DT, RF, gradient boosting trees, and lightGBM.	Among the trained models, lightGBM outperformed the other models with the highest accuracy of 83.4% and AUC of 96.1%.
[37]	Yiming Ma, Qi Lu, Feng Yuan and Hongliang Chen	2023	Comparison of the effectiveness of different machine learning algorithms in predicting new fractures after PKP for osteoporotic vertebral compression fractures	DT, RF, SVM, GBM, NNET, regularized discriminant analysis (RDA) and LR	Except for decision tree, machine learning algorithms showed better performance than logistic regression, and among them, random forest demonstrated the highest predictive capability.
[38]	Leila Fasihi, BakhtyarTartibian, Rasoul Eslami2 & Hossein Fasihi	2022	Artificial intelligence used to diagnose osteoporosis from risk factors in clinical data and proposing sports protocols	DT, RF, KNN, SVM, GB, Extra trees, AB, ANN, MLP, hyperparameter tuning and AURC	GB algorithm performed best with an AUROC of 0.95. In terms of exercise recommendation, the RF algorithm exhibited the highest performance in diagnosing exercise for healthy individuals, as well as those with osteopenia and osteoporosis, with AUROCs of 0.96 and 0.99 in women and men, respectively.
[39]	Róża Dzierżak and Zbigniew Omiotek	2022	Application of Deep Convolutional Neural Networks in the Diagnosis of Osteoporosis	six DCNN architectures which are pre-trained (MobileNetV2, Xception, InceptionResNetV2, VGG16, VGG19 and ResNet50)	VGG16 model which have the lowest topological depths showed the highest results with 95% ACC, 96% TPR and 94% TNR.

[40]	Younghn Kwon, Juyeon Lee, Joo Hee Park, Yoo Mee Kim, Se Hwa Kim, Young Jun Won, and Hyung-Yong Kim	2022	Osteoporosis pre-screening using ensemble machine learning in postmenopausal Korean women	RF, AdaBoost, and GBM, were trained on three different models: A, checkup features, B, survey features, and C, both checkup and survey features	Model C achieved the highest accuracy rates for AdaBoost, GBM, and RF with scores of 84.9%, 82.9%, and 83.2% respectively, and AUROC scores of 92.1%, 90.8%, and 91.9%.
[41]	Miso Jang, Mingyu Kim, Sung Jin Bae, Seung Hun Lee, Jung-Min Koh, and Namkug Kim	2021	Opportunistic Osteoporosis Screening Using Chest Radiographs With Deep Learning: Development and External Validation With a Cohort Dataset	OsPor-screen model that was tested in two dataset	82.40% accuracy for the internal dataset and 77.69% for the external.
[14]	Ryoungwoo Jang, Jae Ho Choi, Namkug Kim, Jae Suk Chang, Pil WhanYoon and Chul-Ho Kim	2021	Prediction of osteoporosis from simple hip radiography using deep learning algorithm	DNN	The model got accuracy of 81.2%, , sensitivity of 91.1%, and specificity of 68.9%.
[42]	Wen-Yu Ou Yang, Cheng-Chien Lai, Meng-Ting Tsou and Lee-Ching Hwang	2021	Development of Machine Learning Models for Prediction of Osteoporosis from Clinical Health Examination Data	SVM, KNN, RF, LoR and ANN AUROC was used to compare the performance of each model.	RF achieved the highest value of AUROC for both men and women datasets
[43]	Marrium Anam, Vasaki Ponnusamy, Muzammil Hussain, Muhammad Waqas Nadeem, Mazhar Javed, Hock Guan Goh and Sadia Qadeer	2021	Osteoporosis Prediction for Trabecular Bone using Machine Learning: A Review	Two Sequences MRI, Trabecular Biomechanical Strength, Bone Strength MRI, Radial Bias Function, Otsu Threshold, FDT based techniques, 3D micro–MR Imaging	certain algorithms, such as ensemble models and deep learning architectures, yielded superior results in comparison to traditional machine learning approaches
[44]	Norio Yamamoto, Shintaro Sukegawa, Akira Kitamura, Ryosuke Goto, Tomoyuki Noda, Keisuke Nakano,	2020	Deep Learning for Osteoporosis Classification Using Hip Radiographs and Patient Clinical Covariates	5 CNN models which are ResNet18, ResNet34, GoogleNet, EfficientNet b3 and EfficientNet b4. The model tested in two dataset one was image based only and the other one was with clinical data	The model with highest accuracy was EfficientNet b3 with 0.8850 for the dataset with clinical data.

	Kiyofumi Takabatake, Hotaka Kawai, Hitoshi Nagatsuka, Keisuke Kawasaki, Yoshihiko Furuki and Toshifumi Ozaki				
[45]	Xinghu Yu, Chao Y, Liangbi Xiang	2016	Application of artificial neural network in the diagnostic system of osteoporosis	Multilayer perceptron (MLP)-based ANN, backpropagation and the hyperparameters	Achieved a high level of accuracy in diagnosing osteoporosis with 90%
[46]	Tae Keun Yoo, Sung Kean Kim, Deok Won Kim, Joon Yul Choi, Wan Hyung Lee, Ein Oh, and Eun-Cheol Park	2013	Osteoporosis risk prediction for bone mineral density assessment of postmenopausal women using machine learning	SVM, RF, ANN, LR	SVM achieved the highest performance for predicting osteoporosis risk with 76.7% accuracy, 82.7% AUC.

Table 3 Osteoporosis Literature Review Summary

N	Author/s	Year	Title	Technique/s	Results
[47]	Maxime Darrin, Ashwin Samudre, Maxime Sahun, ScottAtwell, Catherine Badens, Anne Charrier, Emmanuèle Helfer, AnnieViallat, Vincent Cohen-Addad & Sophie Gifard-Roisin	2023	Classification of red cell dynamics with convolutional and recurrent neural networks: a sickle cell disease case study	. Uniform downsampling, Similarity downsampling, data cleaning, Cell characterization, hyperparameters tuning, confusion matrix, CNN and CRNN.	. high accuracy in classifying red blood cells between the two groups with accuracy of 97% and an F1-score of 0.94
[48]	Oluwafisayo Ayoade, Tinuke Oladele, Lucky Imoize, and Jerome Adeloye	2023	An Ensemble Models for the Prediction of Sickle Cell Disease from Erythrocytes Smears.	Python programming language, MLR, XGBoost, RF	MLR=87%, XGBoost=90%, and RF=93%.
[49]	Dylan Nguyen, Lukas Abraham, and Ashesh Amatya	2023	Application of Deep Learning Models into the Prediction of Interleukin-6 and -8 Cytokines in Sickle Cell Anemia Patients	Independent sample t-test, Chi-square test, Mann-Whitney test ANN, Backpropagation Neural Network and Gradient Descent	achieved an impressive accuracy of 90.9% and an r-squared value of 0.88
[50]	Dimas Chaerul Ekty Saputra, Khamron Sunat and Tri Ratnaningsih	2023	A New Artificial Intelligence Approach Using Extreme Learning Machine as the Potentially Effective Model to Predict and Analyze the Diagnosis of Anemia	ELM	The model achieved high results with a 99.21% accuracy rate, 98.44% sensitivity, 99.30% precision, and an F1 score reaching 98.84%.

[51]	Rajan Vohra, Abir Hussain, Anil Kumar Dudyala, Jankisharan Pahareeya, Wasiq Khan	2022	Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting	DT, LR, MLP, NB, RF and SVM	The results of the MLP model showed the highest performance, achieving an accuracy of 99.35% on the SMOTE dataset.
[52]	Chayashree Patgiri and Amrita Ganguly	2022	Machine Learning Techniques for Automatic Detection of Sickle Cell Anemia using Adaptive Thresholding and Contour-based Segmentation Method	Artificial Neural Network (ANN) and Support Vector Machine (SVM).	ANN got the highest accuracy of 99%
[53]	Muhammad Shahzad, Arif Iqbal Umar, Syed Hamad Shirazi, Zakir Khan, Asfandyar Khan, Muhammad Assam, Abdullah Mohamed and El-Awady Attia	2022	Identification of Anemia and Its Severity Level in a Peripheral Blood Smear Using 3-Tier Deep Neural Network	3-TierDCFNet	The Tier-III model achieved the highest result with 89.29% accuracy, 95.96% recall, 95.87% F1-Score, and 96.34% specificity.
[54]	Sarthak Srivastava, Radhika N. K., Rajesh Srinivasan, Nishanth K M Nambison, and Sai Siva Gorthi	2021	diagnosis of sickle cell anemia using AutoML on UV-VIS absorbance spectroscopy data	spectroscopy-based test, High-performance liquid chromatography (HPLC), feature selection, AutoML, hyperparameter tuning, Auto-sklearn, TPOT, Hyperopt-Sklearn and Auto-Keras	accurately identifies the presence of sickle hemoglobin with a sensitivity of 100% and a specificity of 93.84%.

[55]	Sagar Yeruva, M. Sharada Varalakshmi, B. Pavan Gowtham, Y. Hari Chandana and PESN. Krishna Prasad	2021	Identification of Sickle Cell Anemia Using Deep Neural Networks	MLP, SVM, RF,KNN, Logistic Regression and DT	MLP got the highest accuracy of 99%
[56]	Hajara Abdulkarim Aliyu, Mohd Azhar Abdul Razak, Rubita Sudirman, and Norhafizah Ramli	2020	A deep learning AlexNet model for classification of red blood cells in sickle cell anemia.	deep-learning AlexNet model	Using deep-learning AlexNet model, the proposed framework achieved a classification prediction accuracy of 95.92% for identifying abnormalities in red blood cells of sickle cell anemia patients.
[57]	Laith Alzubaidi, Mohammed A. Fadhel, Omran Al-Shamma, Jinglan Zhang and Ye Duan	2020	Deep Learning Models for Classification of Red Blood Cells in Microscopy Images to Aid in Sickle Cell Anemia Diagnosis	lightweight deep learning and data segmentations, in addition to 3 different datasets. SVM classifier was added to the model to enhance the performance.	The model alone got accuracy of 99.54%, with the addition of SVM it increased slightly to be 99.98%
[58]	Akram Mohammed, Pradeep Podila, Robert Davis, Kenneth Ataga, Jane Hankins and Rishikesan Kamaleswaran	2019	Machine learning predicts early-onset acute organ failure in critically ill patients with sickle cell disease.	Multi-layer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR).	The RF model accurately predicted organ failure up to six hours prior to its onset, achieving an accuracy of 94.57%, with a sensitivity of 90.24% and specificity of 98.9%.
[59]	Laith Alzubaidi, Omran Al- Shamma, Mohammed A. Fadhel, Laith	2019	Classification of Red Blood Cells in Sickle Cell Anemia Using Deep Convolutional Neural Network	CNN and ECOC	The CNN model with ECOC achieved accuracy of 92.06%

	Farhan and Jinglan Zhang <sup>5</sup>				
[60]	Mengjia Xu, Dimitrios P. Papageorgiou, and Sabia Abidi	2017	A deep convolutional neural network for classification of red blood cells in sickle cell anemia	The authors employed deep convolutional neural networks CNNs.	The mean evaluation accuracy for classification was 87.50%
[61]	Jameela Ali Akrimi, Azizah Suliman, Loay E. George and Abdul Rahim Ahmad	2014	Classification red blood cells using support Vector Machine', Proceedings of the 6th International Conference on Information Technology and Multimedia.	Support Vector Machine (SVM).	Support Vector Machine achieves high accuracy rates, with an accuracy of 99.9%.

Table 4 Sickle Cell Anemia Literature Review Summary

## **Chapter 3: Software Project Management Plans (SPMP)**

This chapter is divided into four major parts, the first being the project overview, which includes the system's purpose, scope, objectives, assumptions, constraints, and risks. In addition, it consists of the project's deliverables, budget summary, and plan evolution. The second part of this chapter covers the project organization, where the internal structure, external structure, roles, and responsibilities are identified. The third part of this chapter includes the managerial process plan, which defines the startup plan, work plan, project tracking plan, reporting, and closeout plan. Following this, plans for the technical process are presented, encompassing the process model, utilization of tools, methods, and techniques. Moreover, it includes product acceptance and infrastructure.

### **Project Overview**

This part of the document encompasses the core purpose and objectives of the system, as well as its defined scope, potential risks, underlying assumptions, and constraints. Additionally, it addresses the project's anticipated deliverables, budgetary considerations, scheduling, and outlines the plan for its evolution.

#### **3.1.1 Purpose, Scope, and Objectives**

##### **3.1.1.1 Purpose**

Chronic diseases present a significant threat to human life, leading to higher mortality rates and disabilities. The resemblance in symptoms between chronic diseases and other ailments often leads to inaccurate diagnoses. Thus, it is imperative to employ advanced technologies in treating chronic diseases to enhance the overall quality of life for the population. This project is centered on the preemptive diagnosis of Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia in Saudi Arabia using Machine Learning (ML) and data mining techniques, aiming to identify these conditions before their symptoms manifest. Introducing such a method could optimize supportive care, thereby reducing the potential risks associated with late detection of the targeted chronic diseases.

##### **3.1.1.2 Scope**

In this project, we will utilize online datasets for epileptic seizure, osteoporosis, and sickle cell anemia. These datasets will serve as the foundational information for our research and analysis. Leveraging this comprehensive resource, we aim to conduct in-depth studies and implement data-driven methodologies to address specific objectives within our project.

##### **3.1.1.3 Objectives**

The outlined goals for the project implementation include:

- Building upon the progress made in prior stages of the project.
- Enabling early diagnosis of Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia through a dedicated website.
- Investigating Saudi Arabian datasets for the purpose of constructing predictive models.

- Leveraging the potential of data mining techniques and Machine Learning in the realm of medicine.
- Attaining optimal diagnostic outcomes for the specified diseases by employing appropriate preprocessing and Machine Learning methods on the provided datasets.

### **3.1.2 Assumptions, Constraints, and Risks**

#### **3.1.2.1 Assumptions**

The project implementation will be guided by the following underlying assumptions:

- Data will be sourced from selected Saudi hospitals based on availability.
- All team members are proficient in Python programming for applying data mining techniques.
- Team members possess expertise in data pre-processing and constructing ML models.
- The project will adhere to the established timeline without any delays.
- Team members are equipped to tackle obstacles and handle unforeseen challenges.
- Diverse ML techniques will be explored to identify the most effective ones.
- Various techniques will be investigated to employ the best ML approaches.
- Regular weekly meetings will be held to review progress and outcomes.
- Essential development tools are accessible for the project's execution.

#### **3.1.2.2 Constraints**

The project implementation will be bound by the following constraints:

- The project is restricted to a specific number of participants.
- Access to available datasets is limited and presents a challenge.
- The project must adhere to the designated deadline for completion.
- The scope of the project is confined to straightforward clinical tabular data.

#### **3.1.2.3 Risks**

Anticipated risks that could potentially impede the project's progress include:

- Potential delays in acquiring datasets from selected hospitals in Saudi Arabia.
- Receiving a dataset that is unsuitable for use.
- Risk of losing crucial files during work.
- Possibility of failing to meet designated submission deadlines.
- Balancing work on this project alongside other course-related projects may pose a challenge for team members.
- There's a risk of team members not fully comprehending the project requirements, which could potentially slow down the pace of work.

### 3.1.3 Project Deliverables

Table 5 shows the deliverables for the project.

<b>Deliverable</b>	<b>To whom</b>	<b>Delivery Media</b>	<b>Date</b>
<b>Project Proposal</b>	Dr. S.O. Olatunji (Aadam) Mr. Aftab khan.	Softcopy	20 <sup>st</sup> August 2023
<b>Literature Review</b>	Dr. S.O. Olatunji (Aadam) Mr. Aftab khan.	Softcopy	23 <sup>th</sup> September 2023
<b>Submit Software Project Management Plan (SPMP)</b>	Dr. S.O. Olatunji (Aadam), Mr. Aftab khan.	Softcopy	11 <sup>st</sup> October 2023
<b>Mid Semester Report</b>	Dr. S.O. Olatunji (Aadam), Mr. Aftab khan.	Softcopy	11 <sup>st</sup> October 2023
<b>Submit System Requirements Specification (SRS)</b>	Dr. S.O. Olatunji (Aadam), Mr. Aftab khan.	Softcopy	4 <sup>th</sup> November 2023
<b>Submit Software Design Specification (SDS)</b>	Dr. S.O. Olatunji (Aadam), Mr. Aftab khan.	Softcopy	18 <sup>th</sup> November 2023
<b>First Semester Final Report</b>	Dr. S.O. Olatunji (Aadam), Mr. Aftab khan.	Softcopy	6 <sup>th</sup> December 2023
<b>Final Presentation</b>	Dr. S.O. Olatunji (Aadam), Mr. Aftab khan.	Softcopy	13 <sup>th</sup> December 2023

Table 5 Project Deliverables

### 3.1.4 Schedule and Budget Summary

Given the predominant use of open-source tools in this project, it enjoys economic efficiency. The anticipated expenses encompass the software utilized for program design and development, as well as potential publication fees. Table 6 outlines the various phases of the project, alongside their projected durations.

<b>The Project Phase</b>	<b>Duration</b>
<i>Proposal Discussion</i>	2 days
<i>Define Chronic Diseases</i>	1 weeks
<i>Datasets Chronic Diseases</i>	1 week
<i>Literature Review</i>	2 week
<i>Mid Semester Report</i>	5 weeks
<i>Software Project Management Plan (SPMP)</i>	17 days
<i>Methodology</i>	1 week
<i>System Requirements Specification (SRS)</i>	2 week
<i>Software Design Specification (SDS)</i>	2 week
<i>First Semester Final Report</i>	18 days
<i>Implementing/Coding</i>	14 weeks
<i>Software Test Plan (STP)</i>	2 weeks
<i>User Manual</i>	4 days
<i>Final Project Report</i>	5 days
<i>Presentation</i>	4 days

Table 6 The project Schedule

### 3.1.5 Evolution of the Plan

This section constitutes the inaugural iteration of the Software Project Management Plans (SPMPs), adhering to IEEE standards. Should no alterations be deemed necessary, the team will proceed with the implementation of these proposed plans. Any revisions will be made exclusively after a thorough team-wide discussion, unanimous consensus, and upon receiving explicit endorsement from our supervisors, Dr. S.O. Olatunji (Aadam) and Mr. Aftab Khan.

### 3.1.6 Definitions

Table 7 contains definitions for the primary terminology used in the work.

Term	Definition
Chronic Disease	Chronic illnesses are generally characterized as conditions persisting for a year or longer, necessitating continuous medical care or imposing restrictions on daily activities, or both. Notably, conditions like heart disease, cancer, and diabetes stand as the primary contributors to global mortality and impairment [62].
Software Project Management Plan (SPMP)	pertains to the utilization of expertise, know-how, methodologies, and resources in overseeing project endeavors to meet specified project criteria. This typically involves delineated stages of project management, comprising project commencement, project blueprinting, project implementation, project oversight, and project finalization [63].
Software Requirements Specification (SRS)	A Software Requirements Specification (SRS) is a documented account specifying the expected behaviors and performance criteria of the software. Moreover, it delineates the functionalities and attributes that need to align with the demands of diverse stakeholders, encompassing both end-users and business entities [64].
Software Design Specification (SDS)	The Software Design Specification (SDS) is a comprehensive account of the individual components and subsystems that will be included in the final product [65].
Software Test Plan (STP)	A Test Plan is an elaborate document outlining objectives, test strategies, resources, deadlines, and schedule for a project. It acts as a blueprint for ensuring correct software functionality, overseen by test managers [66].
Waterfall Model	The Waterfall Model is a traditional software development approach, characterized by sequential phases. It offers well-defined documentation and clear project milestones [67].

Table 7 Terms and Definitions

### 3.1.7 Document Structure

The structure of the document is as follows:

- **Project Overview:** Offers a brief overview of the project's objectives, scope, and aims. It covers foundational assumptions, limitations, potential risks, deliverables, along with an outline of the timeline and financial resources. It also includes acronyms, references, and the overall structure of the document.

- **Project Organization:** This section outlines the internal structure of the project, its external affiliations, and the specific responsibilities of each team member.
- **Management Process Strategies:** This category is divided into five components. The initial part outlines the project's initiation plan, involving tasks like staffing, staff estimates, and training for project personnel. The second part serves as an overview of the project's action plan, including task breakdown structure, scheduling, resource allocation, and budget distribution. The third part specifies the approach for monitoring the project, covering aspects like managing requirements, scheduling, assurance of quality, submitting reports, and metrics for the project. The last part addresses the risk management approach, with the final portion introducing the project's closure strategy.
- **Technical Process Strategies:** Outlines the methods and approaches employed for technical development within the project. It encompasses the selection of programming languages, frameworks, and architecture design choices.
- **Supporting Process Strategies:** It includes publications that provide assistance for process development.

## 3.2 Project Organization

### 3.2.1 External Interfaces

Pre-emptive chronic disease diagnosis project supervisor is Mr. Mohammad Aftab Khan which target some chosen Saudi hospitals. Figure 2 illustrates external interfaces.

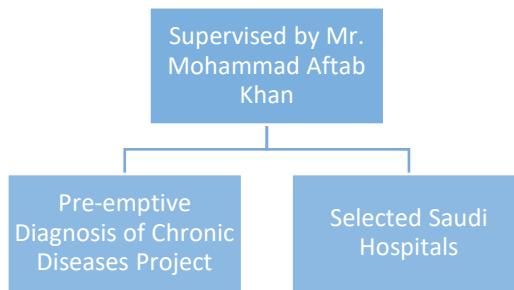


Figure 2 External Interface

### 3.2.2 Internal Structure

The project is supervised by Mr. Mohammad Aftab Khan supervises with the collaboration of Dr. S.O. Olatunji (Aadam) to ensure that everything gets completed on schedule. The team members working on this project: Rahaf Yaan Allah, Fai Al-anazi, Razan Alshammari, Fatimah Alkhathim, and Shahad Alghamdi. Each team member is given a specific role according to their skills, knowledge, and expertise in order to produce optimal outcomes. The project's internal organizational structure is illustrated in Figure 3.

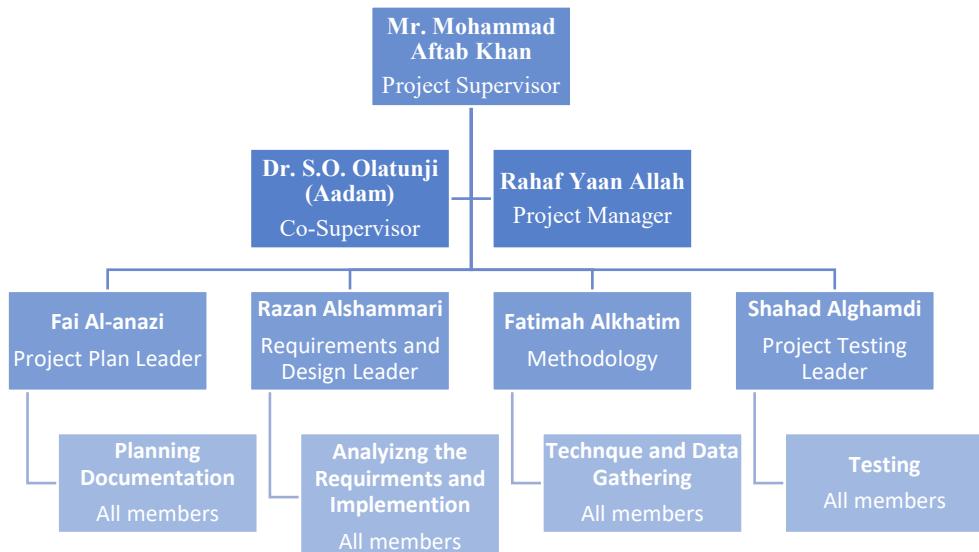


Figure 3 Internal Structure

### 3.2.3 Roles and Responsibilities

The team leader distributes tasks to each team member fairly and monitors their completion. Each team member takes part in each stage of the development process. The team members' tasks and responsibilities are summarized in Table 8.

Role	Responsibilities	Member
<b>Project Manager</b>	<ul style="list-style-type: none"> <li>◆ Direct contact with the supervisor.</li> <li>◆ Define the project's objectives.</li> <li>◆ Distribute the work among team members in an equitable manner.</li> <li>◆ Keep track of the work's progress.</li> <li>◆ Ensure that all tasks are completed on time.</li> </ul>	Rahaf Yaan Allah
<b>Project Plan Leader</b>	<ul style="list-style-type: none"> <li>◆ Keep track of the project's timeline.</li> <li>◆ Support the project manager.</li> <li>◆ Control the threats.</li> <li>◆ Assign project resources.</li> </ul>	Leader: Fai Al-anazi All team members
<b>Requirements and Design Leader</b>	<ul style="list-style-type: none"> <li>◆ Prepare and distribute the requirements.</li> <li>◆ Evaluate the requirements.</li> <li>◆ Complete all design documents.</li> <li>◆ Create a system prototype that satisfies the specifications.</li> </ul>	Leader: Razan Alshammari All team members
<b>Methodology</b>	<ul style="list-style-type: none"> <li>◆ Strategy and data collection.</li> <li>◆ Create a development strategy.</li> <li>◆ Gathering and assessing requirements.</li> </ul>	Leader: Fatimah Alkhatim All team members
<b>Project Testing Leader</b>	<ul style="list-style-type: none"> <li>◆ Test and troubleshoot the system.</li> <li>◆ Complete the test report.</li> <li>◆ Specify the test data and operations.</li> </ul>	Leader: Shahad Alghamdi All team members

Table 8 Roles and Responsibilities

### 3.3 Managerial Process Plans

Plans for project start-up, project work, project tracking, risk management, and project close-out. will be defined in this section.

#### 3.3.1 Start-up Plan

The resources and tools needed to start the project, including an estimation plan, staffing plan, and training plan, will be described in this section. Depending on the size and scope of the project, incorporation between these plans and other plans may take place.

##### 3.3.1.1 Estimates

The cost of the program will be free, based on our experience and the system's application domain. It should be noted that each project stage includes an estimation of cost, resources, and time.

- **Cost Estimate:** The stages of designing, implementing, and testing are free of charge. Depending on the size of the paper, the documentation stage may cost up to 250SR for all documents. Depending on the journal chosen, publication expenses are also included. Additionally, given that some sessions were held at the campus, transportation expenses might add to the expense.
- **Resource Estimate:** Taking into account human resources, including their work and efforts. Additionally, there are other resources available, such as computers and members' training costs for paid courses.
- **Time Estimate:** Within the upcoming seven months, the full project is anticipated to be finished. Each task has a beginning time and an end time. The team will work to resolve any issues and will compensate for any additional time if necessary.

##### 3.3.1.2 Staffing

Five students from Imam Abdulrahman bin Faisal University's 9th level, notably from the college of Computer Science and Information Technology, will work together to construct the project. To complete the project by the end of the academic year, each member will be given a separate duty based on their expertise and experience. All team members should be proficient in fundamental technical abilities like data analysis and programming. In order to facilitate cooperation, they should also possess soft skills including writing, research, decision-making, time management, and communication abilities. Table 9 lists each phase along with the number of people and time required to complete it.

Project Phase	Human Resource	Duration
Project Initiation	All team member	2 weeks
Planning		1 week
Methodology / Requirements Specification		3 weeks
Designing		2 week
Implementation		14 weeks
Testing		2 weeks

Documentation		18 days
---------------	--	---------

Table 9 Human resources

### 3.3.1.3 Project Staff Training

This project is being developed in seven steps, and each phase demands a particular level of expertise and knowledge from all team members in order to be successfully completed. For each role in the team, Table 10 specifies the skills that must be mastered and/or acquired.

Role	Skills
Project Manager	<ul style="list-style-type: none"> <li>• Technical writing skills.</li> <li>• Team working skills.</li> <li>• Communication skills.</li> <li>• Leadership skills.</li> <li>• Problem-solving skills.</li> <li>• Critical thinking skills.</li> <li>• Time management skills.</li> <li>• Decision-making skills.</li> <li>• Negotiation skills</li> </ul>
Project Plan Manager	<ul style="list-style-type: none"> <li>• Communication skills.</li> <li>• Technical writing skills.</li> <li>• Decision making skills.</li> <li>• Analytical skills.</li> <li>• Time management skills.</li> </ul>
Requirements Analysts	<ul style="list-style-type: none"> <li>• Communication skills.</li> <li>• Analytical skills.</li> <li>• Time management skills.</li> <li>• Quick learning skills.</li> <li>• Technical skills.</li> </ul>
Project Design Manager	<ul style="list-style-type: none"> <li>• Creative and imagination skills.</li> <li>• Software design skills.</li> <li>• Reasoning skills.</li> </ul>
Database Designer	<ul style="list-style-type: none"> <li>• Design system ER, EER and mapping.</li> <li>• Knowledge about DBMS.</li> <li>• Knowledge about NoSQL.</li> <li>• Team working skills.</li> </ul>
Project Testing Manager	<ul style="list-style-type: none"> <li>• Decision making skills.</li> <li>• Time management skills.</li> <li>• Analytical skills.</li> <li>• Critical thinking skills.</li> <li>• Team working skills.</li> </ul>
Technical Writing	<ul style="list-style-type: none"> <li>• Research skills.</li> <li>• Writing skills.</li> </ul>

Table 10 Skills needed for each Role.

### 3.3.2 Work Plan

The division of project work into separate activities, timetables, available resources, and budget will all be covered in this section.

### 3.3.2.1 Work Breakdown Structure

Table 11 shows the work breakdown structure for this project.

No.	Work Activity	Resources		Duration	Deliverable	Status
		Software	Hardware			
1	Project Initiation			2 weeks	Mid-Semester Report	Completed
1.1	Initial meeting for the senior project	MS Word	Personal Computer (PC)	1 hours		
1.2	Searching for the most common chronic diseases in Saudi Arabia	MS Word	PC	4 days		
1.3	Choosing the chronic diseases	Browser and MS Word	PC	7 days		
1.4	Searching for datasets	Browser and excel	PC	2 days		
1.5	Updating the chosen diseases	Browser and MS Word	PC	1 days		
2	Introduction chapter			1 week	Mid-Semester Report	Completed
2.1	Writing the introduction chapter	MS Word	PC	5 days		
2.2	Reviewing and editing the introduction chapter	MS Word	PC	2 days		
3	Literature reviews chapter			2 weeks	Mid-Semester Report	Completed
3.1	Writing literature reviews chapter	MS Word	PC	12 days		
3.2	Reviewing and editing the literature reviews chapter	MS Word	PC	2 days		
4	Software Project Management Plan			17 days	SPMP	Completed
4.1	Work plan	MS Word	PC	3 hours		
4.2	Writing the SPMP	MS Word	PC	2 weeks		

4.3	Reviewing the SPMP	MS Word	PC	2 days		
4.4	Submitting the SPMP	MS Word	PC	1 hour		
5	Submitting the Mid-Semester Report			1 hour	Mid-Semester Report	Completed
6	Methodology			1 week	Methodology	Completed
6.1	Writing the methodology	MS Word	PC	5 days		
6.2	Reviewing the methodology	MS Word	PC	2 day		
6.3	Submitting the methodology	MS Word	PC	1 hour		
7	Software Requirements Specification			2 weeks	SRS	Completed
7.1	Collecting the requirements	MS Word	PC	2 days		
7.2	Analyzing the requirements	MS Word	PC	2 days		
7.3	Determining the requirements	MS Word	PC	4 days		
7.4	Submitting the draft of the SRS	MS Word	PC	1 day		
7.5	Updating the requirements	MS Word	PC	3 days		
7.6	Submitting the SRS	MS Word	PC	1 hour		
8	Software Design Specification			2 weeks	SDS	Completed
8.1	Designing initial interfaces	Axure or balsamic	PC	2 days		
8.2	Designing the database for the system	drow.io	PC	4 days		
8.3	Developing the interfaces	Axure or balsamic	PC	4 days		
8.4	Completing the SDS	MS Word	PC	2 days		
8.5	Reviewing SDS	MS Word	PC	2 days		
8.6	Submitting SDS	MS Word	PC	1 hour		
9	Final report			18 days	Final report (for first semester)	Completed
9.1	Reviewing the final report	MS Word	PC	13 days		

9.2	Preparing the final presentation	PowerPoint	PC	4 days		
9.3	Submitting the final report	MS Word	PC	1 hour		
9.4	Submitting the final presentation	PowerPoint	PC	1 hour		
10	Implementation the system			14 weeks	Progress Implementation	Completed
10.1	Pre-processing the dataset	Python	PC	1 week		
10.2	Modeling the classifiers	Python	PC	2 weeks		
10.3	Analyzing the results of the classifiers	Python	PC	1 week		
10.4	Building machine learning models	Browser	PC	2 weeks		
10.5	Developing the database	NoSQL developer	PC	3 weeks		
10.6	Developing the interface	Text editor and server	PC	1 week		
10.7	Coding	Text editor and server	PC	4 weeks		
11	Software Testing Plan			2 weeks	STP	Completed
11.1	Unit testing	Text editor and server	PC	4 days		
11.2	Integration testing	Text editor and server	PC	3 days		
11.3	Testing the whole system	Text editor and server	PC	2 days		
11.4	Completing the STP	MS Word	PC	3 days		
11.5	Reviewing the STP	MS Word	PC	2 days		
11.6	Submitting the STP	MS Word	PC	1 hour		
12	Submitting the final report			18 days	Final report and the code of the project	Completed
12.1	Reviewing the final report	MS Word	PC	5 days		
12.2	Developing the user manual	MS Word	PC	4 days		
12.3	Preparing the final presentation	MS Word and PowerPoint	PC	4 days		

Table 11 Work Breakdown Structure

### 3.3.2.2 Schedule Allocation

The duration, start date, and finish date for each activity are displayed in Table 12. All academic year long, adherence to the schedule is anticipated.

No.	Work Activity	Start date	Duration	End date
1	Project Initiation	20 <sup>th</sup> August 2023	2 weeks	2 <sup>nd</sup> September 2023
1.1	Initial meeting for the senior project	20 <sup>th</sup> August 2023	1 hours	20 <sup>th</sup> August 2023
1.2	Searching for the most common chronic diseases in Saudi Arabia	20 <sup>th</sup> August 2023	4 days	23 <sup>rd</sup> August 2023
1.3	Choosing the chronic diseases	24 <sup>th</sup> August 2023	7 days	30 <sup>th</sup> August 2023
1.4	Searching for datasets	31 <sup>st</sup> August 2023	2 days	1 <sup>st</sup> September 2023
1.5	Updating the chosen diseases	2 <sup>nd</sup> September 2023	1 days	2 <sup>nd</sup> September 2023
2	Introduction chapter	3 <sup>rd</sup> September 2023	1 week	9 <sup>th</sup> September 2023
2.1	Writing the introduction chapter	3 <sup>rd</sup> September 2023	5 days	7 <sup>th</sup> September 2023
2.2	Reviewing and editing the introduction chapter	8 <sup>th</sup> September 2023	2 days	9 <sup>th</sup> September 2023
3	Literature reviews chapter	10 <sup>th</sup> September 2023	2 weeks	23 <sup>rd</sup> September 2023
3.1	Writing literature reviews chapter	10 <sup>th</sup> September 2023	12 days	21 <sup>st</sup> September 2023
3.2	Reviewing and editing the literature reviews chapter	22 <sup>nd</sup> September 2023	2 days	23 <sup>rd</sup> September 2023
4	Software Project Management Plan	24 <sup>th</sup> September 2023	17 days	11 <sup>th</sup> October 2023
4.1	Work plan	24 <sup>th</sup> September 2023	3 hours	24 <sup>th</sup> October 2023
4.2	Writing the SPMP	25 <sup>th</sup> September 2023	2 weeks	8 <sup>th</sup> October 2023
4.3	Reviewing the SPMP	9 <sup>th</sup> October 2023	2 days	10 <sup>th</sup> October 2023
4.4	Submitting the SPMP	11 <sup>th</sup> October 2023	1 hour	11 <sup>th</sup> October 2023
5	Submitting the Mid-Semester Report	11 <sup>th</sup> October 2023	1 hour	11 <sup>th</sup> October 2023
6	Methodology	15 <sup>th</sup> October 2023	1 week	21 <sup>st</sup> October 2023
6.1	Writing the methodology	15 <sup>th</sup> October 2023	5 days	19 <sup>th</sup> October 2023
6.2	Reviewing the methodology	20 <sup>th</sup> October 2023	2 days	21 <sup>st</sup> October 2023
6.3	Submitting the methodology	21 <sup>st</sup> October 2023	1 hour	21 <sup>st</sup> October 2023

7	Software Requirements Specification	22 <sup>nd</sup> October 2023	2 weeks	4 <sup>th</sup> November 2023
7.1	Collecting the requirements	22 <sup>nd</sup> October 2023	2 days	23 <sup>rd</sup> October 2023
7.2	Analyzing the requirements	24 <sup>th</sup> October 2023	2 days	25 <sup>th</sup> October 2023
7.3	Determining the requirements	26 <sup>th</sup> October 2023	4 day	29 <sup>th</sup> October 2023
7.4	Submitting the draft of the SRS	30 <sup>th</sup> October 2023	1 day	30 <sup>th</sup> October 2023
7.5	Updating the requirements	1 <sup>st</sup> November 2023	3 days	3 <sup>rd</sup> November 2023
7.6	Submitting the SRS	4 <sup>th</sup> November 2023	1 hour	4 <sup>th</sup> November 2023
8	Software Design Specification	5 <sup>th</sup> November 2023	2 weeks	18 <sup>th</sup> November 2023
8.1	Designing initial interfaces	5 <sup>th</sup> November 2023	2 days	6 <sup>th</sup> November 2023
8.2	Designing the database for the system	7 <sup>th</sup> November 2023	4 days	10 <sup>th</sup> November 2023
8.3	Developing the interfaces	11 <sup>th</sup> November 2023	4 days	14 <sup>th</sup> November 2023
8.4	Completing the SDS	15 <sup>th</sup> November 2023	2 days	16 <sup>th</sup> November 2023
8.5	Reviewing SDS	17 <sup>th</sup> November 2023	2 days	18 <sup>th</sup> November 2023
8.6	Submitting SDS	18 <sup>th</sup> November 2023	1 hour	18 <sup>th</sup> November 2023
9	Final report	19 <sup>th</sup> November 2023	18 days	6 <sup>th</sup> December 2023
9.1	Reviewing the final report	19 <sup>th</sup> November 2023	13 days	1 <sup>st</sup> December 2023
9.2	Preparing the final presentation	2 <sup>nd</sup> December 2023	4 days	5 <sup>th</sup> December 2023
9.3	Submitting the final report	6 <sup>th</sup> December 2023	1 hour	6 <sup>th</sup> December 2023
9.4	Submitting the final presentation	6 <sup>th</sup> December 2023	1 hour	6 <sup>th</sup> December 2023
10	Implementation the system	14 <sup>th</sup> January 2024	14 weeks	20 <sup>th</sup> April 2024
10.1	Pre-processing the dataset	14 <sup>th</sup> January 2024	1 week	20 <sup>th</sup> January 2024
10.2	Modeling the classifiers	21 <sup>st</sup> January 2024	2 weeks	3 <sup>rd</sup> February 2024
10.3	Analyzing the results of the classifiers	4 <sup>th</sup> February 2024	1 week	10 <sup>th</sup> February 2024
10.4	Building machine learning models	11 <sup>th</sup> February	2 weeks	24 <sup>th</sup> February

		2024		2024
10.5	Developing the database	25 <sup>th</sup> February 2024	3 weeks	16 <sup>th</sup> March 2024
10.6	Developing the interface	17 <sup>th</sup> March 2024	1 week	23 <sup>rd</sup> March 2024
10.7	Coding	24 <sup>th</sup> March 2024	4 weeks	20 <sup>th</sup> April 2024
11	Software Testing Plan	21 <sup>st</sup> April 2024	2 weeks	4 <sup>th</sup> May 2024
11.1	Unit testing	21 <sup>st</sup> April 2024	4 days	25 <sup>th</sup> April 2024
11.2	Integration testing	26 <sup>th</sup> April 2024	3 days	28 <sup>th</sup> April 2024
11.3	Testing the whole system	29 <sup>th</sup> April 2024	2 days	30 <sup>th</sup> April 2024
11.4	Completing the STP	1 <sup>st</sup> May 2024	3 days	3 <sup>rd</sup> May 2024
11.5	Reviewing the STP	4 <sup>th</sup> May 2024	1 days	4 <sup>th</sup> May 2024
11.6	Submitting the STP	4 <sup>th</sup> May 2024	1 hour	4 <sup>th</sup> May 2024
12	Submitting the final report	5 <sup>th</sup> May 2024	18 days	22 <sup>nd</sup> May 2024
12.1	Reviewing the final report	5 <sup>th</sup> May 2024	5 days	9 <sup>th</sup> May 2024
12.2	Developing the user manual	10 <sup>th</sup> May 2024	4 days	13 <sup>th</sup> May 2024
12.3	Preparing the final presentation	19 <sup>th</sup> May 2024	4 days	22 <sup>nd</sup> May 2024

Table 12 Schedule Allocation

### 3.3.2.3 Resource Allocation

Both human and non-human resources are part of the resources. Hardware, software, and written resources are categorized as non-human resources, whereas the team members are considered human resources. The infrastructure plan for non-human resources is shown in Table 13.

Infrastructure Plan	
Hardware	<ul style="list-style-type: none"> <li>• Personal Computer.</li> <li>• Printer.</li> <li>• Mobile Phone.</li> </ul>
Software	<ul style="list-style-type: none"> <li>• Microsoft Office: Word, PowerPoint, Teams, and Excel.</li> <li>• Zoom</li> <li>• Text editor.</li> <li>• Web browser.</li> <li>• Drow.io.</li> <li>• Axure or Balsamiq.</li> <li>• NoSQL Developer.</li> </ul>

	<ul style="list-style-type: none"> <li>• Python.</li> <li>• OneDrive.</li> </ul>
Written resources	<ul style="list-style-type: none"> <li>• Papers about chronic diseases.</li> <li>• Mendeley.</li> <li>• Grammarly.</li> <li>• Textbooks.</li> </ul>

*Table 13 Resource Allocation*

#### **3.3.2.4 Budget Allocation**

Since most of the necessary tools are open source, the project's development budget will be reasonably priced. However, the total cost of printing hard copies for all deliverables may be around 250SR. Publication costs, which can range from 3,100 to 6,000 for each paper, may also be taken into account.

#### **3.3.3 Project Tracking Plan**

The project's plans for requirements management, schedule control, quality control, reporting, and project metrics are detailed in the section below.

##### **3.3.3.1 Requirements Management**

The requirements must be prioritized and precisely described in order to enable easy tracking of the project's progress. Any adjustments to the requirements will be made as necessary. If a modification is required, it is important to comprehend how much of an influence it will have on the project. The requirements must be altered with the support of the entire team and advisers, and it must be done with a great deal of work if the change is mandatory and critical. However, the time element will play a role in the execution of the change if it is deemed less significant and has no impact on the project's continuation.

##### **3.3.3.2 Schedule Control**

Consistent updates are made to Table # in Section 3.3.2.2, where the performance of the tasks and the project as a whole must be observed. The advisers and team members meet once a week to go over the project's status, issues, and concerns. If the plan is altered or delayed for any reason, the cause must be disclosed, and a backup plan must be created to sustain the original plan. Last but not least, effective teamwork depends on open lines of communication.

##### **3.3.3.3 Quality Control**

The following criteria are used to evaluate quality:

- All project requirements are met.
- Use procedures and tests to gauge the success criteria.
- Early error detection and rectification.

### 3.3.3.4 Reporting

The following points are part of the reporting method:

- Regular team meetings to follow up on work that is done in a variety of ways. For instance, online and in-person sessions are both documented and sent to the advisors each week.
- The advisers and I meet once a week to talk about the project's development and updates.

### 3.3.3.5 Project Metrics

The metrics that will be used to gauge the project's success are listed in Table 14.

Metrics	Comment	Frequency
Time	The project progress is compared with the schedule in sections 3.3.2.1 and 3.3.2.2	Continuously
Value	The project's progress is checked according to the requirements specification to ensure that requirements have been achieved	Continuously
Scope	Tracking change requests is necessary to keep the project on time, where the change measures according to the initial scope.	After each task
Quality	All imperfections and quality problems should be fixed to ensure that the system is working successfully.	After each task

Table 14 Project Metrics

### 3.3.4 Risk Management Plan

The project's principal goal is to identify all potential dangers. Table 15 lists some potential risks along the probability, prevention, impact, and actions that will be taken.

No.	Potential risk	Probability	Prevention	Action	Impact
1	Change of meeting time	Low	Inform all the members before the meeting.	Set a meeting time that is agreed upon between members each week.	Time
2	Short time to complete the project	Moderate	Follow the plan as much as possible and collaborate to finalize the work.	Work more hours to complete the project.	Time
3	Delay delivery	High	Work extra hours to meet the deadline.	Set a well-structured plan for management and avoid last-minute submission.	Time
4	Change in the requirements	Moderate	- Work hard and take enough time to collect and	More work and redivision the work between the rest of the team.	Time

			understand the requirements. - Explain the impact of the changes that affect the users.		
5	One of the members drop the course	Moderate	Motivate the members to complete the project.	Additional work if needed.	Time

Table 15 Risk Management Plan

### 3.3.5 Project Closeout Plan

Before the final deadline, all team members must take the following actions according to the project close-out plan:

- Verifying that all system functions and requirements are met.
- Archiving all source code, papers, and project materials.
- Verifying the most recent version of every document and the entire system before submission.
- Giving the supervisor a printed copy and a digital copy of all paperwork that needs to be authorized.

## 3.4 Technical Process Plans

This section includes the process model, methodologies, tools, and techniques. Moreover, it also covers infrastructure and product acceptance.

### 3.4.1 Process Model

The chosen software process model for the deliverables is the Waterfall model. It organizes the project into independent sequential phases. Figure 4 describes the project process model's stages.

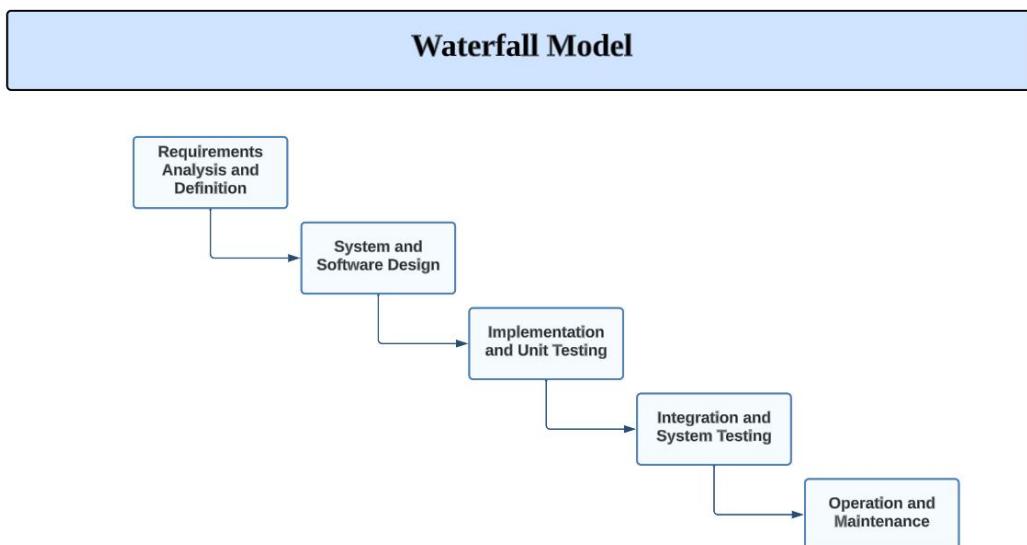


Figure 4 The Waterfall Model

Table 16 shows the waterfall phases, along with their goals and deliverable dates.

Phase	Goal	Deliverable
Requirements Analysis and Definition	<ul style="list-style-type: none"> <li>• Requirements Collecting.</li> <li>• Requirements Analyzing.</li> <li>• Requirements Specification.</li> </ul>	Deliver the SRS document. 28 <sup>th</sup> of October 2023
	<ul style="list-style-type: none"> <li>• Design the interfaces for the system and its structure and show how users interact with it.</li> </ul>	11 <sup>th</sup> November 2023
	<ul style="list-style-type: none"> <li>• Working on the practical part and writing the code using Python programming language.</li> </ul>	13 <sup>th</sup> April 2024
	<ul style="list-style-type: none"> <li>• Test and debug each component separately.</li> </ul>	27 <sup>th</sup> April 2024
	<ul style="list-style-type: none"> <li>• The maintenance phase is out of the scope of the project.</li> </ul>	11 <sup>th</sup> May 2024

Table 16 The Waterfall Phases

### 3.4.2 Methods, Tools, and Techniques

Various tools will be utilized in each phase to conduct the software. Table 17 outlines the variety of tools utilized in each phase.

Phase	Tool
Planning	<ul style="list-style-type: none"> <li>• Microsoft Teams.</li> <li>• Microsoft Excel.</li> <li>• Microsoft PowerPoint.</li> <li>• Microsoft Word.</li> <li>• Zoom.</li> <li>• WhatsApp.</li> </ul>
	<ul style="list-style-type: none"> <li>• Microsoft Word.</li> <li>• Smart Art.</li> </ul>
	<ul style="list-style-type: none"> <li>• Microsoft Word.</li> <li>• Draw.io.</li> <li>• MongoDB.</li> </ul>
	<ul style="list-style-type: none"> <li>• Python.</li> <li>• Axure or Balsamiq.</li> </ul>

Table 17 Used Tools

### 3.4.3 Infrastructure

Table 18 outlines the essential requirements in the development process such as hardware, network, software, and operating system, and other facilities.

Hardware	<ul style="list-style-type: none"> <li>• A personal Laptop for each team member.</li> <li>• A printer.</li> <li>• A mobile phone for each team member.</li> </ul>
----------	---

<b>Software</b>	<ul style="list-style-type: none"> <li>Microsoft Teams for planning.</li> <li>Microsoft Office Programs.</li> <li>Google Collab.</li> <li>Anaconda.</li> <li>Axure or Balsamiq for designing the interface.</li> <li>Draw.io for the diagrams.</li> <li>Creately.</li> <li>MongoDB DBMS to extract the datasets.</li> </ul>
<b>Operating System</b>	Any type of OS could be utilized.
<b>Network</b>	<ul style="list-style-type: none"> <li>Cloud services.</li> <li>High-speed Wireless Network.</li> <li>Anti-viruses and Firewalls.</li> </ul>
<b>Facilities</b>	<ul style="list-style-type: none"> <li>Office.</li> <li>Resources such as printers and scanners.</li> </ul>

Table 18 The Essential Requirements

#### 3.4.4 Product Acceptance

The project should use the best machine learning methods to get the highest accuracy feasible. Therefore, Dr. Sunday Olatunji (Aadam), Mr. Mohammad Aftab Alam Khan, and the reviewers will continuously provide input about the work process and approval to move on to the next step to ensure that all project stages are successfully completed.

#### 3.5 Supporting Process Plans

Various documents will be submitted at various points in the project's schedule. The documentation will be written and reviewed by the entire team. The list of papers that will be created is shown in Table 19.

Document Type	Format Standard	Prepare Document	Review Document
Introduction and literature review	Provided by the supervisor	All team members	Project supervisor: Dr. S.O. Olatunji (Aadam), Mr. Mohammad Aftab Alam Khan
Mid semester report	Provided by the supervisor		
Software Project Management Plan (SPMP)	IEEE Standard 1058-1998		
Software Requirements Specification (SRS)	IEEE Standard 1058-1998		
Methodology	Provided by supervisor.		
Software Design Specification (SDS)	IEEE Standard 1058-1998		
First Semester Final Report	Provided by supervisor.		

Software Test Plan	IEEE Standard 1058-1998		
User Manual	IEEE Standard 1058-1998		
Senior Project Source Code and Report	Provided by supervisor.		

Table 19 Supporting Process Plans

### 3.6 Resource Scheduling

Resource scheduling is an ongoing process through all project stages. The resources required for each task might be human resources (our team member), hardware such as PCs and laptops, or software like VS code. The availability of these resources is essential to complete each task on time without delay. And thus, managing these resources has a significant impact on the project progress. Table 20 represents the resources needed for each task, availability, and Priority and dependency. More details about when each task starts and end on section 3.3.2.2.

No.	Task	Resources	Availability of resources	Priority and dependency
1	Online meetings	<ul style="list-style-type: none"> <li>• All team member</li> <li>• Microsoft teams</li> </ul>	Available	Moderate priority and depends on the project progress
2	Writing introduction chapter	<ul style="list-style-type: none"> <li>• All team member</li> <li>• laptops</li> <li>• Microsoft word</li> <li>• Web browser</li> </ul>	Available	High priority and needed to be complete before starting with Writing Literature reviews
3	Writing Literature reviews	<ul style="list-style-type: none"> <li>• All team member</li> <li>• laptops</li> <li>• Microsoft word</li> <li>• Web browser</li> </ul>	Available	High priority
4	Writing SPMP	<ul style="list-style-type: none"> <li>• All team member</li> <li>• laptops</li> <li>• Microsoft word</li> </ul>	Available	High priority
5	Writing methodology and SRS	<ul style="list-style-type: none"> <li>• All team member</li> <li>• laptops</li> <li>• Microsoft word</li> <li>• Web browser</li> </ul>	Available	High priority
6	Writing SDS	<ul style="list-style-type: none"> <li>• All team member</li> <li>• laptops</li> <li>• Microsoft word</li> <li>• Web browser</li> </ul>	Available	High priority
7	Writing implementation chapter	<ul style="list-style-type: none"> <li>• All team member</li> <li>• laptops</li> <li>• Microsoft word</li> </ul>	Available	High priority depends on implementing the models and website codes
8	Writing STP	<ul style="list-style-type: none"> <li>• All team member</li> <li>• laptops</li> <li>• Microsoft word</li> </ul>	Available	High priority and depend on implementation

9	Writing conclusion chapter	<ul style="list-style-type: none"> <li>• All team member</li> <li>• laptops</li> <li>• Microsoft word</li> </ul>	Available	High priority and depend on all report chapters and project progress
10	Implementing the models	<ul style="list-style-type: none"> <li>• All team member</li> <li>• laptops</li> <li>• CoLab for coding</li> <li>• Datasets to construct the models</li> </ul>	Available	High priority and needed to be completed before writing implementation chapter
11	Implementing the website	<ul style="list-style-type: none"> <li>• All team member</li> <li>• Laptops</li> <li>• Visual Studio Code</li> <li>• MongoDB</li> </ul>	Available	High priority and needed to be completed before writing implementation chapter
12	Presentation	<ul style="list-style-type: none"> <li>• All team member</li> <li>• Microsoft PowerPoint</li> </ul>	Available	High priority depends on all project progress

Table 20 resource scheduling

## **Chapter 4: Methodology / Software Requirements Specification (SRS)**

The operational framework and methods that will be used to develop the machine learning (ML) models are included in this chapter. It also contains a thorough mathematical description of the machine learning methods that will be trained.

### **4.1 Methodology**

#### **4.1.1 Operational Framework**

Our project will implement ten computational intelligence techniques, Random Forest (RF), Adaptive Boosting (AdaBoost), Support Vector Machines (SVM), Support Vector Machine with Gaussian (SVM\_Gaussian), Logistic Regression (LR), k-Nearest Neighbors (k-NN), Decision Tree (DT), Gradient Boosting (GBoost), Gaussian Naive Bayes (GaussianNB), and Extreme Gradient Boosting (XGBoost). Using the specified performance metrics, the outcomes of each approach will be contrasted with those of previous studies.

The general steps of building systems based on computational intelligence are covered in this part. Part 4.1.2 goes on to detail the various computational intelligence approaches, including RF, AdaBoost, SVM, SVM\_Gaussian, LR, KNN, DT, GBoost, GaussianNB, and XGBoost. The performance metrics in section 4.1.3 will be used to evaluate the suggested models' performance, and part 4.1.4 will provide an overview of the entire chapter.

The framework utilized to build the diagnosis models is shown in Figure 5. The method is divided into nine stages, or milestones: gathering data, pre-processing, creating and testing the model, tweaking hyper-parameters, validating the model with the best hyper-parameters, ensemble, feature selection, prediction using the best features and hyper-parameters, and interpreting the results.

##### **4.1.1.1 Step 1: Data Gathering**

The project is focused on developing proactive models for three chronic illnesses—Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia—by leveraging online datasets. Accessing hospital datasets poses significant challenges, primarily due to the paramount need to safeguard patient privacy. Ethical approval processes for obtaining these datasets might be time-consuming, involving in-depth scrutiny by experts to ensure alignment with project objectives. In pursuit of our goals, three distinct online datasets will be utilized. To uphold strict patient privacy standards, sensitive information like national identification, contact details, and addresses have been deliberately excluded from the requested datasets. The acquired datasets are, therefore, limited to demographic information and basic clinical laboratory test results.

###### **4.1.1.1.1 Epileptic Seizure**

The Epileptic Seizure dataset will be obtained from Mendeley data [68]. It contains numerical and categorical variables. If securing the Saudi dataset proves challenging or if it doesn't meet

our needs, an alternative online dataset will be utilized to build the predictive model necessary for the study.

#### **4.1.1.1.2 Osteoporosis**

The dataset for the Osteoporosis project includes a combination of numerical and nominal data. It is available in Mendeley data [69]. Numerical data encompasses attributes like age, weight, height, BMI, Maximum Walking Distance, and Number of Pregnancies, while nominal data includes categorical information such as gender, Joint Pain, and binary indicators for various health-related factors.

#### **4.1.1.1.3 Sickle Cell Anemia**

The Sickle Cell Anemia dataset is a comprehensive collection of both numerical and nominal data. It includes essential health-related information obtained from Al Fashir Teaching Hospital North Darfur State, Sudan [70].

### **4.1.1.2 Step 2: Dataset Pre-processing**

The model's performance is intricately tied to the data's quality employed for its training. Consequently, constructing an effective model necessitates data preprocessing, which encompasses the management of discrepancies and absent data points within the initial dataset. A multitude of preprocessing approaches can be harnessed to enhance the model's efficacy, including addressing missing values, data scaling, applying sampling techniques, and standardizing the data. Furthermore, techniques like encoding categorical variables play a pivotal role in converting the raw data into a format comprehensible to algorithms, given that these algorithms fundamentally rely on mathematical computations.

### **4.1.1.3 Step 3: Model Building and Testing**

We will construct prediction models using five machine learning algorithms: RF, AdaBoost, SVM, SVM\_Gaussian, LR, KNN, DT, GBoost, GaussianNB, and XGBoost. In this phase, we create the initial models by training and testing the classifiers with their default hyper-parameters using the pre-processed data. Subsequently, we evaluate the model's performance to guide us in further refining these models in the following phase.

### **4.1.1.4 Step 4: Hyper-parameter Tuning**

Hyper-parameter tuning is a crucial process involving the systematic adjustment of a machine learning algorithm's hyper-parameters to optimize model performance. In this project, we'll employ the grid search technique for hyper-parameter tuning. GridSearchCV operates by establishing a grid-like search space, wherein hyperparameters are defined with a range of values. It systematically evaluates all possible combinations to identify the hyperparameters that result in the highest model performance.

### **4.1.1.5 Step 5: Validation with Optimal Hyper-parameters**

In this stage, the models configured with the most effective hyper-parameters undergo validation, which is achieved through data partitioning techniques like train/test splitting or K-fold cross-validation. The validation process assesses the models to ascertain whether they

meet the desired performance standards. If the individual models fall short of expectations, an ensemble technique will be deployed. In the event of satisfactory model performance, feature selection will be promptly conducted.

#### **4.1.1.6 Step 6: Ensemble Technique**

To enhance the generalization capability of an individual model, the ensemble method is employed, unifying multiple base models into one. While various ensembles are conceivable, the most prevalent methods encompass bagging, boosting, and stacking. It's important to note that this stage is only initiated when the model's performance falls short of expectations. In bagging, numerous homogeneous weak learners are united, each independently trained on distinct subsets of the dataset, and their predictions are subsequently averaged to generate the ultimate output. Conversely, boosting operates sequentially, training homogeneous weak learners based on the preceding model's output. The final result is ascertained in a manner akin to bagging. In contrast, stacking concurrently amalgamates diverse weak learners and harnesses their outcomes to educate another algorithm called the meta-classifier, which supervises the ultimate outcome's determination.

#### **4.1.1.7 Step 7: Feature Selection**

Feature selection involves cherry-picking a subset of attributes to enhance a machine learning model's learning process. In this project, we commence by computing the correlation coefficient between each attribute and the target feature, arranging them in a descending order of importance. We employ a recursive feature elimination technique to iteratively cull the least influential half of the attributes until a single, optimal attribute remains. Our elimination process unfolds as follows:

1. We initialize with all attributes ( $V$  features) and train the classifier.
2. Next, we calculate each attribute's correlation with the target feature and select the top  $V/2$  attributes based on their correlation values.
3. The process repeats until just one attribute persists.
4. We determine the most effective feature subset that maximizes accuracy.

Additionally, we explore the select k best algorithm, utilizing the chi-square test to gauge the relationship between each feature and the class variable. The chi-square test evaluates the discrepancy between observed and expected category frequencies, and our selection unfolds as follows:

1. We compute the chi-square score for each feature, quantifying their association with the class variable.
2. Features are ranked based on their descending chi-square scores, with higher scores indicating stronger connections to the class variable.
3. We choose the top  $k$  features with the highest chi-square scores, allowing flexibility in determining  $k$  based on domain knowledge or validation set performance.

Furthermore, we delve into the genetic algorithm for feature selection:

1. Features are encoded as binary strings, and new feature subsets are bred through genetic operations, including mutation and crossover.
2. Fitness functions assess each subset's performance, preserving high-performing subsets for future generations.
3. The process repeats until convergence, ensuring a refined feature selection for model enhancement.

#### 4.1.1.8 Step 8: Prediction with best Feature Subset and Hyper-parameters

This phase involves constructing the final models using the most effective feature subset and hyperparameters determined in earlier stages.

#### 4.1.1.9 Step 9: Result Interpretation

In this phase, the prediction outcomes are thoroughly examined and explained, considering the predefined performance metrics detailed in section 4.1.3.

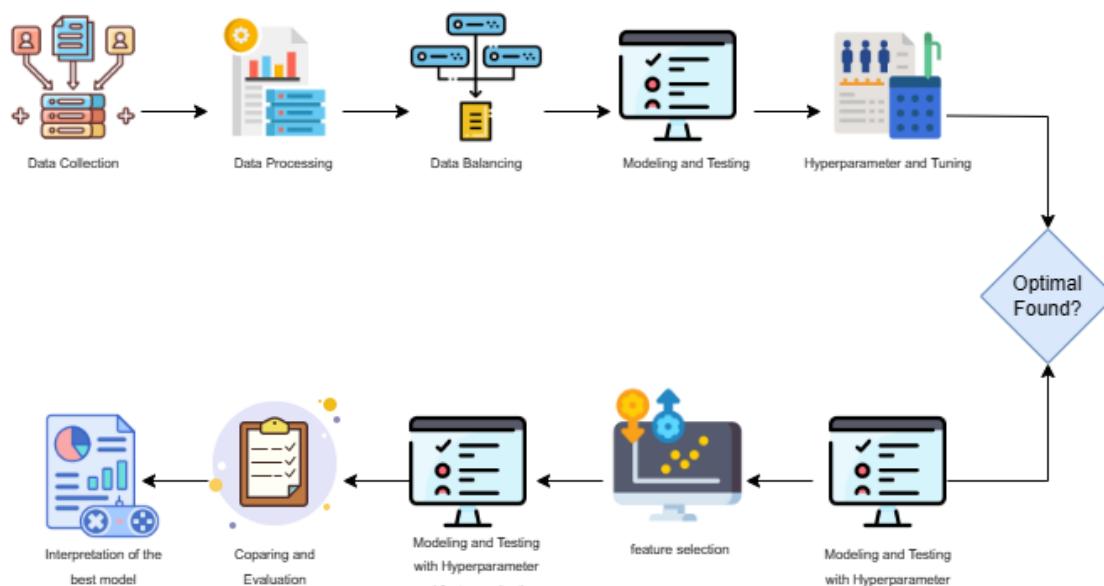


Figure 5 Operational Framework

### 4.1.2 The Proposed Individual Computational Intelligence Approaches

This section provides a theoretical overview of the computational intelligence methods employed, encompassing RF, AdaBoost, SVM, SVM\_Gaussian, LR, KNN, DT, GBoost, GaussianNB, and XGBoost.

#### 4.1.2.1 Random Forest (RF)

Random Forest (RF) is a well-established machine learning method that forms an integral part of the supervised learning approach introduced by Breiman in 2001 [71]. RF finds applications in both classification and regression tasks. It capitalizes on the concept of ensemble learning; wherein diverse classifiers are amalgamated to address complex problems and enhance model

performance. RF operates as a meta-learner, comprising multiple individual learners in the form of trees. To arrive at a final decision for a given set of inputs, RF leverages classifications from numerous random trees. Notably, each classification carries equal weight in the voting process [72]. The voting mechanism used by the RF classifier to determine the final output is expressed in Equation (1), incorporating data point  $x$ , the combined outcomes,  $C_n$  from the  $n$ -th sample, and the overall result  $\hat{y}_f$  [73]. Achieving a balanced dataset involves random sampling, where the high-quality and low-quality variants identified in the previous phase serve as the training data for RF. Given the relative scarcity of low-quality variants, high-quality variants are selected at random, ensuring their sample size matches that of low-quality variants. Consequently, the balanced dataset comprises twice the number of samples as the low-quality variants. The features used to train an RF are pivotal for characterizing datasets [74].

$$\text{Equation (1): } \hat{y}_f = \text{mode} \{C_1(x), C_2(x), \dots, \dots, C_n(x)\}$$

RF generates  $n$  distinct trees employing a variety of feature subsets. Each tree yields a classification outcome, and the final decision of the classification model hinges on a majority vote, as depicted in Figure 6. The class garnering the highest number of votes is assigned to the sample [71]. RF offers several advantages, including its capability to handle a considerable number of missing values through either median replacement or proximity-weighted mean replacement. However, it is worth noting that RF is often associated with the downside of slow prediction generation due to the decision tree's intricacies [72].

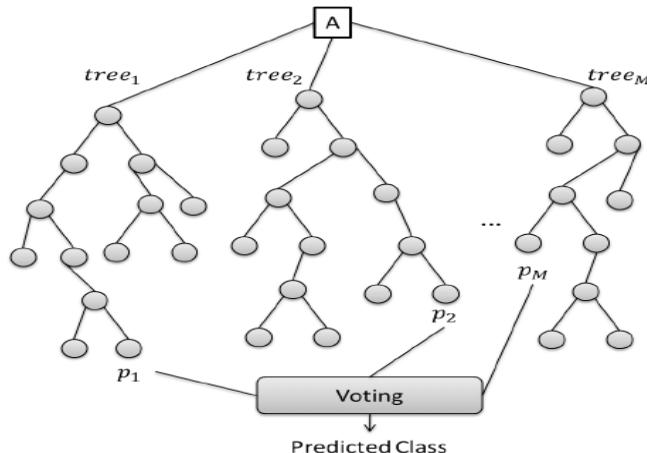


Figure 6 Majority Voting in Random Forest [72]

#### 4.1.2.2 Support Vector Machines (SVM)

Support Vector Machine (SVM) made its debut in the late 1990s, introduced by Vapnik, Cortes, and Boser, and has since garnered substantial recognition within the machine learning community [75]. SVM stands as a widely employed supervised machine learning classifier, serving as a solution for both classification and regression tasks. In practice, SVM finds its primary application in addressing classification challenges [76]. Its adeptness in dealing with

real-world data makes it a preferred choice when compared to alternative techniques [77]. SVM further excels in processing high-dimensional data, rendering it suitable for various applications like spam detection and medical diagnostics, both of which are instances of binary classification tasks [78], [79].

The SVM algorithm delves into the realm of training examples belonging to specific classes and leverages support vectors to craft a boundary that effectively segregates the data into two distinct classes. This boundary takes on the moniker of a hyperplane, representing a subspace of  $p-1$  dimensions [80]. The quest for the optimal hyperplane entails the maximization of the margin, denoting the separation between the hyperplane and the support vectors. The visual depiction of support vectors is showcased in Figure 7 as the dataset's extreme points.

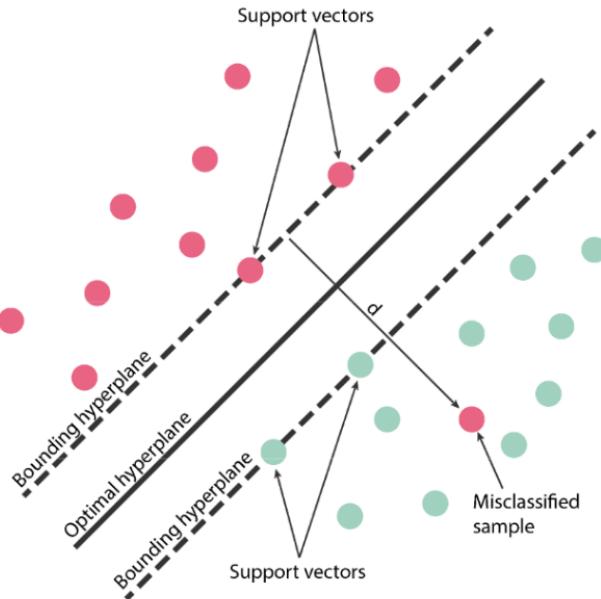


Figure 7 Hyperplane that separates two classes with support vectors points [81]

Equation (2) is used to construct the hyperplane, while Equations (3) and (4) are used to arrange the points to the left and right of the hyperplane [82].

$$\text{Equation (2): } w^T x_i + b = 0$$

$$\text{Equation (3): } w^T x_i + b > 0, y_i = 1$$

$$\text{Equation (4): } w^T x_i + b < 0, y_i = -1$$

Where  $w$  represents the weight vector,  $x$  is the input vector, and  $b$  represents the bias.

A margin maximization is necessary to determine the best hyperplane for a better classification. It is possible to obtain it by reducing the weight vector. Thus, generalization control is obtained. The constrained optimization problem for finding the maximum marginal hyperplane is shown in Equation (5) [82].

Equation (5): minimize  $= \frac{1}{2} ||w||^2$   
 subject to  $y_i(\langle w \cdot x \rangle + b) \geq 1$

To handle the restricted optimization issue, the Lagrangian duality theory is introduced. By computing the derivative after adding a scalar  $a$ , we get [83]:

Equation (6):  $w = \sum_{i=1}^N a_i y_i x_i$   
 and

Equation (7):  $L = \sum_{i=1}^N a_i - \frac{1}{2} \sum_{i,j=1}^N a_i a_j y_i y_j x_i \cdot x_j$   
*subject to*  $\sum_{i=1}^N a_i y_i = 0$  and  $a_i \geq 0$

As mentioned earlier, linearly separable data can be effectively divided using a hyperplane. Nevertheless, tackling the separation of non-linear data necessitates a more advanced approach. In response to this challenge, SVM was extended to handle non-linear data through the introduction of the Kernel Function, a concept known as the "kernel trick." This ingenious technique elevates non-linear data to a higher-dimensional space, enabling their separation with greater efficacy [84]. Figure 8 visually illustrates the application of a Kernel Function, demonstrating how it maps non-linearly separable data into a higher-dimensional realm.

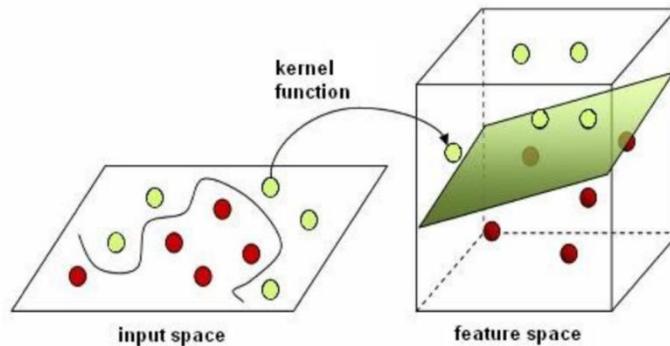


Figure 8 Mapping nonlinearly separable to a higher dimension using kernel function [85]

Below is a list of Kernel Functions' equations that are widely utilized in different applications [86], [87].

Linear kernel function:

$$\text{Equation (8): } K(x_i, x_j) = x_i^T x_j$$

Polynomial:

$$\text{Equation (9): } K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$$

RBF Kernel function (Exponential Radial Basis):

$$\text{Equation (10): } K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0$$

Sigmoid Kernel function (Multi-Layers Perceptron):

$$\text{Equation (11): } K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$$

#### 4.1.2.3 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm, introduced by Evelyn Fix and Joseph Hodges in 1951 and further developed by Thomas Cover [88], operates as a non-parametric, instance-based learner. KNN, often categorized as a "lazy learner," retains the training dataset rather than immediately deriving a model from it. The fundamental concept behind this classifier involves making predictions for hidden data points based on their proximity to neighboring data points. Prediction accuracy relies on the calculation of distances between data points. In the initial step of applying KNN, the number of nearest neighbors with the shortest distance to the target point is determined. Subsequently, a suitable class or value is assigned to the new data point through a majority voting mechanism [89]. The Minkowski distance formula, as presented in Equation (12), calculates the distance between two points ( $x$  and  $y$ ), with 'k' representing the number of neighbors and 'p' allowing flexibility in computing either Euclidean distance (when 'p' equals 2, as shown in Equation 13) or Manhattan distance (when 'p' equals 1, as represented in Equation 14) [90].

$$\text{Equation (12): } d(x, y) = \left( \sum_{i=1}^k |x_i - y_i| \right)^{1/p}$$

$$\text{Equation (13): } d(x, y) = \sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

$$\text{Equation (14): } d(x, y) = \sum_{i=1}^k |x_i - y_i|$$

A major advantage of using the K-Nearest Neighbors (KNN) algorithm is the absence of a need to construct a model or perform intricate tuning. Unlike eager learners, which require model training, KNN is a lazy learner, employing all available data points during the prediction process. Nevertheless, this approach entails significant computational costs due to the storage of all training data and results in high memory consumption [91].

#### 4.1.2.4 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an ensemble gradient boosting classifier based on decision trees. It was first developed in 2016 by Chen and Guestrin and considered a highly precise and scalable algorithm to solve classification as well as regression problems [92]. It is an upgraded version of the gradient boosting machine, and its main idea is to train numerous subtrees successively from an initial tree to lower the error of the previous model. XGBoost function as below.

Suppose the used dataset is presented as follows:

$$\text{Equation (15): } D = \{(x_i, y_i) \text{ where } x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$$

Where  $m$  represents the number of independent features,  $x_i$ , and  $y_i$  is the target class of the specified instance  $i$ . At first, the predicted value,  $\hat{y}$ , of a specified sample,  $i$ , is calculated by Equation (16).

$$\text{Equation (16): } \hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F$$

Where  $K$  denotes for the total number of trees,  $f_k(x_i)$  represents the score of the  $k^{th}$  tree, and  $F$  is the  $ki$  function's space containing all trees [93]. Next, Equation (17) is used to calculate the objective function of XGBoost,  $\mathcal{L}$ , with the intention of minimizing it to discover the functions of the regression tree model,  $f_k$ .

$$\text{Equation (17): } \mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

The difference between the prediction  $\hat{y}_i$  and the actual value  $y_i$  is calculated using the training loss function,  $l(y_i, \hat{y}_i)$ . As a means of avoiding the overfitting issue the model's complexity is penalized, by using the term  $\Omega$ , as shown below.

$$\text{Equation (18): } \Omega(f_k) = \gamma T + \frac{1}{2} \lambda ||w||^2$$

When  $\gamma$  and  $\lambda$  are the regularization parameters,  $T$  and  $w$  are the leaf counts and individual leaf scores, respectively [94].

By incorporating several algorithmic improvements, XGBoost improves Gradient Boosting Trees and has some distinctive advantages. First, by applying the second derivative and carrying out the second-order Taylor expansion, it can accelerate the model's training's convergence rate, enabling faster model training and more effective use of memory resources. Second, as shown above in Equation (18), the regularization terms added to the objective function controls the tree model's complexity to produce a more basic model and prevent overfitting. Third, XGBoost offers a great degree of flexibility and enables users to specify unique optimization goals and evaluation criteria [93].

#### 4.1.2.5 Gradient Boosting (GBoost)

Gradient Boosting (GBoost) is a powerful machine learning algorithm that belongs to the ensemble learning family, widely used for both classification and regression tasks [95]. It operates by sequentially training a series of weak learners, typically decision trees, and correcting the errors of the previous models to improve overall predictive performance. The idea of gradient boosting originated in the observation by Leo Breiman [96] that boosting can be interpreted as an optimization algorithm on a suitable cost function. By combining multiple

models, GBoost leverages their collective strength, making it robust against overfitting and capable of handling complex, high-dimensional data. Moreover, GBoost introduces an element of adaptability by assigning varying weights to different training samples based on their prediction errors, emphasizing challenging instances. This results in highly accurate and resilient models that have found applications in diverse domains such as online advertising, anomaly detection, and ecological modeling.

#### **4.1.2.6 Logistic Regression (LR)**

Logistic Regression is widely used in statistical software and machine learning for predicting the outcomes of binary variables. It functions by fitting a logistic curve to the data, allowing for the estimation of probabilities that range between 0 and 1. This model is particularly effective in scenarios like email filtering (spam vs. non-spam) or disease diagnosis (sick vs. healthy), where the outputs are distinctly categorized [97].

#### **4.1.2.7 Decision Tree (DT)**

A Decision Tree is a straightforward yet powerful tool for both classification and regression tasks, allowing models to predict outcomes based on decisions from a series of questions asked about the observed features. The simplicity of decision trees helps in understanding the decision-making logic at a glance, making it a preferred choice for initial data analysis tasks. This technique divides a dataset into smaller subsets while at the same time, an associated decision tree is incrementally developed [98].

#### **4.1.2.8 Adaptive Boosting (AdaBoost)**

AdaBoost, or Adaptive Boosting, enhances the capability of simple models to improve their accuracy by focusing more on examples that previous models misclassified. This method sequentially adjusts the emphasis on problematic observations, thus allowing weak learners to focus more on difficult parts of the data. It's especially effective for binary classification problems and demonstrates how combining multiple weak learners can achieve high accuracy [99].

#### **4.1.2.9 Support Vector Machine with Gaussian (SVM\_Gaussian)**

Support Vector Machines with Gaussian kernels are an extension of linear models that are designed to capture complex relationships in data by transforming them into higher-dimensional spaces where a linear separation is possible. This approach is well-suited for datasets where the decision boundary is not linear and can handle various feature types, making it versatile for many practical applications [100].

#### **4.1.2.10 Gaussian Naive Bayes (GaussianNB)**

The Gaussian Naive Bayes algorithm extends the Naive Bayes framework to accommodate continuous data that follows a Gaussian distribution. It's particularly useful in fields like document classification and medical diagnosis, where it handles the independent features with Gaussian distributions effectively. GaussianNB is appreciated for its simplicity and speed in making predictions [101].

#### 4.1.3 Performance Evaluation of the Proposed Models

The Models' performance will be evaluated based on several performance metrics, including accuracy, precision, recall, and f-measure. Accuracy can be defined as the percentage of the correctly predicted result. Precision can be defined as the percentage of the other subjects that were correctly predicted as other subjects. Recall can be defined as the percentage of the infected subjects that were correctly predicted as infected subjects. F-measure combines the properties of precision and recall in a single score [40]. These measures play a vital role in evaluating the model's performance, while some measurements rely on others. For instance, to calculate accuracy, both precision and recall must be known. TP denotes True Positive, TN denotes True Negative, FP denotes False Positive, and FN denotes False Negative [102].

$$\text{Equation (19) Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Equation (20) Precision} = \frac{TP}{TP+FP}$$

$$\text{Equation (21) Recall} = \frac{TP}{TP+FN}$$

$$\text{Equation (22) F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

As previously mentioned, metrics, the FN, which can be defined as the wrong predicted negative test result, and the FP, which can be defined as the wrong predicted positive test result, are important to be considered to avoid misdiagnosis that represents more risk in FN, which may put people's lives in risk.

#### 4.1.4 Summary

This section proposes several computational intelligence techniques with general procedure steps to develop them. These techniques will be implemented in the requested datasets from Saudi hospitals to detect the possibilities of developing Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia.

The first technique is RF and it is a powerful ensemble learning method for classification and regression tasks, combining multiple decision trees with equal-weighted voting. It handles imbalanced datasets and missing values but may have slower predictions due to complex trees. The second algorithm is SVM, it is a widely used supervised classifier, primarily for classification tasks, excelling in real-world and high-dimensional data applications. SVM constructs an optimal hyperplane to separate classes, maximize the margin between the hyperplane and the support vectors. The third proposed technique is KNN, which is a non-parametric, lazy learner for prediction based on data point proximity. It calculates distances to determine the neighbors and relies on majority voting for classification. The fourth technique

is XGBoost, a powerful ensemble gradient boosting classifier based on decision trees. It enhances training speed, controls model complexity, and allows customized optimization goals. The fifth proposed algorithm is GBoost, which is an ensemble algorithm used for classification and regression problems. It corrects errors by sequentially training weak learners (often decision trees), handling complex data, and emphasizing challenging instances for robust, accurate models in various applications. The sixth algorithm called logistic regression which commonly used for binary classification. seventh is decision tree, which is a simple classification algorithm used for both classification and regression, it makes the predication depending on the answers to several questions. The eighth technique is adaboost, it combines multiple weak learners to achieve higher accuracy. Ninth is SVM\_Gaussian that can detect complex relationship on the dataset. Lastly, the tenth algorithm is GaussianNB which knowingly used in medical field. The evaluation metrics will be performed after the implementation are precision, accuracy, f-measure, recall, and others to determine the results effectiveness.

## 4.2 Software Requirements Specification (SRS)

### 4.2.1 Introduction

The Software Requirements Specification (SRS) gives a detailed overview of the software system development process.

#### 4.2.1.1 Overview

The SRS chapter is divided into the following six sections to address all the project's requirements:

- **Section 1 Introduction:** The section introduces the SRS and an outline of the chapter's structure.
- **Section 2 Overall Description:** The section displays the system's viewpoint and features. Additionally, it describes the traits of the intended users of the system.
- **Section 3 Specific Requirements:** The section explains the functional needs of system users as well as the requirements for the system's external interfaces.
- **Section 4 Performance Requirements:** This section details the system's response time, availability, fault tolerance, recovery, and capacity performance requirements.
- **Section 5 Design Constraints:** The section outlines the design restrictions for the program, such as the implementation language and hardware and software limitations.
- **Section 6 Software System Attributes:** The section discusses the software system attributes, including usability, reliability, and availability. In addition, it includes the safety requirements, information security, privacy level, maintainability, and portability of the system.

#### 4.2.2 Overall description

This section will provide a general description of the system. It contains the product perspective, product functions, and the user characteristics of the system.

#### **4.2.2.1 Product Perspective**

The updated system is a stand-alone application that diagnoses the following diseases: Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia.

In the previous phases, it was built to diagnose Diabetes, Coronary Heart Disease, Chronic Kidney Disease, Schizophrenia, Asthma, Breast Cancer, Rheumatoid Arthritis, Thyroid Cancer, Alzheimer's, Hypothyroidism, Attention Deficit Hyperactivity Disorder, Glaucoma, Lung Cancer, Prostate Cancer, Cervical Cancer, MS, Schizophrenia, Hepatitis C, Depression, Liver Cirrhosis, Chronic Obstructive Pulmonary, and Parkinson's Disease. The collected dataset will be used to create classification models for patients' early diagnoses using machine learning techniques.

The system contains five users, and their roles are as follows:

- The admins can manage all users, update their own profile, update the diagnosis models, and manage the system's database.
- The medical specialists can update their own profile information, enter the patient's information to be diagnosed, view their diagnosis history, and print patients' results.
- The laboratory specialist can update their own profile information and enter the patients' demographic and laboratory test data for either an overall diagnosis or a single disease diagnosis.
- The registered users can create and delete their account, update their own profile information, enter their information to be diagnosed, view their diagnosis history, and print their results.
- The guest users can diagnose themselves by entering their information without logging in. They can also print their results.

#### **4.2.2.2 Product Functions**

The system aims to preemptively diagnose patients for the presence of previously and newly chosen chronic diseases. For each user, the system will provide many functionalities. The main functions of the system are:

- **Account:**
  - The users should be able to log into their account profile or log out to exit the system.
  - The admins, laboratory and medical specialists can update their email and password.
  - The registered users can update their email, password, name, gender, and date of birth.
  - The admins can manage the system users by adding users of type admin, laboratory specialists, or medical specialists.
  - Registered users can manage their own accounts by having the ability to create and delete their accounts.
  - In a case where a user forgets their password, a new password will be sent to the user's email address.
- **Diagnosis:**

- Laboratory specialists can issue a predictive diagnosis of the patients for each of the chronic diseases.
  - Laboratory specialists can issue an overall predictive diagnosis for two or more chronic diseases at once.
  - The medical specialists, registered users, and guest can fill the predictive diagnosis form and view the results for each of the chronic diseases.
  - All users except for admins and laboratory specialists are allowed to print diagnosis results.
- **View patients' diagnosis history:**
    - The medical specialists will be able to view patients' diagnosis history for each or all chronic disease/s, including information about the accuracy of the model used for the diagnosis.
    - The registered users will be able to view their own diagnosis history.
  - **Rebuilding and updating the diagnostic models:**
    - The admins will be able to rebuild or update the prediction model for each disease by uploading the new dataset and generating the model, and this process will be stored in the system. Figure 9 illustrates the system use-case diagram.

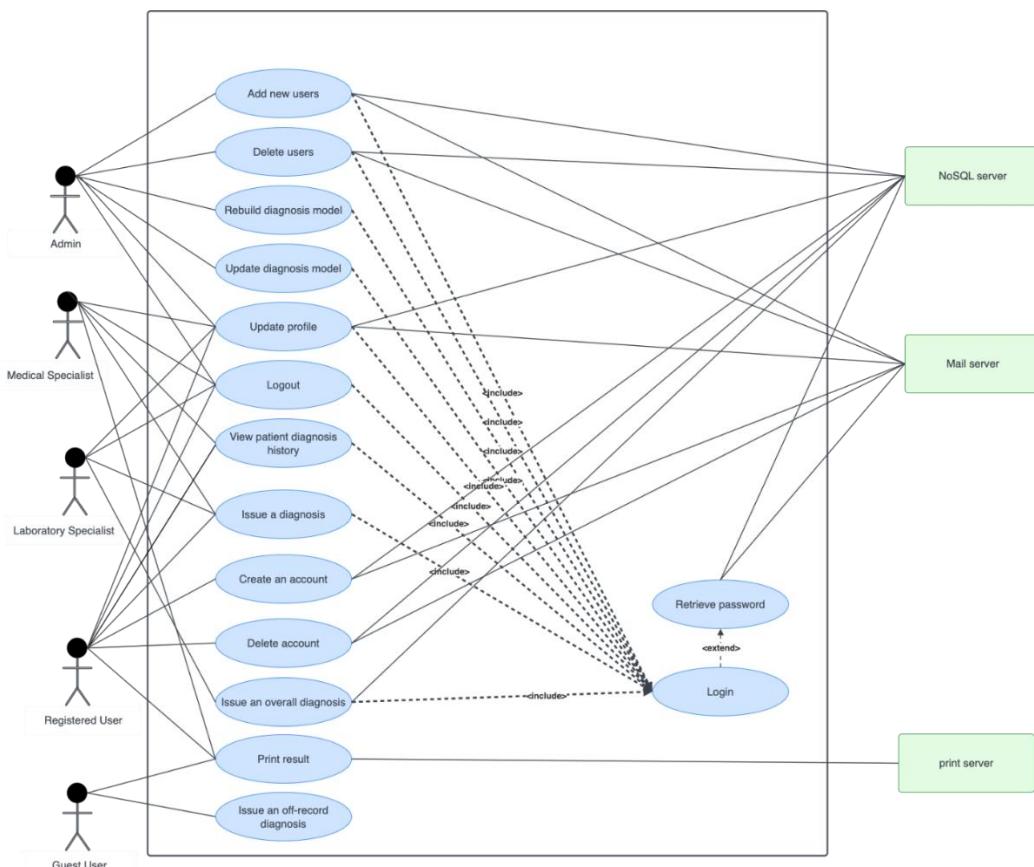


Figure 9 The system use-case diagram

#### 4.2.2.3 User Characteristics

The admins, medical specialists, laboratory specialist, registered users, and guest users are the system's users. The general characteristics of the users are shown in Table 21. The system use-case diagram

User	General Characteristics	
	Education Level	Technical Expertise
Admin	Holds a technical bachelor's degree	<ul style="list-style-type: none"><li>Has technical expertise.</li><li>Has data mining skills.</li><li>Trained to use the system.</li></ul>
Medical Specialist	Holds a technical bachelor's degree in medicine	<ul style="list-style-type: none"><li>At least has basic knowledge of using computers.</li><li>Trained to use the system.</li><li>Have background knowledge about the chronic diseases' clinic tests.</li></ul>
Laboratory Specialist	Holds a technical bachelor's degree in medical laboratory	<ul style="list-style-type: none"><li>At least has basic knowledge of using computers.</li><li>Trained to use the system.</li><li>Able to provide detailed clinical test results and demographical data.</li></ul>
Registered user	Any level of education	<ul style="list-style-type: none"><li>At least has basic knowledge of using computers.</li></ul>
Guest users	Any level of education	<ul style="list-style-type: none"><li>At least has basic knowledge of using computers.</li></ul>

Table 21 User Characteristics

#### 4.2.3 Specific Requirements

This section will provide a detailed description of the system's requirements including the external interfaces requirements and the functional requirements of each system user.

##### 4.2.3.1 External Interfaces Requirements

This section describes the external interfaces that will be used in our system, specifically, the users, hardware, and software interfaces.

###### 4.2.3.1.1 User Interface

The user interface is the main interface of our system, where the user interacts with the system in a friendly and convenient manner to perform the functionalities. User interfaces have a consistent layout and carry out high-security techniques.

#### 4.2.3.1.1 Login Interface

The login interface is the first interface displayed to the user. The user will be able to enter his username and password to access the system if the entered information is valid. Table 22 shows a brief description of the main fields in the Login Interface.

Field Name	Format	Level	Input/Output	Comment
Username	Text	Required	Input	Unique
Password	Encrypted Text	Required	Input	The entered password must be at least 8 alphanumeric

Table 22 The Main Fields of the Login Interface

#### 4.2.3.1.2 Create Account Interface

Any non-medical staff are allowed to register in the system using this interface to follow up on their medical reports. Table 23 shows a brief description of the main fields in the Create Account Interface.

Field Name	Format	Level	Input/Output	Comment
Username	Text	Required	Input	-
Name	Text	Required	Input	-
Email	Text	Required	Input	Unique, and must include the @ symbol
MRN	Integer	Required	Output	Auto-generated unique MRN
Date of Birth	Date	Required	Input	-
Gender	Text	Required	Input	-
Password	Encrypted Text	Required	Input	The entered password must be at least 8 alphanumeric
Confirm Password	Encrypted Text	Required	Input	Must match the entered password.

Table 23 The Main Fields of the Create Account Interface

#### 4.2.3.1.3 Profile Interface

The profile tab is shown on all users' homepage, and once the user clicks on the tab, the profile interface will be displayed. The admins, medical specialists, laboratory specialists, and registered users can change their email and password through this interface, where the users are directed to the Change Email and Change Password interfaces. The registered users can also update their name, gender, and birthdate. The registered users can also delete their accounts. Table 24 shows a brief description of the main fields in the Profile Interface.

Field Name	Format	Level	Input/Output	Comment
Name	Text	Required	Input/Output	It can be changed by registered users only.
Role	Text	Required	Output	-
Email	Text	Required	Output	Unique, and must include the @ symbol

Table 24 The Main Fields of the Profile Interface

#### 4.2.3.1.1.4 Change Email Interface

The user can easily change his email by entering the old one and updating it with a new email. Table 25 shows a brief description of the main fields in the Change Email Interface.

Field Name	Format	Level	Input/Output	Comment
Current Email	Text	Required	Input	Unique
New Email	Text	Required	Input	Unique, and must include the @ symbol
Confirm New Email	Text	Required	Input	Must match the new email.

Table 25 The Main Fields of the Change Email Interface

#### 4.2.3.1.1.5 Change Password Interface

The user can easily change his password by entering the old one and updating it with a new password. Table 26 shows a brief description of the main fields in the Change Password Interface.

Field Name	Format	Level	Input/Output	Comment
Current Password	Encrypted Text	Required	Input	-
New Password	Encrypted Text	Required	Input	The entered password must be at least 8 alphanumeric.
Confirm New Password	Encrypted Text	Required	Input	Must match the new password.

Table 26 The Main Fields of the Change Password Interface

#### 4.2.3.1.1.6 Admin Interface

This interface will be displayed to the system admin after logging into the system. The system admin has complete control in updating the diagnosis model to include more data. Besides, the system admin can manage the system users by adding and deleting users. This interface is separated into two tabs: The Manage Users Tab and The Update Model Tab.

### ❖ Manage Users Tab

In this tab, the admin will be able to remove or add a user to the system. Table 27 shows a brief description of the main fields in the Manage Users Tab.

Field Name	Format	Level	Input/Output	Comment
Username	Text	Required	Output	-
Name	Text	Required	Output	-
Email	Text	Required	Input/Output	Unique, and must include the @ symbol
Role	Text	Required	Output	All types of users will be listed.

Table 27 The Main Fields of the Manage Users Tab in the Admin Interface

### ❖ Rebuild and Update Model Tab

In this tab, the admin will be able to rebuild and update the diagnosis system model. Table 28 shows a brief description of the main fields in the Rebuild and Update Model Tab.

Field Name	Format	Level	Input/Output	Comment
Datafile.csv	CSV file	Required	Input	-
Disease	Text	Required	Input	The following options: (Diabetes Mellitus, Chronic Kidney Disease, Coronary Heart Disease, Asthma, Thyroid Cancer, Schizophrenia, Glaucoma, Alzheimer's Disease, Lung Cancer, Rheumatoid Arthritis, Hypothyroidism, Prostate Cancer, Cervical Cancer, Multiple Sclerosis, Liver Cirrhosis, Chronic Obstructive Pulmonary Disease, Parkinson's Disease, Hepatitis C, Depression, Epileptic Seizure, Osteoporosis and Sickle Cell Anemia).
Training Percentage	Percentage	Required	Input	-
Correctly Classified Instances	Percentage	Required	Output	-

Incorrectly Classified Instances	Percentage	Required	Output	-
----------------------------------	------------	----------	--------	---

Table 28 The Main Fields of the Rebuild and Update Model Tab in the Admin Interface

#### 4.2.3.1.1.7 Medical Specialist Interface

This interface will be displayed if a medical specialist enters a valid username and password. The medical specialist will be able to search for a specific patient, diagnose a patient, view any patient's history, and print the patient's results. Table 29 shows a brief description of the main fields in the Medical Specialist Interface.

Field Name	Format	Level	Input/Output	Comment
MRN	Integer	Required	Input	Unique for hospital
Patient Name	Text	Required	Output	If the patient exists
Test ID	Integer	Required	Input/Output	If the patient exists
Diagnose Date	Date	Required	Output	If the patient exists
Diagnose Type	Text	Required	Output	If the patient exists
Diagnose Result	Text	Required	Output	It will display: (Positive or Negative)

Table 29 The Main Fields of the Medical Specialist Interface

#### 4.2.3.1.1.8 Laboratory Specialist Interface

This interface will be displayed if a laboratory specialist enters a valid username and password. The laboratory specialist will be able to store laboratory and demographical data to the database by filling the laboratory test form and will also be able to perform individual or overall disease diagnosis.

##### ❖ Laboratory Test Tab

In this tab, the laboratory specialist will be able to choose the diseases to be diagnosed for the patient, enter their information, fill in the required fields, and finally store the information in the database and perform the prediction. Table 30 shows a brief description of the main fields in the laboratory Test Tab.

Field Name	Format	Level	Input/Output	Comment
MRN	Integer	Required	Input	Unique for hospital
Disease To be Diagnosed	Check box	Required	Input	The following options: (All, Diabetes Mellitus, Chronic Kidney Disease, Coronary Heart Disease, Asthma, Thyroid Cancer, Schizophrenia, Glaucoma, Alzheimer's Disease, Lung Cancer,

				Rheumatoid Arthritis, Hypothyroidism, Prostate Cancer, Cervical Cancer, Multiple Sclerosis, Liver Cirrhosis, Chronic Obstructive Pulmonary Disease, Parkinson's Disease, Hepatitis C, Depression, Epileptic Seizure, Osteoporosis and Sickle Cell Anemia).
Demographical Data	Texts, Integers and Combo-boxes	Required	Input	Demographical Data Fields to be filled are shown depending on the chosen options from the field "Disease To be Diagnosed"
Blood Tests	Texts, Integers and Combo-boxes	Required	Input	Blood Tests Fields to be filled are shown depending on the chosen options from the field "Disease To be Diagnosed"
Symptoms	Texts, Integers and Combo-boxes	Required	Input	Symptoms Fields to be filled are shown depending on the chosen options from the field "Disease To be Diagnosed"
Other tests	Texts, Integers and Combo-boxes	Required	Input	Other tests Fields to be filled are shown depending on the chosen options from the field "Disease To be Diagnosed"
Store and predict	Button	Required	Input	Filled information will be stored in the database and according to the chosen disease, predictions will be made.

Table 30 The Main Fields of the Laboratory Specialist Interface

#### 4.2.3.1.9 Registered User Interface

This interface will be displayed if the registered non-medical staff user logs into the system. The registered user will view their medical history, run a medical diagnosis, and view and print their results. Moreover, they can delete their accounts. Table 31 shows a brief description of the main fields in the Registered User Interface.

Field Name	Format	Level	Input/Output	Comment
MRN	Text	Required	Output	-
Test ID	Integer	Required	Output	If the diagnosis history exists
Diagnose Date	Date	Required	Output	If the diagnosis history exists
Diagnose Type	Text	Required	Output	The following options: (Diabetes Mellitus, Chronic Kidney Disease, Coronary Heart Disease, Asthma, Thyroid Cancer, Schizophrenia, Glaucoma, Alzheimer's Disease, Lung Cancer, Rheumatoid Arthritis, Hypothyroidism, Prostate Cancer, Cervical Cancer, Multiple Sclerosis, Liver Cirrhosis, Chronic Obstructive Pulmonary Disease, Parkinson's Disease, Hepatitis C, Depression, Epileptic Seizure, Osteoporosis and Sickle Cell Anemia).
Diagnose Result	Text	Required	Output	It will display: (Positive or Negative)

Table 31 The Main Fields of the Registered User Interface

#### 4.2.3.1.10 Diagnosis Interface

The diagnosis interface allows the user to fill out a form to detect the presence of Diabetes Mellitus, Chronic Kidney Disease, Coronary Heart Disease, Asthma, Thyroid Cancer, Schizophrenia, Glaucoma, Alzheimer's Disease, Lung Cancer, Rheumatoid Arthritis, Hypothyroidism, Prostate Cancer, Cervical Cancer, Multiple Sclerosis, Liver Cirrhosis, Chronic Obstructive Pulmonary Disease, Parkinson's Disease, Hepatitis C, Depression, Epileptic Seizure, Osteoporosis or Sickle Cell Anemia. After entering the requested fields, the

user clicks on the "Diagnose" button to view the results. The interface contains twenty-two tabs, each dedicated to a single disease.

#### ❖ Diagnose Diabetes Mellitus Tab

This tab is utilized in preemptively diagnosing Diabetes. Table 32 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Sugar Level	Decimal	Required	Input	-
Hematocrit Level	Decimal	Required	Input	-
Mean Platelet Volume (MPV)	Decimal	Required	Input	-

Table 32 The Main Fields to Diagnose Diabetes Mellitus

#### ❖ Diagnose Chronic Kidney Disease Tab

This tab is utilized in preemptively diagnosing Chronic Kidney Disease (CKD). Table 33 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Blood Urea Nitrogen	Decimal	Required	Input	-
Creatinine	Decimal	Required	Input	-

Table 33 The Main Fields to Diagnose Chronic Kidney Disease

#### ❖ Diagnose Coronary Heart Disease Tab

This tab is utilized in preemptively diagnosing Coronary Heart Disease (CHD). Table 34 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Gender	Drop-down list	Required	Input	-
Age	Integer	Required	Input	-
Mean Corpuscular Hemoglobin (MCH)	Decimal	Required	Input	-
Mean Corpuscular Hemoglobin Concentration (MCHC)	Decimal	Required	Input	-
Red Cell Distribution Width (RDW)	Decimal	Required	Input	-
Platelet Count	Decimal	Required	Input	-
MPV	Decimal	Required	Input	-
Hemoglobin	Decimal	Required	Input	-

Neutrophil Granulocyte Instrument	Decimal	Required	Input	-
Basophil Instrument %	Decimal	Required	Input	-
Basophil Instrument Absolute	Decimal	Required	Input	-
Neutrophil Granulocyte Instrument Absolute	Decimal	Required	Input	-
Mononucleosis Absolute	Decimal	Required	Input	-
Potassium	Decimal	Required	Input	-
Anion Gap	Decimal	Required	Input	-
Gamma-glutamyl transpeptidase (GGTP)	Decimal	Required	Input	-
Serum glutamic-oxaloacetic transaminase (SGOT)	Decimal	Required	Input	-
Serum glutamic-pyruvic transaminase (SGPT)	Decimal	Required	Input	-

Table 34 The Main Fields to Diagnose Coronary Heart Disease

#### ❖ Diagnose Asthma Disease Tab

This tab is utilized in preemptively diagnosing Asthma Disease. Table 35 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Gender	Drop-down list	Required	Input	-
Age	Integer	Required	Input	-
Basophil Instrument %	Decimal	Required	Input	-
Hematocrit	Decimal	Required	Input	-
Hemoglobin	Decimal	Required	Input	-
MCH	Decimal	Required	Input	-
MCHC	Decimal	Required	Input	-
MPV	Decimal	Required	Input	-
White Blood Cell count	Decimal	Required	Input	-

Table 35 The Main Fields to Diagnose Asthma Disease

### ❖ Diagnose Thyroid Cancer Tab

This tab is utilized in preemptively diagnosing Thyroid Cancer. Table 36 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Gender	Drop-down list	Required	Input	-
Age	Integer	Required	Input	-
Hematocrit	Decimal	Required	Input	-
MCHC	Decimal	Required	Input	-
MPV	Decimal	Required	Input	-
Red Blood Cell count	Decimal	Required	Input	-
White Blood Cell count	Decimal	Required	Input	-

Table 36 The Main Fields to Diagnose Thyroid Cancer

### ❖ Diagnose Schizophrenia Tab

This tab is utilized in preemptively diagnosing Schizophrenia. Table 37 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Age	Integer	Required	Input	-
White Blood Cell	Decimal	Required	Input	-
Hemoglobin	Decimal	Required	Input	-
Hematocrit	Decimal	Required	Input	-
Mean Corpuscular Volume (MCV)	Decimal	Required	Input	-
MCH	Decimal	Required	Input	-
MCHC	Decimal	Required	Input	-
Platelet	Decimal	Required	Input	-
MPV	Decimal	Required	Input	-
Aspartate Aminotransferase (AST)	Decimal	Required	Input	-
Total Protein	Decimal	Required	Input	-
Gamma Glutamyl transferase (GGT)	Decimal	Required	Input	-

Table 37 The Main Fields to Diagnose Schizophrenia

### ❖ Diagnose Glaucoma Tab

This tab is utilized in preemptively diagnosing Glaucoma. Table 38 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
At	Decimal	Required	Input	-
Ean	Decimal	Required	Input	-
Mhci	Decimal	Required	Input	-
Vasi	Decimal	Required	Input	-
Varg	Decimal	Required	Input	-
Vars	Decimal	Required	Input	-
Tmi	Decimal	Required	Input	-

Table 38 The Main Fields to Diagnose Glaucoma

#### ❖ Diagnose Alzheimer's Disease Tab

This tab is utilized in preemptively diagnosing Alzheimer's Disease. Table 39 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Gender	Drop-down list	Required	Input	-
Age	Integer	Required	Input	-
Pulse Ox	Decimal	Required	Input	-
Respiratory Rate	Decimal	Required	Input	-
BP - Diastolic	Decimal	Required	Input	-
White Blood Cell count	Decimal	Required	Input	-
Red Blood Cell count	Decimal	Required	Input	-
Hemoglobin	Decimal	Required	Input	-
Hematocrit	Decimal	Required	Input	-
MCV	Decimal	Required	Input	-
MCH	Decimal	Required	Input	-
RDW	Decimal	Required	Input	-
MPV	Decimal	Required	Input	-

Table 39 The Main Fields to Diagnose Alzheimer's Disease

#### ❖ Diagnose Lung Cancer Tab

This tab is utilized in preemptively diagnosing Lung Cancer. Table 40 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Gender	Drop-down list	Required	Input	-
Age	Integer	Required	Input	-
Smoking	Drop-down list	Required	Input	-
Yellow Fingers	Drop-down list	Required	Input	-
Anxiety	Drop-down list	Required	Input	-

Wheezing	Drop-down list	Required	Input	-
Peer Pressure	Decimal	Required	Input	-
Chronic Disease	Drop-down list	Required	Input	-
Fatigue	Drop-down list	Required	Input	-
Allergy	Drop-down list	Required	Input	-
Coughing	Drop-down list	Required	Input	-
Alcohol	Drop-down list	Required	Input	-
Shortness Of Breath	Drop-down list	Required	Input	-
Swallowing Difficulty	Drop-down list	Required	Input	-
Chest Pain	Drop-down list	Required	Input	-

Table 40 The Main Fields to Diagnose Lung Cancer

#### ❖ Diagnose Rheumatoid Arthritis Disease Tab

This tab is utilized in preemptively diagnosing Rheumatoid Arthritis. Table 41 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Gender	Drop-down list	Required	Input	-
Age	Integer	Required	Input	-
Albumin	Decimal	Required	Input	-
Alkaline phosphatase	Decimal	Required	Input	-
Blood Urea Nitrogen	Decimal	Required	Input	-
Chloride	Decimal	Required	Input	-
Carbon dioxide	Decimal	Required	Input	-
Creatinine	Decimal	Required	Input	-
Direct Bilirubin	Decimal	Required	Input	-
GGTP	Decimal	Required	Input	-
Hemoglobin	Decimal	Required	Input	-
Hematocrit Level	Decimal	Required	Input	-
Potassium	Decimal	Required	Input	-
Lactic Acid Dehydrogenase	Decimal	Required	Input	-
MCH	Decimal	Required	Input	-

MPV	Decimal	Required	Input	-
MCHC	Decimal	Required	Input	-
MCV	Decimal	Required	Input	-
Sodium	Decimal	Required	Input	-
Platelet	Decimal	Required	Input	-
Red Blood Cell Count	Decimal	Required	Input	-
RDW	Decimal	Required	Input	-
SGOT	Decimal	Required	Input	-
SGPT	Decimal	Required	Input	-
Total Bilirubin	Decimal	Required	Input	-
Total Protein	Decimal	Required	Input	-

Table 41 The Main Fields to Diagnose Rheumatoid Arthritis

#### ❖ Diagnose Hypothyroidism Tab

This tab is utilized in preemptively diagnosing Hypothyroidism. Table 42 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Age	Integer	Required	Input	-
BP - Systolic	Decimal	Required	Input	-
Respiratory Rate	Decimal	Required	Input	-
MCV	Decimal	Required	Input	-
Pulse Ox	Decimal	Required	Input	-

Table 42 The Main Fields to Diagnose Hypothyroidism

#### ❖ Diagnose Prostate Cancer Tab

This tab is utilized in preemptively diagnosing Prostate Cancer. Table 43 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Perimeter	Decimal	Required	Input	-
Area	Decimal	Required	Input	-
Smoothness	Decimal	Required	Input	-
Compactness	Decimal	Required	Input	-

Table 43 The Main Fields to Diagnose Prostate Cancer

#### ❖ Diagnose Cervical Cancer Tab

This tab is utilized in preemptively diagnosing Cervical Cancer. Table 44 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
STDs: Number of Diagnosis	Decimal	Required	Input	-

STDs Condylomatosis	Drop-down List	Required	Input	-
STDs Syphilis	Drop-down List	Required	Input	-
STDs HIV	Drop-down List	Required	Input	-
STDs HPV	Drop-down List	Required	Input	-
Dx	Drop-down List	Required	Input	-
Dx: CIN	Drop-down List	Required	Input	-
Dx: HPV	Drop-down List	Required	Input	-

Table 44 The Main Fields to Diagnose Cervical Cancer

#### ❖ Diagnose Multiple Sclerosis Tab

This tab is utilized in preemptively diagnosing Multiple Sclerosis. Table 45 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Age	Integer	Required	Input	-
Alanine Transaminase (ALT)	Decimal	Required	Input	-
Lactate Dehydrogenase (LDH)	Decimal	Required	Input	-
Creatinine	Decimal	Required	Input	-
Blood Urea Nitrogen	Decimal	Required	Input	-
Total Bilirubin	Decimal	Required	Input	-
GGT	Decimal	Required	Input	-
Alkaline Phosphatase	Decimal	Required	Input	-
AST	Decimal	Required	Input	-
Platelet	Decimal	Required	Input	-
BP - Systolic	Decimal	Required	Input	-

Table 45 The Main Fields to Diagnose Multiple Sclerosis

#### ❖ Diagnose Liver Cirrhosis Tab

This tab is utilized in preemptively diagnosing Liver Cirrhosis. Table 46 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Gender	Drop-down list	Required	Input	-

Age	Integer	Required	Input	-
N_Days	Decimal	Required	Input	-
Hepatomegaly	Drop-down list	Required	Input	-
Spiders	Drop-down list	Required	Input	-
Edema	Decimal	Required	Input	-
Cholesterol	Decimal	Required	Input	-
Copper	Decimal	Required	Input	-
SGOT	Decimal	Required	Input	-
Platelet	Decimal	Required	Input	-
Prothrombin	Decimal	Required	Input	-
Ascites	Drop-down list	Required	Input	-
Serum Bilirubin	Decimal	Required	Input	-
Albumin	Decimal	Required	Input	-
Alkaline phosphatase	Decimal	Required	Input	-
Triglycerides	Decimal	Required	Input	-
Drug	Decimal	Required	Input	-
Status	Drop-down list	Required	Input	-

Table 46 The Main Fields to Diagnose Liver Cirrhosis

#### ❖ Diagnose Chronic Obstructive Pulmonary Disease Tab

This tab is utilized in preemptively diagnosing Chronic Obstructive Pulmonary. Table 47 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Gender	Drop-down list	Required	Input	-
Age	Integer	Required	Input	-
Smoking	Decimal	Required	Input	-
Imagery part minimum	Decimal	Required	Input	-
Imagery part average	Decimal	Required	Input	-
Real part minimum	Decimal	Required	Input	-
Real part average	Decimal	Required	Input	-

Table 47 The Main Fields to Diagnose Chronic Obstructive Pulmonary Disease

#### ❖ Diagnose Parkinson's Disease Tab

This tab is utilized in preemptively diagnosing Parkinson's Disease. Table 48 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment

Gender	Drop-down list	Required	Input	-
Age	Integer	Required	Input	-
Anion Gap	Decimal	Required	Input	-
ALT	Decimal	Required	Input	-
LDH	Decimal	Required	Input	-
White Blood Cells	Decimal	Required	Input	-
Red Blood Cells	Decimal	Required	Input	-
Hemoglobin	Decimal	Required	Input	-
Hematocrit	Decimal	Required	Input	-
Sodium	Decimal	Required	Input	-
Potassium	Decimal	Required	Input	-
Chloride	Decimal	Required	Input	-
Carbon Dioxide	Decimal	Required	Input	-
Creatinine	Decimal	Required	Input	-
Total Protein	Decimal	Required	Input	-
Albumin	Decimal	Required	Input	-
Blood Urea Nitrogen	Decimal	Required	Input	-
Total Bilirubin	Decimal	Required	Input	-
Direct Bilirubin	Decimal	Required	Input	-
GGT	Decimal	Required	Input	-
MCV	Decimal	Required	Input	-
MCH	Decimal	Required	Input	-
MCHC	Decimal	Required	Input	-
Alkaline Phosphatase	Decimal	Required	Input	-
RDW	Decimal	Required	Input	-
AST	Decimal	Required	Input	-

Table 48 The Main Fields to Diagnose Parkinson's Disease

#### ❖ Diagnose Hepatitis C Tab

This tab's purpose is to detect Hepatitis C. Table 49 illustrates the fields demanded from the user to diagnose Hepatitis C.

Field Name	Format	Level	Input/output	Comment
Age	Integer	Required	Input	-
Total Protein	Decimal	Required	Input	-
Total Bilirubin	Decimal	Required	Input	-
Direct Bilirubin	Decimal	Required	Input	-
GGT	Decimal	Required	Input	-
Alkaline Phosphatase	Decimal	Required	Input	-
Lymphocyte - Instrument %	Decimal	Required	Input	-

Neutrophil Granulocyte - Instrument Absolute	Decimal	Required	Input	-
Platelet	Decimal	Required	Input	-
Basophil - Instrument %	Decimal	Required	Input	-
BP – Systolic	Decimal	Required	Input	-
Fall Risk - Morse	Decimal	Required	Input	-
Body Mass Index	Decimal	Required	Input	-
International Normalized Ratio	Decimal	Required	Input	-

Table 49 The Main Fields to Diagnose Hepatitis C

#### ❖ Diagnose Depression Tab

Diagnosing Depression can be done using this tab. Table 50 presents the demanded fields to detect this disease.

Field Name	Format	Level	Input/output	Comment
Age	Integer	Required	Input	-
Household Size	Integer	Required	Input	-
Education Level	Integer	Required	Input	-
Value of livestock	Integer	Required	Input	-
Value of durable goods	Integer	Required	Input	-
Value of savings	Integer	Required	Input	-
Land owned	Integer	Required	Input	-
Consumed Alcohol	Decimal	Required	Input	-
Consumed Tobacco	Decimal	Required	Input	-
Education expenditure	Decimal	Required	Input	-
Non-ag business flow expenses, monthly	Decimal	Required	Input	-
Livestock sales and meat revenue, monthly	Decimal	Required	Input	-
Total expenses, monthly	Decimal	Required	Input	-
Whole days without food	Decimal	Required	Input	-
Non-durable Investments	Decimal	Required	Input	-
Amount received using M-Pesa	Decimal	Required	Input	-
Marital status	Drop-down list	Required	Input	-
Children	Integer	Required	Input	-
hh_children	Integer	Required	Input	-

Non-agricultural business owner	Drop-down list	Required	Input	-
Saved money using M-Pesa	Drop-down list	Required	Input	-
Early Survey	Drop-down list	Required	Input	-

Table 50 The Main Fields to Diagnose Depression

#### ❖ Diagnose Epileptic Seizure Disease Tab

This tab is utilized in preemptively diagnosing Epileptic Seizure Disease. Table 51 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Gender	Drop-down list	Required	Input	-
NumberOfNonPsych Comorbidities	Integer	Required	Input	-
NumberOfPrior AEDs	Integer	Required	Input	-
Asthma	Drop-down list	Required	Input	-
Migraine	Drop-down list	Required	Input	-
Chronic Pain	Drop-down list	Required	Input	-
Diabetes	Drop-down list	Required	Input	-
non metastatic cancer	Drop-down list	Required	Input	-
NumberOfNonSeizureNonPsych Medication	Integer	Required	Input	-
NumberOfCurrent AEDs	Integer	Required	Input	-
Baseline	Decimal	Required	Input	-
MedianDurationOfSeizures	Decimal	Required	Input	-
NumberOfSeizureTypes	Integer	Required	Input	-
InjuryWithSeizure	Drop-down list	Required	Input	-
Catamenial	Drop-down list	Required	Input	-
Trigger of sleep deprivation	Drop-down list	Required	Input	-
Aura	Drop-down list	Required	Input	-

IctalEyeClosure	Drop-down list	Required	Input	-
IctalHallucinations	Drop-down list	Required	Input	-
Oralautomatisms	Drop-down list	Required	Input	-
Incontinence	Drop-down list	Required	Input	-
LimbAutomatisms	Drop-down list	Required	Input	-
IctalTonic-clonic	Drop-down list	Required	Input	-
MuscleTwitching	Drop-down list	Required	Input	-
HipThrusting	Drop-down list	Required	Input	-
Post-ictalFatigue	Drop-down list	Required	Input	-
HeadInjury	Drop-down list	Required	Input	-
PsychTraumaticEvents	Drop-down list	Required	Input	-
Concussionw/oLOC	Drop-down list	Required	Input	-
Concussionw/LOC	Drop-down list	Required	Input	-
SevereTBI	Drop-down list	Required	Input	-
Opioids	Drop-down list	Required	Input	-
SexAbuse	Drop-down list	Required	Input	-
PhysicalAbuse	Drop-down list	Required	Input	-
Rape	Drop-down list	Required	Input	-

Table 51 The Main Fields to Epileptic Seizure

#### ❖ Diagnose Osteoporosis Disease Tab

This tab is utilized in preemptively diagnosing Osteoporosis Disease. Table 52 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
------------	--------	-------	--------------	---------

Gender	Drop-down list	Required	Input	-
Age	Integer	Required	Input	-
Weight	Integer	Required	Input	-
Hight	Integer	Required	Input	-
Diabest	Drop-down list	Required	Input	-
Hypothyroidism	Drop-down list	Required	Input	-
SeizerDisorder	Drop-down list	Required	Input	-
Alcohol	Drop-down list	Required	Input	-
Smoking	Drop-down list	Required	Input	-
EstrogenUse	Drop-down list	Required	Input	-
JointPain	Drop-down list	Required	Input	-
HistoryOfFracture	Drop-down list	Required	Input	-
Dialysis	Drop-down list	Required	Input	-
Family History of Osteoporosis	Drop-down list	Required	Input	-
Maximum Walking distance	integer	Required	Input	-
Daily Eating habits	Drop-down list	Required	Input	-
BMI	Decimal	Required	Input	-
Site	Drop-down list	Required	Input	-
Obesity	Drop-down list	Required	Input	-

Table 52 The Main Fields to Diagnose Osteoporosis

#### ❖ Diagnose Sickle Cell Anemia Disease Tab

This tab is utilized in preemptively diagnosing Sickle Cell Anemia Disease. Table 53 shows the required fields to perform the diagnosis procedure.

Field Name	Format	Level	Input/output	Comment
Sex	Drop-down list	Required	Input	-

Tribe	Decimal	Required	Input	-
Hemoglobin (HB)	Decimal	Required	Input	
Packed cell volume (PCV)	Decimal	Required	Input	-
Red Blood Cells (RBCs)	Decimal	Required	Input	-
Mean Cell Volume (MCV)	Decimal	Required	Input	-
Mean Cell Hemoglobin (MCH)	Decimal	Required	Input	-
Mean Cell Hemoglobin Concentration (MCHC)	Decimal	Required	Input	-
Total White Blood Cells (TWBCs)	Decimal	Required	Input	-
Platelet Counts (PLTs)	Decimal	Required	Input	-

Table 53 The Main Fields to Diagnose Sickle Cell Anemia

Table 54 shows the additional fields needed from the laboratory specialist to use the interface. A laboratory specialist clicks on the “Retrieve” button to check if the patient exists before filling all features needed to diagnose.

Field Name	Format	Level	Input/Output	Comment
MRN	Integer	Required	Input/Output	-

Table 54 The Main Fields in the Diagnosis Interface

#### 4.2.3.1.1.11 Result Interface

The Result Interface will exhibit the diagnosis results of the intended diseases when the medical specialists, registered user, or the guest click on the “Diagnosis” button. The fields of this interface are presented in Table 55.

Field Name	Format	Level	Input/Output	Comment
Result	Text	Required	Output	Unique
Disease	Text	Required	Output	Diabetes Mellitus, Chronic Kidney Disease, Coronary Heart Disease, Asthma, Thyroid Cancer, Schizophrenia, Glaucoma, Alzheimer's Disease, Lung Cancer, Rheumatoid Arthritis, Hypothyroidism, Prostate Cancer, Cervical Cancer, Multiple Sclerosis, Liver Cirrhosis, Chronic Obstructive Pulmonary Disease, Parkinson's Disease, Hepatitis C, Depression, Epileptic Seizure, Osteoporosis and Sickle Cell Anemia.

Accuracy	Decimal	Required	Output	-
----------	---------	----------	--------	---

Table 55 The Main Field of The Result Interface

#### 4.2.3.1.2 Hardware Interface

The intended system does not require any hardware interfaces, as it does not communicate directly with any hardware. Alternatively, the system interacts with the underlying operating system and is runnable on any operating system and processor.

#### 4.2.3.1.3 Software Interface

The requirements for the software interfaces include:

- The system is required to run on Windows and macOS X.
- The system must be able to read and write information from MongoDB DBMS.
- The system must be able to read CSV file extensions.
- The system must be able to read DICOM file extensions.

#### 4.2.3.1.4 Communication Interface

The system communicates with the database using MongoDB Connector/Python, a Python driver, to communicate with MongoDB servers. The system will send an automatic verification email to the user who creates an account in the system.

### 4.2.3.2 Functional Requirements

Functional requirements are classified based on the users' roles: admin, medical specialists, laboratory specialists, registered users, and guest users. In addition, there are common functionalities among all these users.

#### 4.2.3.2.1 Common Functionalities

Many functions are shared among all users, including logging into the system, retrieving the password, updating the profile, viewing the diagnosis history, diagnosing, and printing result.

##### 4.2.3.2.1.1 Login

Table 56 shows the “Login” functionality.

<b>Actors</b>	<ul style="list-style-type: none"> <li>• Admins.</li> <li>• Medical specialists.</li> <li>• Laboratory specialists.</li> <li>• Registered users.</li> <li>• Guest.</li> </ul>
<b>Description</b>	<ul style="list-style-type: none"> <li>• Admins, medical specialists, laboratory specialists, and registered users should log in to the system by entering their username and password.</li> <li>• Guests should be able to access the system without a username or password.</li> </ul>

	<ul style="list-style-type: none"> <li>The system validates the entered information based on the data stored in the database.</li> <li>A temporary password will be provided if the user is newly registered to the system or has forgotten the password.</li> <li>If the user logs in with a temporary password before the password expires, the user should be redirected to the Change Password interface. However, if the temporary password is expired, the user will be redirected to the Retrieve Password interface to create a new temporary password.</li> <li>The system should redirect the user to the Retrieve Password interface for the third invalid login.</li> </ul>
<b>Data</b>	<ul style="list-style-type: none"> <li>Username.</li> <li>Password.</li> </ul>
<b>Stimulus</b>	The user command will be issued by clicking on the Login button.
<b>Response</b>	<ul style="list-style-type: none"> <li>The homepage shows whether the user has logged in successfully using the long-term password.</li> <li>If any user accesses the system with a temporary password, they will be redirected to the Change Password interface.</li> </ul>
<b>Abnormal conditions</b>	<ul style="list-style-type: none"> <li>Invalid entries for either the username or password or for both the username and the password.</li> <li>The user should be redirected to the Retrieve Password interface for the third invalid login.</li> </ul>

Table 56 Login Functionality

Figure 10 demonstrates the activity diagram of the “Login” functionality. the activity diagram of the “Login” functionality.

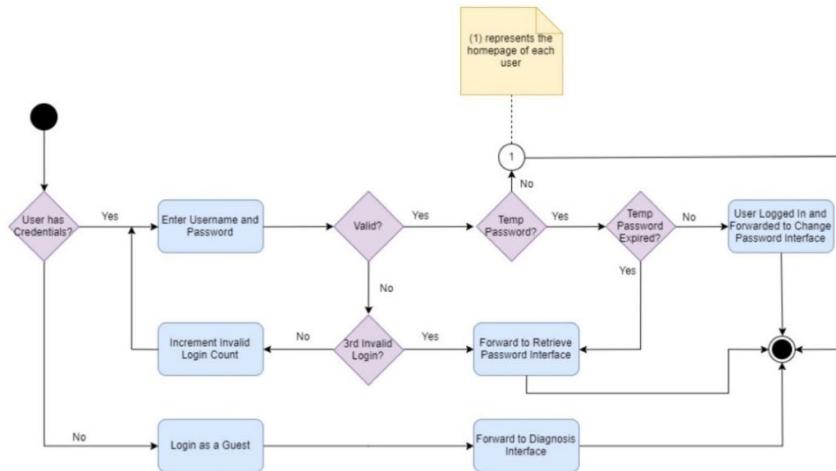


Figure 10 The Activity Diagram of the Login Functionality

#### 4.2.3.2.1.2 Retrieve Password

Table 57 shows the “Retrieve Password” functionality where users can retrieve their forgotten passwords.

<b>Actors</b>	<ul style="list-style-type: none"> <li>Admins.</li> </ul>
---------------	---

	<ul style="list-style-type: none"> <li>Medical specialists.</li> <li>Laboratory specialists.</li> <li>Registered users.</li> </ul>
Description	<ul style="list-style-type: none"> <li>If the users forgot their password, they could reset it by entering the registered email. The system sends an email containing a temporary random password to that existing email.</li> <li>The user should communicate with the admin for the third invalid entry of the email.</li> <li>The user can log into the system using the temporary password.</li> </ul>
Data	<ul style="list-style-type: none"> <li>Email.</li> <li>Temporary Password.</li> </ul>
Stimulus	The user command is issued once the user enters a valid email then clicks on the “Send” button.
Response	<ul style="list-style-type: none"> <li>A temporary password will be stored in the database. After that, it will be sent to the user via email.</li> </ul>
Abnormal conditions	<ul style="list-style-type: none"> <li>If an invalid email format is typed.</li> <li>If the user enters an nonexistent email.</li> </ul>

Table 57 Retrieve Password Functionality

Figure 11 demonstrates the activity diagram of the “Retrieve Password” functionality.

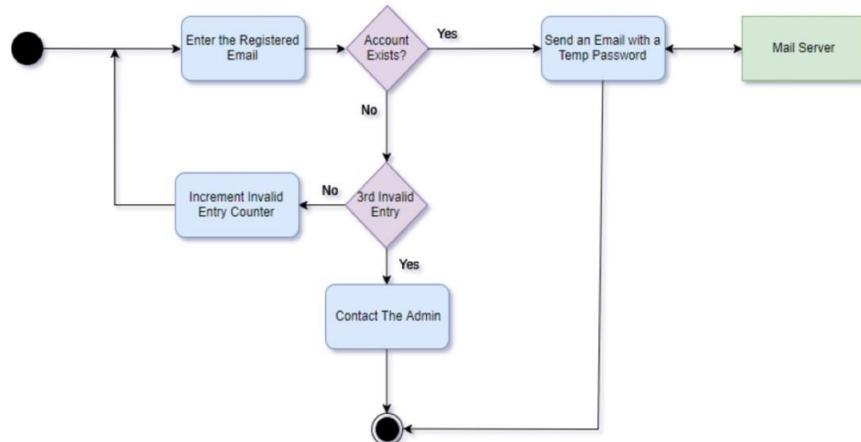


Figure 11 The Activity Diagram of the Retrieve Password Functionality

#### 4.2.3.2.1.3 Update Profile

Table 58 shows the “Update Profile” functionality where users can view their personal information and update their profile.

Actors	<ul style="list-style-type: none"> <li>Admins.</li> <li>Medical specialists.</li> <li>Laboratory specialists.</li> <li>Registered users.</li> </ul>
--------	---

<b>Description</b>	<ul style="list-style-type: none"> <li>Users can update their profiles by changing their email or password. Registered users can also change their name, birth date, and gender.</li> <li>The system verifies the validity of the recently entered information.</li> </ul>
<b>Data</b>	<p>All actors can change their:</p> <ul style="list-style-type: none"> <li>Email.</li> <li>Password.</li> </ul> <p>Registered users can also change their:</p> <ul style="list-style-type: none"> <li>Name.</li> <li>Birthdate.</li> <li>Gender.</li> </ul>
<b>Stimulus</b>	<ul style="list-style-type: none"> <li>The user command can be issued by clicking on the Change Email or Change Password buttons and submitting a new email or password.</li> <li>For registered users, a user command can be issued by clicking on the “Save Changes” button after entering a new name, birthdate, or gender.</li> </ul>
<b>Response</b>	Changes will be updated if all the fields are validated successfully.
<b>Abnormal conditions</b>	<ul style="list-style-type: none"> <li>If the current email is invalid.</li> <li>If the new email format is invalid.</li> <li>If the new email and confirm email do not match.</li> <li>If the current password is invalid.</li> <li>If the new password and confirm password do not match.</li> <li>If the new password does not follow the password constraints (at least eight alphanumeric characters).</li> <li>If the newly entered name includes numeric digits.</li> </ul>

Table 58 Update Profile Functionality

Figure 12 demonstrates the activity diagram of the “Update Profile” functionality.

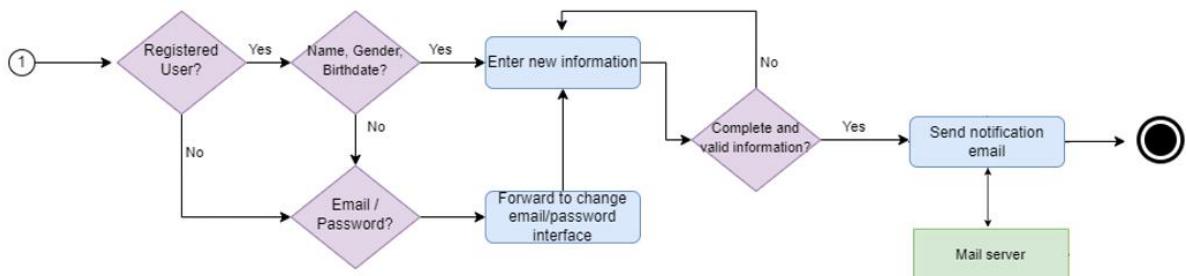


Figure 12 The Activity Diagram of the Update Profile Functionality

#### 4.2.3.2.1.4 View Diagnosis History

Table 59 illustrates the “View Diagnosis History” functionality. It enables medical specialists and registered users to view their diagnosis history.

<b>Actors</b>	<ul style="list-style-type: none"> <li>Medical specialist.</li> <li>Registered user.</li> </ul>
<b>Description</b>	The results of previous diagnoses must be accessible to medical specialists and registered users.
<b>Data</b>	<p>For medical specialists:</p> <ul style="list-style-type: none"> <li>MRN.</li> </ul> <p>For registered users:</p> <ul style="list-style-type: none"> <li>None.</li> </ul>
<b>Stimulus</b>	<p>For medical specialists:</p> <ul style="list-style-type: none"> <li>The user command is issued by entering the patient's MRN and clicking search.</li> </ul> <p>For registered users:</p> <ul style="list-style-type: none"> <li>The user command is issued by clicking diagnose history.</li> </ul>
<b>Response</b>	<p>For medical specialists:</p> <ul style="list-style-type: none"> <li>The previous diagnosis results of the intended patient will be displayed if any record was found in the database.</li> </ul> <p>For registered users:</p> <ul style="list-style-type: none"> <li>The previous diagnosis history will be retrieved from the database if it exists.</li> </ul>
<b>Abnormal conditions</b>	<p>For medical specialists:</p> <ul style="list-style-type: none"> <li>Invalid MRN.</li> </ul> <p>For registered users:</p> <ul style="list-style-type: none"> <li>None.</li> </ul>

Table 59 View Diagnosis History Functionality

Figure 13 shows the activity diagram of “View Patient’s History” functionality.

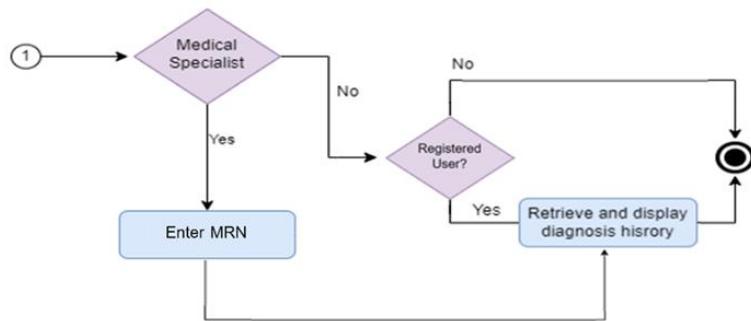


Figure 13 The Activity Diagram of the View History Functionality

#### 4.2.3.2.1.5 Diagnose

Table 60 shows the “Diagnose” functionality.

Actors	<ul style="list-style-type: none"> <li>• Medical specialists.</li> <li>• Laboratory Specialists.</li> <li>• Registered users.</li> <li>• Guests.</li> </ul>
Description	<ul style="list-style-type: none"> <li>• Medical specialists should be able to diagnose patients with any of the investigated diseases.</li> <li>• Registered users and guest users must be able to perform the diagnosis process themselves.</li> <li>• The guest users can use the system to diagnose themselves without registering with a username or a password.</li> <li>• The data entered by the laboratory specialists is stored in the database.</li> <li>• The diagnosis result will be stored along with the model accuracy.</li> </ul>
Data	<p>For diagnosing Diabetes Mellitus:</p> <ul style="list-style-type: none"> <li>• Sugar Level</li> <li>• Hematocrit Level</li> <li>• Mean Platelet Volume (MPV)</li> </ul> <p>For diagnosing CKD:</p> <ul style="list-style-type: none"> <li>• Blood Urea Nitrogen</li> <li>• Creatinine</li> </ul> <p>For diagnosing CHD:</p> <ul style="list-style-type: none"> <li>• Gender</li> <li>• Age</li> <li>• Mean Corpuscular Hemoglobin (MCH)</li> <li>• Mean Corpuscular Hemoglobin Concentration (MCHC)</li> <li>• Red Cell Distribution Width (RDW)</li> <li>• Platelet Count</li> <li>• MPV</li> <li>• Hemoglobin</li> <li>• Neutrophil Granulocyte Instrument</li> <li>• Basophil Instrument %</li> <li>• Basophil Instrument Absolute</li> <li>• Neutrophil Granulocyte Instrument Absolute</li> <li>• Mononucleosis Absolute</li> <li>• Potassium</li> <li>• Anion Gap</li> <li>• Gamma-glutamyl transpeptidase (GGTP)</li> <li>• Serum glutamic-oxaloacetic transaminase (SGOT)</li> <li>• Serum glutamic-pyruvic transaminase (SGPT)</li> </ul> <p>For diagnosing Asthma Disease:</p> <ul style="list-style-type: none"> <li>• Gender</li> </ul>

- Age
- Basophil Instrument %
- Hematocrit
- Hemoglobin
- MCH
- MCHC
- MPV
- White Blood Cell count

For diagnosing Thyroid Cancer:

- Gender
- Age
- Hematocrit
- MCHC
- MPV
- Red Blood Cell count
- White Blood Cell count

For diagnosing Schizophrenia:

- Age
- White Blood Cell
- Hemoglobin
- Hematocrit
- Mean Corpuscular Volume (MCV)
- MCH
- MCHC
- Platelet
- MPV
- Aspartate Aminotransferase (AST)
- Total Protein
- Gamma Glutamyl transferase (GGT)

For diagnosing Glaucoma:

- At
- Ean
- Mhci
- Vasi
- Varg
- Vars
- Tmi

For diagnosing Alzheimer's Disease:

- Gender
- Age

- Pulse Ox
- Respiratory Rate
- BP - Diastolic
- White Blood Cell count
- Red Blood Cell count
- Hemoglobin
- Hematocrit
- MCV
- MCH
- RDW
- MPV

For diagnosing Lung Cancer:

- Gender
- Age
- Smoking
- Yellow Fingers
- Anxiety
- Wheezing
- Peer Pressure
- Chronic Disease
- Fatigue
- Allergy
- Coughing
- Alcohol
- Shortness Of Breath
- Swallowing Difficulty
- Chest Pain

For diagnosing Rheumatoid Arthritis:

- Gender
- Age
- Albumin
- Alkaline phosphatase
- Blood Urea Nitrogen
- Chloride
- Carbon dioxide
- Creatinine
- Direct Bilirubin
- GGTP
- Hemoglobin
- Hematocrit Level
- Potassium

- Lactic Acid Dehydrogenase
- MCH
- MPV
- MCHC
- MCV
- Sodium
- Platelet
- Red Blood Cell Count
- RDW
- SGOT
- SGPT
- Total Bilirubin
- Total Protein

For diagnosing Hypothyroidism:

- Age
- BP - Systolic
- Respiratory Rate
- MCV
- Pulse Ox

For diagnosing Prostate Cancer:

- Perimeter
- Area
- Smoothness
- Compactness

For diagnosing Cervical Cancer:

- STDs: Number of Diagnosis
- STDs Condylomatosis
- STDs Syphilis
- STDs HIV
- STDs HPV
- Dx
- Dx: CIN
- Dx: HPV

For diagnosing Multiple Sclerosis:

- Age
- Alanine Transaminase (ALT)
- Lactate Dehydrogenase (LDH)
- Creatinine
- Blood Urea Nitrogen

- Total Bilirubin
- GGT
- Alkaline Phosphatase
- AST
- Platelet
- BP – Systolic

For diagnosing Liver Cirrhosis:

- Gender
- Age
- N\_Days
- Hepatomegaly
- Spiders
- Edema
- Cholesterol
- Copper
- SGOT
- Platelet
- Prothrombin
- Ascites
- Serum Bilirubin
- Albumin
- Alkaline phosphatase
- Triglycerides
- Drug
- Status

For diagnosing Chronic Obstructive Pulmonary Disease:

- Gender
- Age
- Smoking
- Imagery part minimum
- Imagery part average
- Real part minimum
- Real part average

For diagnosing Parkinson's Disease:

- Gender
- Age
- Anion Gap
- ALT
- LDH

- White Blood Cells
- Red Blood Cells
- Hemoglobin
- Hematocrit
- Sodium
- Potassium
- Chloride
- Carbon Dioxide
- Creatinine
- Total Protein
- Albumin
- Blood Urea Nitrogen
- Total Bilirubin
- Direct Bilirubin
- GGT
- MCV
- MCH
- MCHC
- Alkaline Phosphatase
- RDW
- AST

For diagnosing Hepatitis C:

- Age
- Total Protein
- Total Bilirubin
- Direct Bilirubin
- GGT
- Alkaline Phosphatase
- Lymphocyte - Instrument %
- Neutrophil Granulocyte - Instrument Absolute
- Platelet
- Basophil - Instrument %
- BP – Systolic
- Fall Risk - Morse
- Body Mass Index
- International Normalized Ratio

For diagnosing Depression:

- Age
- Household Size

- Education Level
- Value of livestock
- Value of durable goods
- Value of savings
- Land owned
- Consumed Alcohol
- Consumed Tobacco
- Education expenditure
- Non-ag business flow expenses, monthly
- Livestock sales and meat revenue, monthly
- Total expenses, monthly
- Whole days without food
- Non-durable Investments
- Amount received using M-Pesa
- Marital status
- Children
- hh\_children
- Non-agricultural business owner
- Saved money using M-Pesa
- Early Survey

For diagnosing Epileptic Seizure:

- NumberOfNonPsych Comorbidities
- NumberOfPrior AEDs
- Asthma
- Migraine
- Chronic Pain
- Diabetes
- non metastatic cancer
- NumberOfNonSeizureNonPsych
- Medication
- NumberOfCurrent AEDs
- Baseline
- MedianDurationOfSeizures
- NumberOfSeizureTypes
- InjuryWithSeizure
- Catamenial
- Trigger of sleep deprivation
- Aura
- IctalEyeClosure
- IctalHallucinations
- Oralautomatisms

- Incontinence
- LimbAutomatisms
- IctalTonic-clonic
- MuscleTwitching
- HipThrusting
- Post-ictalFatigue
- HeadInjury
- PsychTraumaticEvents
- Concussionw/oLOC
- Concussionw/LOC
- SevereTBI
- Opioids
- SexAbuse
- PhysicalAbuse
- Rape

For diagnosing Osteoporosis:

- Gender
- Age
- Weight
- Height
- Diabetes
- Hypothyroidism
- SeizureDisorder
- Alcohol
- Smoking
- EstrogenUse
- JointPain
- HistoryOfFracture
- Dialysis
- Family History of Osteoporosis
- Maximum Walking distance
- Daily Eating habits
- BMI
- Site
- Obesity

For diagnosing Sickle Cell Anemia:

- Sex
- Tribe
- HB
- MCV

	<ul style="list-style-type: none"> <li>• MCH</li> <li>• MCHC</li> <li>• RBC</li> <li>• TWBC</li> <li>• PLT</li> <li>• PCV</li> </ul> <p>For laboratory specialists to diagnose a patient, there are additional data:</p> <ul style="list-style-type: none"> <li>• Patient MRN.</li> </ul>
Stimulus	The user command is issued by clicking on the “Diagnose” button.
Response	<ul style="list-style-type: none"> <li>• The newly entered data is tested using the diagnosis model.</li> <li>• When the medical specialists, registered user, or the guest make a diagnosis, the results of the diagnosis will be displayed.</li> <li>• The results and accuracy of diagnoses are stored in the database by the laboratory specialists, along with the current date.</li> </ul>
Abnormal conditions	Incomplete or invalid information.

Table 60 Diagnose Functionality

Figure 14 illustrates the activity diagram of “Diagnose” functionality.

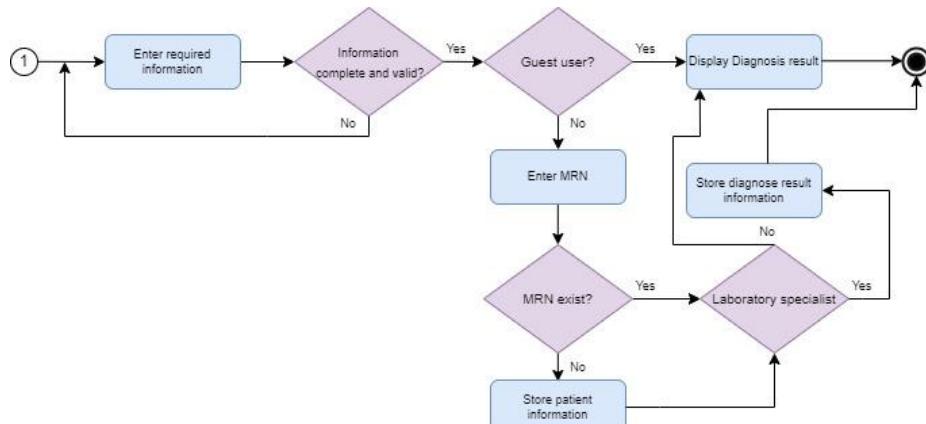


Figure 14 The Activity Diagram of the Diagnose Functionality

#### 4.2.3.2.1.6 Print Result

Users of the system can print their diagnosis results with “Print Result” functionality. Description of “Print Result” functionality is demonstrated in Table 61.

Actors	<ul style="list-style-type: none"> <li>• Medical specialist.</li> <li>• Registered user.</li> <li>• Guest.</li> </ul>
Description	A PDF file will be created and sent to be printed.
Data	Diagnosis result.

<b>Stimulus</b>	The user command is issued by clicking “Print Result” button.
<b>Response</b>	The diagnosis results will be converted to PDF format and sent to the printer.
<b>Abnormal conditions</b>	The printer is incompatible with the system.

Table 61 Print Result Functionality

Figure 15 describes the activity diagram of “Print Result” functionality.

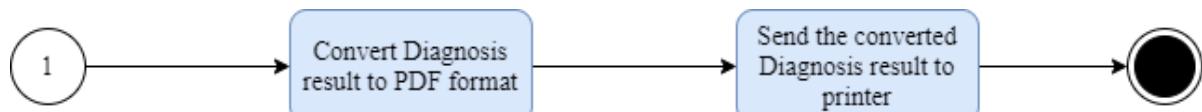


Figure 15 The Activity Diagram of the Print Result Functionality

#### 4.2.3.2.2 User Class 1: Admin

This section describes the admin functionalities.

##### 4.2.3.2.2.1 Rebuild Diagnosis Model

The “Rebuild Diagnosis Model” functionality is demonstrated in Table 62.

<b>Actors</b>	Admin.
<b>Description</b>	The admin can reconstruct the diagnosis model regularly by training the model with the latest patients’ dataset.
<b>Data</b>	<ul style="list-style-type: none"> <li>• Patients’ dataset.</li> <li>• Training set percentage.</li> <li>• Disease type.</li> </ul>
<b>Stimulus</b>	The user command is issued by clicking “Generate Model” button.
<b>Response</b>	<ul style="list-style-type: none"> <li>• Diagnosis model is generated, and its produced accuracy is displayed.</li> <li>• The new diagnosis model’s information is stored.</li> </ul>
<b>Abnormal conditions</b>	The dataset file has an invalid extension.

Table 62 Rebuild Diagnosis Model for Admin Functionality

Figure 16 describes the activity diagram of “Rebuild Diagnosis Model” functionality.

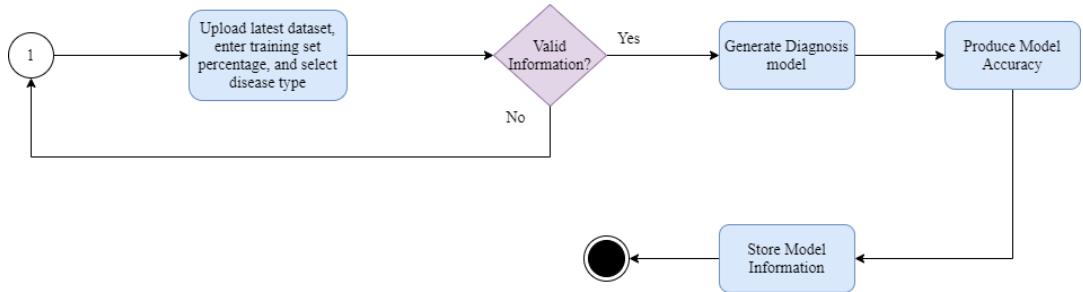


Figure 16 The Activity Diagram of the Rebuild Diagnosis Model Functionality

#### 4.2.3.2.2.2 Replace Diagnosis Model

The “Replace Diagnosis Model” functionality is demonstrated in Table 63.

<b>Actors</b>	Admin.
<b>Description</b>	The admin can replace the previous diagnosis model by uploading the newly constructed diagnosis model with its refined accuracy.
<b>Data</b>	<ul style="list-style-type: none"> <li>• Diagnosis model.</li> <li>• Model accuracy.</li> </ul>
<b>Stimulus</b>	The user command is issued by uploading the new model and clicking “Update Diagnostic Model” button.
<b>Response</b>	The future diagnosis processes will rely on the updated model.
<b>Abnormal conditions</b>	The dataset file has an invalid extension.

Table 63 Replace Diagnosis Model Functionality

Figure 17 describes the activity diagram of “Replace Diagnosis Model” functionality.

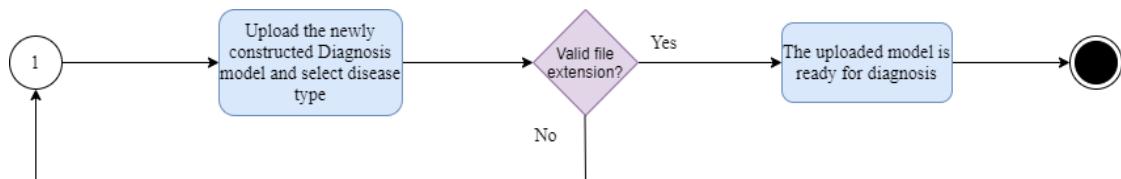


Figure 17 The Activity Diagram of the Replace Diagnosis Model Functionality

#### 4.2.3.2.2.3 Add User

The “Add User” functionality is displayed in Table 64.

<b>Actors</b>	Admin
<b>Description</b>	<ul style="list-style-type: none"> <li>• The admin is allowed to add users including new admins, medical specialists, and laboratory specialists to the system.</li> <li>• A temporary username and password generated by the admin are used to add users.</li> </ul>

	<ul style="list-style-type: none"> <li>The temporary username and password are sent via email to the intended user.</li> </ul>
<b>Data</b>	<ul style="list-style-type: none"> <li>Username.</li> <li>Initial Password.</li> <li>Role.</li> <li>Email.</li> </ul>
<b>Stimulus</b>	The user command is issued after filling all the required information and clicking on “Add” button.
<b>Response</b>	<ul style="list-style-type: none"> <li>The username and password are provided to the new user via an email message.</li> <li>A database entry for the user is made.</li> </ul>
<b>Abnormal conditions</b>	<ul style="list-style-type: none"> <li>Improper formatting of emails.</li> <li>Insufficient or incomplete entry information.</li> </ul>

Table 64 Add User Functionality

Figure 18 presents the activity diagram of “Add User” functionality.

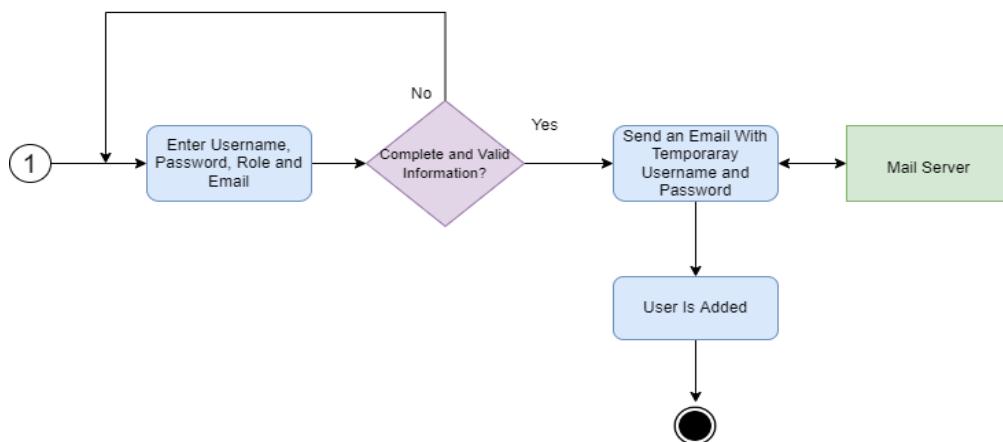


Figure 18 The Activity Diagram of the Add User Functionality

#### 4.2.3.2.2.4 Remove User

Table 65 displays the description of “Remove User” functionality.

<b>Actors</b>	Admin.
<b>Description</b>	<ul style="list-style-type: none"> <li>The admin is allowed to remove any user from the system.</li> <li>When a user is removed all their records will be deleted.</li> </ul>
<b>Data</b>	<ul style="list-style-type: none"> <li>Username.</li> <li>Email.</li> <li>Role.</li> </ul>
<b>Stimulus</b>	The user command is issued by clicking “Delete” button after filling in the required information.

<b>Response</b>	To confirm the deletion of the user's account, a confirmation message is shown, and an email is sent to the user. Additionally, all the data related to the user will be removed from the database
<b>Abnormal conditions</b>	<ul style="list-style-type: none"> <li>An invalid username is entered.</li> <li>An invalid email is entered.</li> <li>An invalid password is entered.</li> </ul>

Table 65 Remove User Functionality

Figure 19 shows the activity diagram of “Remove User” functionality.

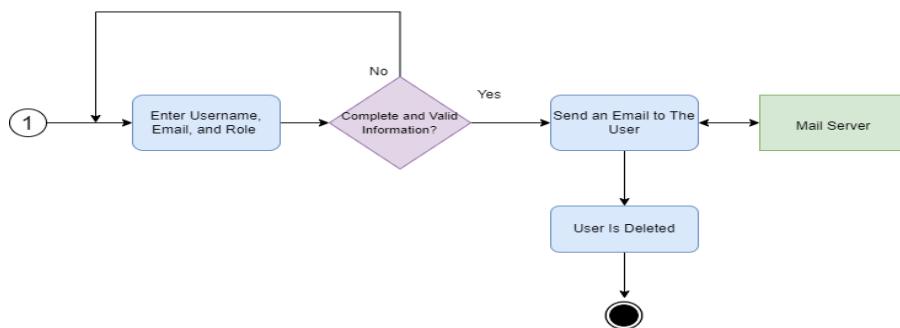


Figure 19 The Activity Diagram of the Remove User Functionality

#### 4.2.3.2.3 User Class 2: Medical Specialists

This section describes the functionalities of the medical specialists.

##### 4.2.3.2.3.1 View Diagnosis History (Any Patient)

Table 66 illustrates the “View Diagnosis History” functionality. It enables medical specialists to view the diagnosis history for any patient.

<b>Actors</b>	Medical specialists.
<b>Description</b>	Medical specialists have the privilege to view the diagnosis history for any patient including all previous or recent diagnosed diseases.
<b>Data</b>	Medical specialists can access the medical history for any patient using MRN.
<b>Stimulus</b>	“View Patient History” button is clicked after entering the MRN.
<b>Response</b>	The doctor will be transferred to page that includes the results of the laboratory tests and previous patient diagnosis.
<b>Abnormal conditions</b>	<ul style="list-style-type: none"> <li>Incorrect input type.</li> <li>MRN does not exist.</li> </ul>

Table 66 View Diagnosis History for Medical Specialists

Figure 20 shows the activity diagram of “View Diagnosis History” functionality for the Medical Specialists.

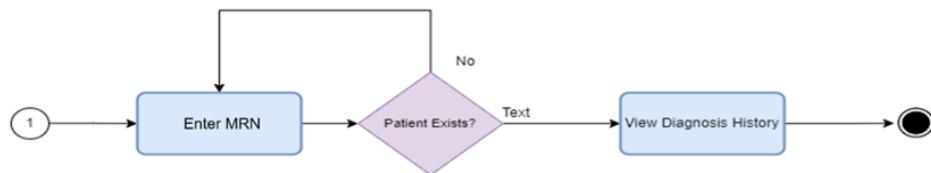


Figure 20 The Activity Diagram of View Diagnosis History Functionality

#### 4.2.3.2.4 User Class 3: Laboratory Specialist

This section describes the functionalities of the laboratory specialist.

##### 4.2.3.2.4.1 Storing Data for Patients and Performing an Overall Diagnosis

Table 67 illustrates “Storing Data” and “Performing an Overall Diagnosis” functionalities. It enables the laboratory specialist to store new records for patients while performing an overall diagnosis.

<b>Actors</b>	Laboratory specialist.
<b>Description</b>	<ul style="list-style-type: none"> <li>A laboratory specialist has the privilege to store data in the database and make an overall prediction for all diseases unlike other users who are only allowed to make one prediction at a time without any changes in the database.</li> <li>The laboratory specialists must fill the MRN for the intended patient.</li> <li>If the laboratory specialist chooses one or more diseases, only the features related to the specified disease/s need to be provided.</li> </ul>
<b>Data</b>	<p>The data is all the results of the laboratory tests for the patient performed by the laboratory specialist including:</p> <ul style="list-style-type: none"> <li>MRN.</li> <li>Diseases to be diagnosed.</li> <li>Demographical data.</li> <li>Blood tests.</li> <li>Symptoms.</li> <li>Other tests.</li> </ul>
<b>Stimulus</b>	The user command is issued after filling all the required information and clicking on “Store and Predict” button.
<b>Response</b>	<ul style="list-style-type: none"> <li>All information will be stored in the database.</li> <li>An overall prediction will be performed, and the results will be sent to the medical specialist.</li> </ul>
<b>Abnormal conditions</b>	<ul style="list-style-type: none"> <li>Incorrect input type.</li> <li>MRN does not exist.</li> </ul>

Table 67 View Storing Data and Performing an Overall Diagnosis

Figure 21 shows the activity diagram of “Storing Data” and “Performing an Overall Diagnosis” functionality for the laboratory specialist.

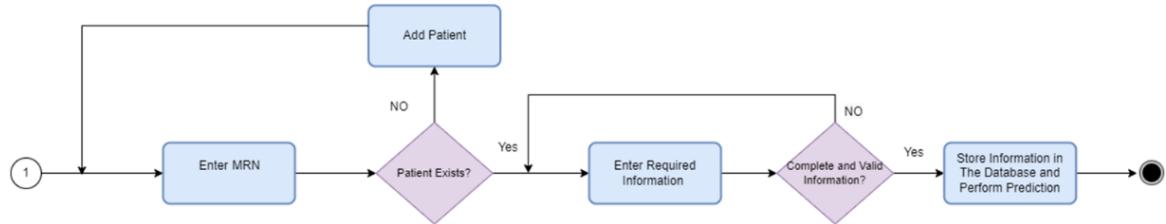


Figure 21 The Activity Diagram of Storing Data and Performing an Overall Diagnosis Functionality

#### 4.2.3.2.5 User Class 4: Registered User

This section describes the functionalities of the registered user.

##### 4.2.3.2.5.1 Create Account

Table 68 shows the description of “Create Account” functionality.

<b>Actors</b>	Registered user.
<b>Description</b>	<ul style="list-style-type: none"> <li>A user should be able to create an account to be a registered user.</li> <li>The required information to create an account include their full name, email, password, birthdate, and gender.</li> </ul>
<b>Data</b>	<ul style="list-style-type: none"> <li>User's full name.</li> <li>Username.</li> <li>Email.</li> <li>Birthdate.</li> <li>Gender.</li> <li>Password.</li> <li>Password Confirmation.</li> </ul>
<b>Stimulus</b>	All required information is filled in, and “Create Account” button is clicked.
<b>Response</b>	<ul style="list-style-type: none"> <li>An account is created if all required information is filled.</li> </ul>
<b>Abnormal conditions</b>	<ul style="list-style-type: none"> <li>Name includes numeric digits.</li> <li>Username already existed.</li> <li>Email format is incorrect.</li> <li>Password and confirm password do not match.</li> <li>Password does not comply with password constraints (at least 8 alphanumeric characters).</li> </ul>

Table 68 Create Account Functionality

Figure 22 shows the activity diagram of “Create Account” functionality.

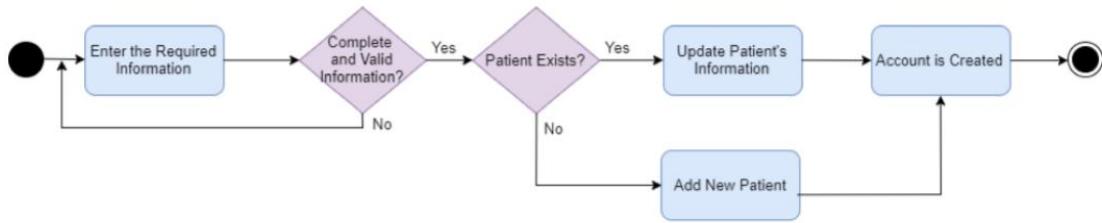


Figure 22 The Activity Diagram of the Create Account Functionality

#### 4.2.3.2.6 User Class 5: Guest User

This section describes the functionalities of Guest user.

##### 4.2.3.2.6.1 Off-Records Diagnosis

Table 69 shows the description of “Off-Record Diagnosis” functionality.

<b>Actors</b>	Guest user.
<b>Description</b>	The guest user will be able to issue an off-record diagnosis for the disease of choice.
<b>Data</b>	In this case the guest user will enter the data for the chosen disease only as described in previous sections.
<b>Stimulus</b>	The user command is issued by clicking on “Diagnose” button.
<b>Response</b>	The result of the predictive model will be displayed which determines the possibility of having the disease.
<b>Abnormal conditions</b>	<ul style="list-style-type: none"> <li>• An invalid input or data type.</li> <li>• Missing information.</li> </ul>

Table 69 Off-Record Diagnosis Functionality

Figure 23 shows the activity diagram of “Off-Record Diagnosis” functionality.

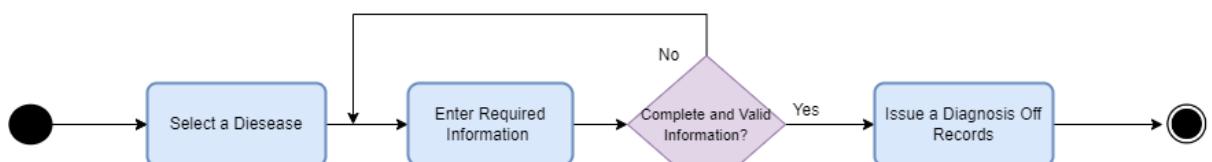


Figure 23 The Activity Diagram of Off-Record Diagnosis Functionality

#### 4.2.4 Performance Requirements

The implemented performance requirements are based on a set of standards to approximate the Response Time, Availability, Fault Tolerance Recovery, and Capacity.

##### 4.2.4.1 Response Time

Building the diagnosis model occupies most of the time roughly 1.35. Accordingly seconds the system shouldn't take more than 2 seconds to respond.

#### **4.2.4.2 Availability**

Accessibility of the system is determined by availability. The proposed system should be accessible for diagnosis 24/7.

#### **4.2.4.3 Fault Tolerance Recovery**

The system must be able to recognize and repair errors brought on by user input, problems with the hardware and operating system that power the system, or both. Below is a description of the steps the system takes to retain its functionality when faults are generated.

- **User Faults:** Incoming user input should be verified by the system, and any invalid input should trigger a warning message. The notification should instruct the user on how to fix the mistakes and violations.
- **System Faults:** The failures should be handled by the system by returning them to a previously valid condition.

#### **4.2.4.4 Capacity**

To fully utilize the capabilities and purposes of the system, a computer must have a minimum of 60 MB of storage space available.

### **4.2.5 Design Constraints**

The system is a web-based system that utilizes the following tools in its processes:

- Excel and Python to pre-process the datasets.
- Python to develop the program along with building diagnosis model.
- MongoDB for the database.

### **4.2.6 Software System Attributes**

#### **4.2.6.1 Usability**

By using universal design principles, the system creates user-friendly interfaces that facilitate quick learning and minimize user error.

#### **4.2.6.2 Reliability**

To ensure the system's reliability, appropriate input validations are used, such as warning the user with error messages regarding the system's current condition. Additionally, the reliability of the predictive diagnosis is reflected by the diagnosis model accuracy.

#### **4.2.6.3 Availability**

The system should be accessible 24/7 to guarantee its viability for users.

#### **4.2.6.4 Safety Requirements**

The system shouldn't interfere with other software, affect its internal components, or damage the server hosting the data in any way.

#### **4.2.6.5 Security and Privacy**

The system stores patient data in a secure manner. Consequently, the system will utilize the following security specifications:

- By archiving the hashed passwords in the database, a secure login mechanism is provided for system access. The system's passwords will then be matched to their hashed counterparts stored in the database.
- After 15 minutes of inactivity, a user will be automatically logged out.

#### **4.2.6.6 Maintainability**

This is when the system handles more available medical data related to Diabetes Mellitus, Chronic Kidney Disease, Coronary Heart Disease, Asthma, Thyroid Cancer, Schizophrenia, Glaucoma, Alzheimer's Disease, Lung Cancer, Rheumatoid Arthritis, Hypothyroidism, Prostate Cancer, Cervical Cancer, Multiple Sclerosis, Liver Cirrhosis, Chronic Obstructive Pulmonary Disease, Parkinson's Disease, Hepatitis C, Depression, Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia. The admin can update the diagnostic model to guarantee its usefulness to the current environment.

#### **4.2.6.7 Portability**

Any browser linked to a server with Python and medical data should be able to run the system.

## Chapter 5: Software Design Specification (SDS)

This chapter illustrates the components and interfaces described in the SRS document. It defines the processes of the system architecture to create an efficient system that meets the client's requirements.

The SDS is conducted in two stages:

- **The initial design phase:** The architecture and data components of the entire system are briefly defined in this phase.
- **Detailed design phase:** A detailed description of the system design is defined in this phase. This section of the SDS highlights purpose, scope, acronyms, and reference of this project.

### 5.1 System Overview

The pre-emptive Diagnosis system for Chronic Diseases serves to benefit society, by making it easier to diagnose a disease early or better still, prevent it from setting in. A user in the system is assigned to a certain task to achieve this goal. The diseases that will be diagnosed are Diabetes Mellitus, Chronic Kidney Disease, Coronary Heart Disease, Asthma, Thyroid Cancer, Schizophrenia, Glaucoma, Alzheimer's Disease, Lung Cancer, Rheumatoid Arthritis, Hypothyroidism, Prostate Cancer, Cervical Cancer, Multiple Sclerosis, Liver Cirrhosis, Chronic Obstructive Pulmonary Disease, Parkinson's Disease, Hepatitis C, Depression, Epileptic Seizure, Osteoporosis, and Sickle Cell Anemia.

#### 5.1.1 System Functionality

The system functionality allows users to have access to different parts of the system based on their privileges. In the system, each user has a specific task. Admin, the medical specialist, laboratory specialist, registered user, and guest user are the five sorts of users. The following subsections will outline the functionalities of each user.

##### 5.1.1.1 Common Functionalities

All users can log into the system to access their functionalities, except for guest users.

##### 5.1.1.2 Admin

- Add new users to the system.
- Delete users from the system.
- Rebuild a diagnosis model.
- Replace the diagnosis model.
- Retrieve the forgotten password.
- Update profile (change email and password).

##### 5.1.1.3 Medical Specialist

- View patients' diagnosis history.
- Diagnose patients.
- Print diagnosis result.

- Update profile (change email and password).
- Retrieve the forgotten password.

#### **5.1.1.4 Laboratory Specialist**

- Diagnose patients.
- Issue an overall diagnosis.
- Update profile (change email and password).
- Retrieve the forgotten password.

#### **5.1.1.5 Registered User**

- Create an account.
- Delete account.
- Diagnose themselves.
- Print diagnosis result.
- View diagnosis history.
- Update profile.
- Retrieve the forgotten password.

#### **5.1.1.6 Guest User**

- Diagnosis themselves.
- Print diagnosis result.

### **5.2 Design consideration**

The design solution requires defining and resolving the existing problems within the system. In the following sections, some of the possible issues encountered will be discussed.

#### **5.2.1 Assumptions and Dependencies**

This subsection demonstrates the assumptions and dependencies of the system, along with their applications. Several concerns may be issued as the following sections outline.

##### **5.2.1.1 Related Software or Hardware**

The proposed system is designed to work on a variety of servers, including Microsoft Edge, Safari, Chrome, and others. It should also be linked to the MongoDB database to process, store, and retrieve data.

##### **5.2.1.2 Operating System**

The proposed system is expected to work well while surfing on Windows 11 and macOS 12. It is also necessary to download the Python Runtime Environment in to be able to run the system using the Python programming language.

### **5.2.1.3 End-User Characteristics**

The proposed system contains five end-users, namely: admin, medical specialist, laboratory specialist, registered user, and guest user. Each user's role and privileges were specified in the SRS (Chapter 4) in section 4.2.2.

### **5.2.1.4 Possible and/or Probable Changes in Functionalities**

Certain adjustments are permissible if they improve the overall quality of the project. Otherwise, no adjustments would be possible because this project is time-limited, and most of the major changes are time-consuming.

## **5.2.2 General Constraints**

This section highlights the global limitations and constraints that have a significant impact on system design software.

### **5.2.2.1 Hardware or Software Environment**

The proposed system is intended to run properly on Windows 11 and macOS 12 operating systems. It requires a connection with the database using MongoDB to store and retrieve the data.

### **5.2.2.2 Interface Requirements**

The system interfaces should be user-friendly, allowing the user to readily comprehend the operating system's functions. The following are design considerations for system development:

- Make use of the simplest and least error-prone input components possible.
- To manage user errors, provide detailed error/warning messages.
- Make use of legible fonts, labels, and colors.

### **5.2.2.3 Data Repository and Distribution Requirements**

All the data used by the system should be stored in the MongoDB database. The system should handle the data carefully, to avoid data loss by any mistake.

### **5.2.2.4 Security Requirements**

The system will ensure users' confidentiality where users can access the system with defined privileges based on their roles.

### **5.2.2.5 Performance Requirements**

All the mentioned performance requirements in the SRS (Chapter 4) section 4.2.4 must be reviewed and tested to ensure they meet the defined requirements.

### **5.2.2.6 Verification and Validation Requirements (Testing)**

The purpose of system validation and verification is to guarantee that the specifications and requirements are satisfied. A test is considered successful if the requirements are fully implemented in the system.

## 5.3 User Interface Design

This section provides an overview of user interfaces, their design rules, screen images, and screen objects and actions.

### 5.3.1 Overview of User Interface

The main interface of the website is the login interface, which empowers every user to get to his/her account. To guarantee system security, each kind of user accesses the system with various privileges.

#### 5.3.1.1 Admin Interface

The admin interface contains the following tabs:

- ❖ **Manage users:**
  - Add new user.
  - Remove existing users.
- ❖ **Rebuild and Update Model:**
  - Upload Data File.
  - Choose the disease type.
  - Pick the training percentage.
  - Generate model.
  - Update Diagnostic Model.
- ❖ **Profile:**
  - Change admin's email or password.

#### 5.3.1.2 Medical Specialist Interface

The medical specialists have two tabs with the following options:

- ❖ **Diagnosis History:**
  - View patients' diagnosis history.
  - Diagnose the patients.
  - Print the diagnosis outcome.
- ❖ **Profile:**
  - Change profile information (email and password).

#### 5.3.1.3 Laboratory specialist interface

The laboratory specialist has two tabs with the following options:

- ❖ **Laboratory Test:**
  - Issue one or an overall diagnosis.
- ❖ **Profile**

- Change profile information (email and password).

#### 5.3.1.4 Registered User Interface

This registered user must have the validity to create an account to access the system. Then, an interface is displayed with the following tabs:

❖ **Diagnosis History:**

- View diagnosis history.
- Start a new diagnosis.
- Print the diagnosis result.

❖ **Profile:**

- Edit their profile information, including name, date of birth, and gender.
- Change email and password.
- Delete account.

#### 5.3.1.5 Guest User Interface

This user is allowed to run a diagnosis and print the generated result.

### 5.3.2 Interface Design Rules

To efficiently satisfy the users' functions, user interfaces should be organized in a simple, productive, and consistent manner. Furthermore, the interfaces should be designed to reduce the system's client confusion.

To ensure interfaces designs meet the universal usability of the system, Shneiderman's eight golden rules will guide us to structure and design these interfaces [103], which include:

- ❖ **Strive for consistency:** The consistency with which interfaces are structured plays an important role in assisting the customer in becoming acquainted with the interfaces. Every interface setup, such as the menu, colors, and typefaces, should be predictable. Furthermore, the terminology used in prompts, error messages, and menus should be unambiguously consistent to ensure comprehension.
- ❖ **Seek universal usability:** Designers must keep in mind that different users will have different experiences with the interfaces. Furthermore, designers must increase the interface's convenience and quality by providing other routes for specialists and reasonable clarifications for novices.
- ❖ **Offer informative feedback:** After a big activity or sequence of activities, designers should offer suitable, human-readable feedback to the user to inform them of what is occurring and where they are.
- ❖ **Design dialogs to yield closure:** Designers should demonstrate the sequence of actions for the user. Furthermore, by identifying the activity's starting and endpoints, users will be able to determine where the activities begin and finish. Furthermore, consumers will be able to determine whether the procedure was successful or not.

- ❖ **Prevent Errors:** The designers should be aware of any potential faults that users may make when using the system. However, if invalid input is entered, the user should be informed of the location of the issue. Furthermore, if a user makes a mistake, the interface should provide a straightforward, useful, and explicit message to help the user correct the mistake. For example, not giving options that are not to be selected, and keeping a strategic distance from alphabetic characters if numbers are necessary, will reduce the likelihood of user errors.
- ❖ **Permit easy reversal of actions:** Designers should give the user a chance to have the option to fix a solitary activity or gathering of activities.
- ❖ **Support internal locus of control:** The designers should structure the interfaces so that giving the user a chance to have the option to control the actions in the system.
- ❖ **Reduce short-term memory load:** The designers ought to know about structuring interfaces that don't give the user a chance to recall data from the past interface.

### 5.3.3 Screen Images

This section clarifies the interfaces of the system, using the Balsamiq prototyping tools to build the basic interface design of the system.

#### 5.3.3.1 System Access Interfaces

##### 5.3.3.1.1 Login Interface

The initial interface of the system is the “Login” interface which is displayed to all users. The users may be the admin, medical specialist, laboratory specialist, registered user, or guest. The first four users can enter the system directly by entering their username and password and clicking the ‘Login’ button. If the users forget their password, they can retrieve their password through the ‘Forget password’ option that takes the user to retrieve the password process which is covered in the next interface. The ‘Create account’ button is for new patients who want to keep records of their diagnostics. Also, the guest user can enter the system and see their diagnostics without saving their record by clicking the ‘Log in as a guest’ option. Figure 24 illustrates the login interface.

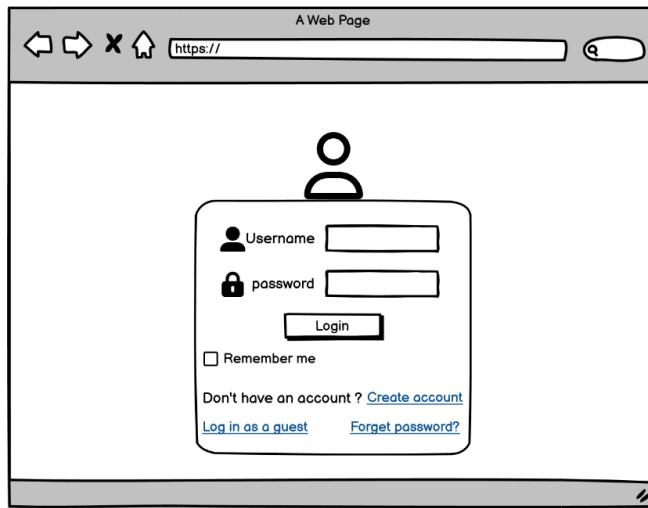


Figure 24 Login Interface

#### 5.3.3.1.2 Retrieve Password Interface

The user can reach the retrieve password interface by clicking the ‘Forgot Password?’ option which is in the login interface. The user can retrieve their password by entering their email in the text field and clicking on the ‘Send’ button. A temporary password will be sent to the user's email to enable them to view the system. Figure 25 presents the “Retrieve Password” interface.

Figure 25 Retrieve Password Interface

#### 5.3.3.1.3 Create Account Interface

The create account interface enables the patients (non-medical staff) to register in the system. Figure 26 presents the create account interface.

A wireframe of a web browser window titled 'A Web Page'. The address bar shows 'https://'. The main content area is titled 'Create Account' with a back arrow. It contains seven input fields: 'Username' (with user icon), 'Name' (with person icon), 'Gender' (dropdown menu with person icon), 'Email' (with envelope icon), 'Password' (with lock icon), and 'Confirm Password' (with lock icon). A 'Sign Up' button is at the bottom.

Figure 26 Create Account Interface

### 5.3.3.2 Admin Interface

When the user enters the system as an admin, they have three tabs in their interface which are a Profile, “Manage Users” and “Rebuild Model”. Each tab has its tasks which will be illustrated in the next subsections.

#### 5.3.3.2.1 Profile Tab

The admin, medical specialist, laboratory specialist, and registered user have a shared profile tab. In this tab, the user can see their information. Also, through this tab, the users can change their email address or password by clicking the “Change Email” or “Change Password” buttons, which forward users to the “Change Email” interface or “Change Password” interface respectively. Figure 27 presents the profile tab interface.

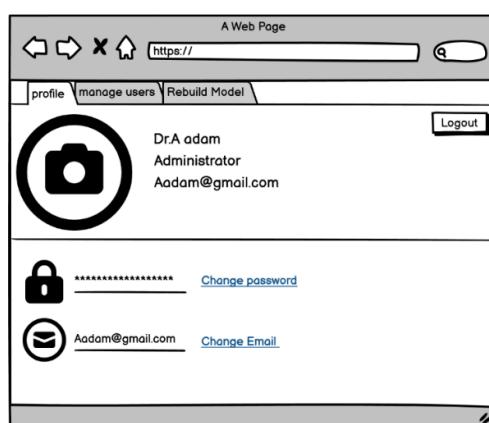


Figure 27 Admin Profile Tab

#### ❖ Change Email

In case the user needs to change their email, they must enter their current email and their new email. The confirmation of the new email should be before changing the email and using it. Figure 28 shows the change email interface.

Figure 28 Change Email Interface

#### ❖ Change Password

In case the user needs to change their password, they must enter their current password and their new password, where the confirmation of the new password should be before changing the password and using it. Figure 29 presents the change email interface.

Figure 29 Change Password Interface

#### 5.3.3.2.2 Manage Users Tab

From the manage user tab, the admin can add and remove users from the system. The "Add User" icon is for adding the users and the "Remove User" icon is for removing the users. Moreover, the search box allows the admin to easily find any user. Figure 30 presents the manage users tab.

A screenshot of a web browser window titled "A Web Page". The address bar shows "https://". The navigation bar includes back, forward, stop, and search buttons. The tabs are "Profile", "Manage users" (which is active), and "Rebuild Model". Below the tabs is a toolbar with icons for "Add User" (a user plus sign) and "Remove User" (a user minus sign). A search bar labeled "Search User" is on the right. The main content area displays a table of users:

Name	Email	Role	
Ibrahim Alnafea	ibrra@gmail.com	Admin	<a href="#">Edit</a>
Abdulaziz Alzahrani	Aziz@gmail.com	Laboratory specialist	<a href="#">Edit</a>
Abdullah Alamri	abdul@hotmail.com	User	<a href="#">Edit</a>
Osamah Aldahlan	os@outlook.com	Medical specialist	<a href="#">Edit</a>
Abdulaziz Alsubaie	abusaud@gmail.com	User	<a href="#">Edit</a>

At the bottom are navigation buttons: "previous" and "next".

Figure 30 Manage Users Tab

#### ❖ Add User

The admin can reach the add a user interface by clicking on the "Add User" icon from Figure 30. This interface allows the admin to add a new user to the system by entering the username, name, email, and choosing the role of the user. Figure 31 presents the "Add User" tab.

A screenshot of a web browser window titled "A Web Page". The address bar shows "https://". The navigation bar includes back, forward, stop, and search buttons. The tabs are "Profile", "Manage users" (active), and "Rebuild Model". Below the tabs is a toolbar with an "Add New User" icon. The main content area displays a form:

**Add New User**

Fields:

- Username
- Name
- Email
- Role:
  - Admin
  - Laboratory Specialist
  - Medical Specialist

Buttons:

- Add User

Figure 31 Add User Tab

#### ❖ Remove User

This interface permits the admin to remove users by filling in the user's name, email, and role. Figure 32 presents the Remove user interface.

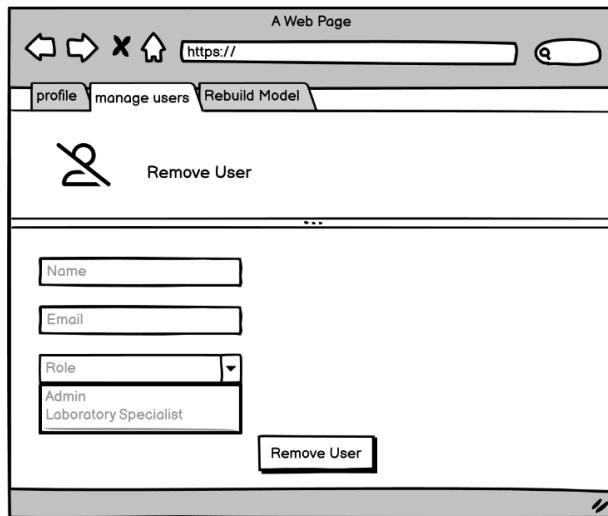


Figure 32 Remove User Tab

#### 5.3.3.2.3 Rebuild Model Tab

The rebuild model tab appears just for the admin user. It allows the admin to upload a data file, which contains the medical records of patients and non-patients with Diabetes Mellitus, Chronic Kidney Disease, Coronary Heart Disease, Asthma, Thyroid Cancer, Schizophrenia, Glaucoma, Alzheimer's Disease, Lung Cancer, Rheumatoid Arthritis, Hypothyroidism, Prostate Cancer, Cervical Cancer, Multiple Sclerosis, Liver Cirrhosis, Chronic Obstructive Pulmonary Disease, Parkinson's Disease, Hepatitis C, and Depression. After, the admin selects the disease from the drop-down list and assigns a training percentage then clicks the "Generate Model". On the upper side of the interface, a summary of the model will be displayed. It is often the best training percentage that is only discovered after multiple trials on the training percentage. Therefore, a record of each tested model is stored in the Models table and whenever clicked, the model summary is changed to reflect the chosen model. Finally, when the admin is satisfied with a model, clicking the "Update Diagnostic Model" Stores the chosen model with the accuracy for future reference. Figure 33 presents the rebuild model tab.

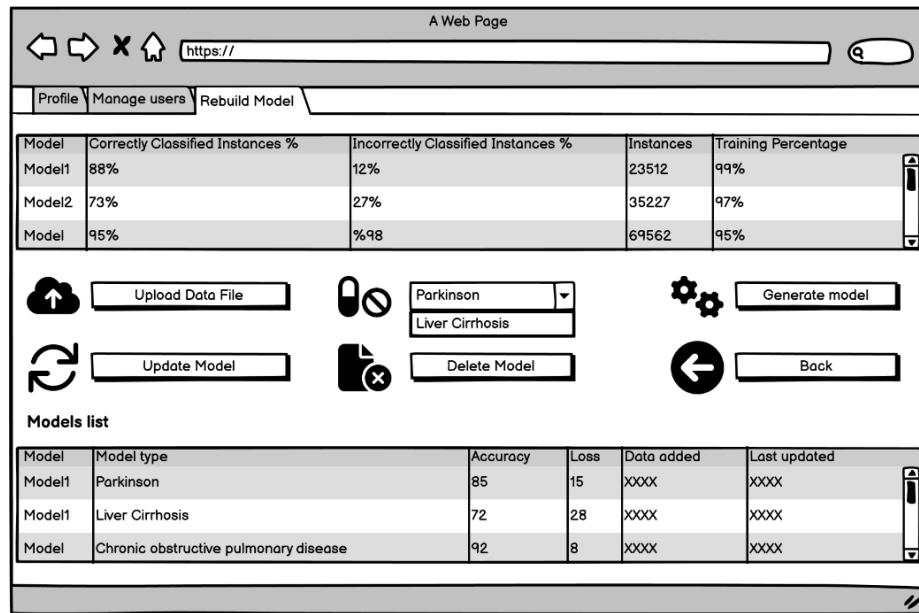


Figure 33 Rebuild Model Tab

### 5.3.3.3 Medical Specialist Interface

When the user enters the system as a medical specialist, they have two tabs in their interface which is a Profile and Diagnosis History. Each tab has its tasks which will be illustrated in the next subsections.

#### 5.3.3.3.1 Profile Tab

The interface of the profile tab is similar to the interface of the admin's profile described in 5.3.3.2.1. Figure 34 demonstrates the medical specialist's profile interface.

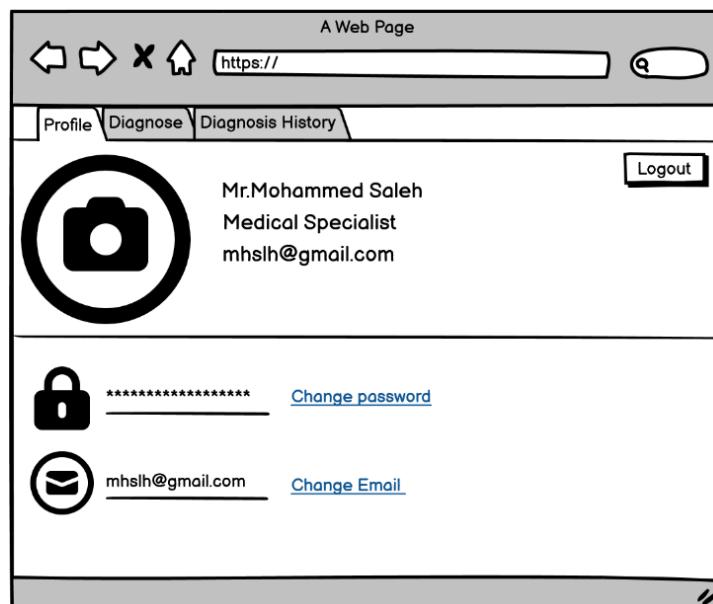


Figure 34 Medical Specialist Profile Interface

### 5.3.3.3.2 Diagnosis History Tab

In the diagnosis history tab, a medical specialist can enter the MRN of a patient's then click "Search" to view the patient's diagnosis history. Figure 35 demonstrates the diagnosis history tab.

The screenshot shows a web browser window titled "A Web Page". At the top, there are navigation icons (back, forward, search, etc.) and a URL bar with "https://". Below the URL bar, a navigation menu has "Profile" and "Diagnosis History" selected. On the left, there is a search bar with "MRN" and "Test ID" dropdowns, and a "Search" button with a magnifying glass icon. In the center, the title "Diagnosis History" is displayed above a table. The table has columns: MRN, Patient Name, Test ID, Diagnosis Date, Disease, Result, and Result Accuracy. The data in the table is as follows:

MRN	Patient Name	Test ID	Diagnosis Date	Disease	Result	Result Accuracy
2351	John Doe	1001423	20/10/2022	Parkinson	Negative	90.30%
5573	Alex James	1001467	07/10/2022	Liver Cirrhosis	Positive	89.10%
8433	Sara Doe	1001846	30/09/2022	COPD	Negative	94.14%
1458	Steve Jobs	1002580	25/09/2022	COPD	Positive	78.80%
9854	Elon Musk	1002478	08/09/2022	Liver Cirrhosis	Positive	91.32%
456	Joe Alex	1002498	26/08/2022	Parkinson	Negative	80.07%

Figure 35 Diagnosis History Tab

### 5.3.3.3.3 Diagnose Diabetes Mellitus Tab

In this tab, the medical specialist enters the patient's diagnosis information and clicks "Diagnose". Figure 36 demonstrates the diagnose Diabetes Mellitus tab.

The screenshot shows a web browser window titled "A Web Page". At the top, there are navigation icons (back, forward, search, etc.) and a URL bar with "https://". Below the URL bar, a navigation menu has "Diagnose Diabetes Mellitus" selected. The main content area is titled "Diagnose Diabetes Mellitus". It contains a section labeled "Diagnosis Information" with three input fields: "Hematoctrit (HTC)", "Glucose", and "Mean platelet volume", each with an associated text input box. At the bottom of the page is a "Diagnose" button.

Figure 36 Diagnose Diabetes Mellitus Tab

#### **5.3.3.3.4 Diagnose Chronic Kidney Disease Tab**

The tab that diagnoses Chronic Kidney Disease (CKD) is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 37 demonstrates the diagnose CKD tab.

A screenshot of a web browser window titled 'A Web Page'. The address bar shows 'https://'. The main content area has a header 'Diagnose Chronic Kidney Disease'. Below it, a section titled 'Diagnosis Information' contains two input fields: 'Blood urea nitrogen' and 'Creatinine', each with a corresponding text input box. At the bottom is a 'Diagnose' button.

Figure 37 Diagnose Chronic Kidney Disease Tab

#### **5.3.3.3.5 Diagnose Coronary Heart Disease Tab**

The tab that diagnoses Coronary Heart Disease (CHD) is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 38 demonstrates the diagnose CHD tab.

A screenshot of a web browser window titled 'A Web Page'. The address bar shows 'https://'. The main content area has a header 'Diagnose Coronary Heart Disease'. Below it, a section titled 'Diagnosis Information' contains seven input fields: 'Sex', 'Age', 'DBili', 'BUN', 'Tbili', 'T Protein', and 'Creat', each with a corresponding text input box. At the bottom is a 'Diagnose' button.

Figure 38 Diagnose Coronary Heart Disease Tab

### 5.3.3.6 Diagnose Asthma Tab

The tab that diagnoses Asthma is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 39 demonstrates the diagnose Asthma tab.

A screenshot of a web browser window titled 'A Web Page'. The address bar shows 'https://'. The tab title is 'Diagnose Asthma'. The main content area is titled 'Diagnose Asthma' and contains a section labeled 'Diagnosis Information'. This section includes input fields for Sex, Age, BASO, HCT, HGB, MCH, MCHC, MCV, and WBC. Below these fields is a 'Diagnose' button.

Figure 39 Diagnose Asthma Tab

### 5.3.3.7 Diagnose Thyroid Cancer Tab

The tab that diagnoses Thyroid Cancer is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 40 demonstrates the diagnose Thyroid Cancer tab.

A screenshot of a web browser window titled 'A Web Page'. The address bar shows 'https://'. The tab title is 'Diagnose Thyroid Cancer'. The main content area is titled 'Diagnose Thyroid Cancer' and contains a section labeled 'Diagnosis Information'. This section includes input fields for Sex, Age, HCT, MCHC, MPV, RBC, and WBC. Below these fields is a 'Diagnose' button.

Figure 40 Diagnose Thyroid Cancer Tab

### 5.3.3.3.8 Diagnose Schizophrenia Tab

The tab that diagnoses Schizophrenia is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 41 demonstrates the diagnose Schizophrenia tab.

A screenshot of a web browser window titled "A Web Page". The address bar shows "https://". The tab title is "Diagnose Schizophrenia". The main content area is titled "Diagnose Schizophrenia Disease". A section labeled "Diagnosis Information" contains a table with the following data:

	Gender	
NEU		
LYM		
HGB		
HCT		
MCHC		
RDW		
UREA		

Below the table is a "Diagnose" button.

Figure 41 Diagnose Schizophrenia Tab

### 5.3.3.3.9 Diagnose Glaucoma Tab

The tab that diagnoses Glaucoma is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 42 demonstrates the diagnose Glaucoma tab.

A screenshot of a web browser window titled "A Web Page". The address bar shows "https://". The tab title is "Diagnose Glaucoma". The main content area is titled "Diagnose Glaucoma Disease". A section labeled "Diagnosis Information" contains a table with the following data:

	at	
ean		
mhc1		
vasi		
varg		
Vars		
tmi		

Below the table is a "Diagnose" button.

Figure 42 Diagnose Glaucoma Tab

#### **5.3.3.3.10 Diagnose Alzheimer's Disease Tab**

The tab that diagnoses Alzheimer's Disease is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 43 demonstrates the diagnose Alzheimer's Disease tab.

A screenshot of a web browser window titled "A Web Page". The address bar shows "https://". The tab title is "Diagnose Alzheimer's". The main content area is titled "Diagnose Alzheimer's Disease". It contains a form titled "Diagnosis Information" with the following fields:

Gender	<input type="text"/>
Age	<input type="text"/>
Pulse	<input type="text"/>
Respiratory_Rate	<input type="text"/>
BP_Diastolic	<input type="text"/>
Wbc	<input type="text"/>
Rbc	<input type="text"/>
...	<input type="text"/>
Hemoglobin	<input type="text"/>

At the bottom of the form is a "Diagnose" button.

Figure 43 Diagnose Alzheimer's Disease Tab

#### **5.3.3.3.11 Diagnose Lung Cancer Tab**

The tab that diagnoses Lung Cancer is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 44 demonstrates the diagnose Lung Cancer tab.

A screenshot of a web browser window titled "A Web Page". The address bar shows "https://". The tab title is "Diagnose Lung Cancer". The main content area is titled "Diagnose Lung Cancer Disease". It contains a form titled "Diagnosis Information" with the following fields:

Gender	<input type="text"/>
Age	<input type="text"/>
Smokin	<input type="text"/>
Yellow Fingers	<input type="text"/>
Anxiety	<input type="text"/>
Peer_Pressure	<input type="text"/>
Chronic Diseases	<input type="text"/>
...	<input type="text"/>
Fatigue	<input type="text"/>

At the bottom of the form is a "Diagnose" button.

Figure 44 Diagnose Lung Cancer Tab

### 5.3.3.3.12 Diagnose Rheumatoid Arthritis Tab

The tab that diagnoses Rheumatoid Arthritis is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 45 demonstrates the diagnose Rheumatoid Arthritis tab.

A screenshot of a web browser window titled "A Web Page". The address bar shows "https://". The main content area has a header "Diagnose Rheumatoid Arthritis Disease". Below it is a section titled "Diagnosis Information" containing a list of medical parameters with corresponding input fields:

Parameter	Input Field
Sex	<input type="text"/>
Age	<input type="text"/>
MCH	<input type="text"/>
MCV	<input type="text"/>
MCHC	<input type="text"/>
RDW	<input type="text"/>
Platelet Count	<input type="text"/>
...	<input type="text"/>
MPV	<input type="text"/>

At the bottom right of the form is a "Diagnose" button.

Figure 45 Diagnose Rheumatoid Arthritis Tab

### 5.3.3.3.13 Diagnose Hypothyroidism Tab

The tab that diagnoses Hypothyroidism is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 46 demonstrates the diagnose Hypothyroidism tab.

A screenshot of a web browser window titled "A Web Page". The address bar shows "https://". The main content area has a header "Diagnose Hypothyroidism Disease". Below it is a section titled "Diagnosis Information" containing a list of medical parameters with corresponding input fields:

Parameter	Input Field
Age	<input type="text"/>
BP-Systolic	<input type="text"/>
Respiratory Rat	<input type="text"/>
MCV	<input type="text"/>
Pulse Ox	<input type="text"/>

At the bottom right of the form is a "Diagnose" button.

Figure 46 Diagnose Hypothyroidism Tab

#### 5.3.3.3.14 Diagnose Prostate Cancer Tab

The tab that diagnoses Prostate Cancer is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 47 demonstrates the diagnose Prostate Cancer tab.

A screenshot of a web browser window titled "A Web Page". The address bar shows "https://". The main content area has a header "Diagnose Prostate Cancer Disease". Below it is a section labeled "Diagnosis Information" containing four input fields: "Perimeter", "Area", "Smoothness", and "Compactness", each with a corresponding horizontal input box. At the bottom is a "Diagnose" button.

Figure 47 Diagnose Prostate Cancer Tab

#### 5.3.3.3.15 Diagnose Cervical Cancer Tab

The tab that diagnoses Cervical Cancer is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 48 demonstrates the diagnose Cervical Cancer tab.

A screenshot of a web browser window titled "A Web Page". The address bar shows "https://". The main content area has a header "Diagnose Cervical Cancer". Below it is a section labeled "Diagnostic Information" containing eight dropdown menus: "STDs: Number of Diagnosis", "STDs: Condylomatosis", "STDs: Syphilis", "STDs: HIV", "STDs: HPV", "Dx", "Dx: CIN", and "Dx: HPV", each with a corresponding dropdown menu box. At the bottom is a "Diagnose" button.

Figure 48 Diagnose Cervical Cancer Tab

#### 5.3.3.3.16 Diagnose Multiple Sclerosis Tab

The tab that diagnoses Multiple Sclerosis is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 49 demonstrates the diagnose Multiple Sclerosis tab.

A screenshot of a web browser window titled 'A Web Page'. The address bar shows 'https://'. The tab title is 'Cervical Cancer | Multiple Sclerosis'. The main content area is titled 'Diagnose Multiple Sclerosis' and contains a 'Diagnostic Information' section. This section includes input fields for Age, WBC, RBC, Hemoglobin, MCH, MCV, MCHC, RDW, and MPV. A 'Diagnose' button is located at the bottom right of the form.

Figure 49 Diagnose Multiple Sclerosis Tab

#### 5.3.3.3.17 Diagnose Liver Cirrhosis Tab

The tab that diagnoses Liver Cirrhosis is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 50 demonstrates the diagnose Liver Cirrhosis tab.

A screenshot of a web browser window titled 'A Web Page'. The address bar shows 'https://'. The tab title is 'Diagnose Liver Cirrhosis'. The main content area is titled 'Diagnose Liver Cirrhosis Disease' and contains a 'Diagnosis Information' section. This section includes dropdown menus for Ascites, Hepatomegaly, Spiders, Edema, Bilirubin, Cholesterol, Albumin, Copper, and Alk\_Phosphatase. A 'Diagnose' button is located at the bottom right of the form.

Figure 50 Diagnose Liver Cirrhosis Tab

### 5.3.3.3.18 Diagnose Chronic Obstructive Pulmonary Disease Tab

The tab that diagnoses Chronic Obstructive Pulmonary Disease (COPD) is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 51 demonstrates the diagnose COPD tab.

A screenshot of a web browser window titled "A Web Page". The address bar shows "https://". The tab title is "Diagnose COPD". The main content area is titled "Diagnose Chronic Obstructive Pulmonary Disease". A section labeled "Diagnosis Information" contains a list of fields with input boxes: Pock History, Smoking, CAT, MWT1, MWT2, FEV1, FVC, HAD, and SGRQ. At the bottom is a "Diagnose" button.

Figure 51 Diagnose Chronic Obstructive Pulmonary Disease Tab

### 5.3.3.3.19 Diagnose Parkinson's Disease Tab

The tab that diagnoses Parkinson's Disease is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 52 demonstrates the diagnose Parkinson's Disease tab.

A screenshot of a web browser window titled "A Web Page". The address bar shows "https://". The tab title is "Diagnose Parkinson's Disease". The main content area is titled "Diagnose Parkinson's Disease". A section labeled "Diagnosis Information" contains a list of fields with input boxes: Anion Gap, ALT (Dimension), LDH, White Blood Cells, MWT2, FEV1, FVC, HAD, and SGRQ. At the bottom is a "Diagnose" button.

Figure 52 Diagnose Parkinson's Disease Tab

### 5.3.3.3.20 Diagnose Hepatitis C Tab

The tab that diagnoses Hepatitis C is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 53 demonstrates the diagnose Hepatitis C tab.

A screenshot of a web browser window titled 'A Web Page'. The address bar shows 'https://'. The tab bar has two tabs: 'Depression' and 'Hepatitis C', with 'Hepatitis C' being the active tab. The main content area is titled 'Diagnose Hepatitis C Disease'. A section labeled 'Diagnostic Information' contains a list of medical parameters with corresponding input fields: Sex, Age, Anion Gap, ALT, LDH, WBC, RBC, Hemoglobin, Hematocrit, Sodium, three ellipsis dots, Respiratory Rate, and BP - Diastolic. Below this list is a 'Diagnose' button.

Figure 53 Diagnose Hepatitis C Tab

### 5.3.3.3.21 Diagnose Depression Tab

The tab that diagnoses Depression is similar to the Diabetes Mellitus diagnosis tab except for the diagnostic information fields. Figure 54 demonstrates the diagnose Depression tab.

A screenshot of a web browser window titled 'A Web Page'. The address bar shows 'https://'. The tab bar has two tabs: 'Depression' and 'Hepatitis C', with 'Depression' being the active tab. The main content area is titled 'Diagnose Depression Disease'. A section labeled 'Diagnostic Information' contains a list of socioeconomic and financial parameters with corresponding input fields: Married, Number of Childrens, House Size, Education, Number of Young Childrens, Value of Stocks, Value of Cellphone, Value of Goods, Value of Savings, Land Owned, three ellipsis dots, Saved money using M-Pesa, and Amount Saved using M-Pesa. Below this list is a 'Diagnose' button.

Figure 54 Diagnose Depression Tab

### 5.3.3.3.22 Diagnose Epileptic Seizure Tab

The tab that diagnoses Epileptic Seizure is similar to the Diabetes Mellitus diagnosis tab also except for the diagnostic information fields. Figure 55 demonstrates the diagnose Depression tab.

A screenshot of a web browser window titled 'A Web Page'. The address bar shows 'https://'. The main content area has a title 'Diagnose Epileptic Seizure' and a sub-section title 'Diagnose Epileptic Seizure Disease'. Below this, there is a section labeled 'Diagnose information' containing the following fields:

Gender	<input type="text"/>
NumberOfNonPsychComorbidities	<input type="text"/>
NumberOfPrio	<input type="text"/>
Asthma	<input type="text"/>
Migraine	<input type="text"/>
Chronic Pain	<input type="text"/>
Diabetes	<input type="text"/>
Rape	<input type="text"/>

At the bottom right of the form is a 'Diagnose' button.

Figure 55 Diagnose Epileptic Seizure Tab

### 5.3.3.3.23 Diagnose Osteoporosis Tab

The tab that diagnoses Osteoporosis is similar to the Epileptic Seizure diagnosis tab except for the diagnostic information fields. Figure 56 demonstrates the diagnose Depression tab.

A screenshot of a web browser window titled 'A Web Page'. The address bar shows 'https://'. The main content area has a title 'Diagnose Osteoporosis' and a sub-section title 'Diagnose Osteoporosis Disease'. Below this, there is a section labeled 'Diagnose information' containing the following fields:

Gender	<input type="text"/>
Age	<input type="text"/>
Weight	<input type="text"/>
Height	<input type="text"/>
Diabetes	<input type="text"/>
Hypothyroidism	<input type="text"/>
SeizerDisorder	<input type="text"/>
Alcohol	<input type="text"/>
Obesity	<input type="text"/>

At the bottom right of the form is a 'Diagnose' button.

Figure 56 Diagnose Osteoporosis Tab

### 5.3.3.3.24 Diagnose Sickle Cell Anemia Tab

The tab that diagnoses Sickle Cell Anemia is similar to the Epileptic Seizure diagnosis tab except for the diagnostic information fields. Figure 57 demonstrates the diagnose Depression tab.

A Web Page  
https://  
Diagnose Sickle Cell Anemia  
Diagnose Sickle Cell Anemia Disease  
Diagnose information  
Sex  
Tribe  
Hemoglobin (HB)  
PCV  
RBCs  
MCV  
MCH  
PLTs  
Diagnose

Figure 57 Diagnose Sickle Cell Anemia Tab

#### ❖ Result

After a medical specialist clicks the “Diagnose”, the result of the diagnosis is displayed as presented in Figure 58. The back icon redirects the user to the homepage.

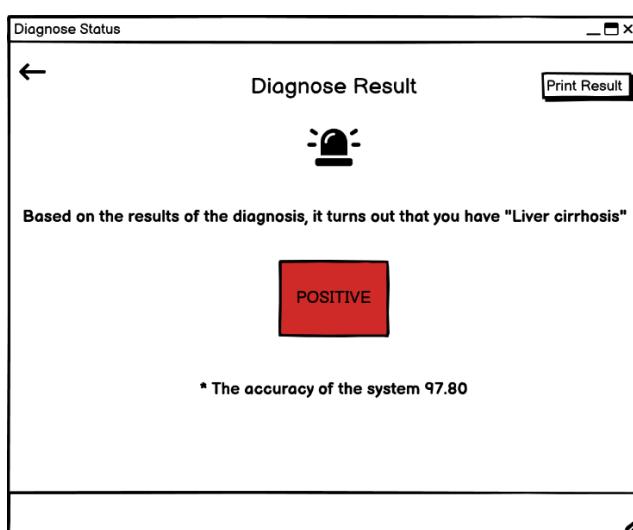


Figure 58 Diagnosis Result Interface

#### 5.3.3.4 Laboratory Specialist Interface

When the user enters the system as a laboratory specialist, they have two tabs in their interface which is a “Profile” and “Laboratory Test”. Each tab has its tasks which will be illustrated in the next subsections.

##### 5.3.3.4.1 Profile Tab

The interface of the profile tab is similar to the interface of the admin’s profile described in 5.3.3.2.1. Figure 59 demonstrates the laboratory specialist’s profile interface.

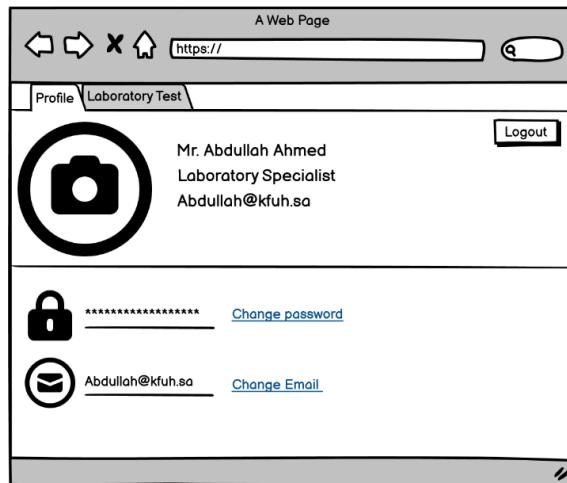


Figure 59 Laboratory Specialist Profile Interface

##### 5.3.3.4.2 Laboratory Test Tab

In the laboratory test tab, a laboratory specialist may choose the diseases to be diagnosed for the patient, enter their information, and fill the required fields. Then click the “Diagnose” to store the information in the database and perform the prediction. Figure 60 demonstrates the laboratory test tab.

A screenshot of a web browser window titled "A Web Page". The address bar shows "https://". Below the address bar is a navigation bar with tabs: "Profile" and "Laboratory Test" (which is selected). The main content area is divided into several sections: "Personal Information" (with fields for MRN and Name, and a "Retrieve" button), "Demographic Data" (with fields for Age, Height, Gender, Weight, BP, and BG), "Blood Tests" (with fields for WBC, Platelets, RBC, BUN, HG, and BG), "Symptoms" (with dropdown menus for Coughing, Anxiety, Fatigue, Wheezing, Allergy, X1, X2, X3, and X4), and "Other Tests" (with fields for X1 and X2). To the right of these sections is a "Disease Type" dropdown menu containing checkboxes for "Parkinson's Disease", "Liver Cirrhosis", "COPD", and "Multiple Sclerosis". A "Diagnose" button is located at the bottom right of the form.

Figure 60 Laboratory Test Tab

### 5.3.3.5 Registered User Interface

The registered user's interface includes two tabs: "Profile" and "Diagnosis History". Each will be explained separately in the following subsections.

#### 5.3.3.5.1 Profile Tab

The "Profile" interface is similar to the admin's profile interface described in section 5.3.3.2.1 with a small difference. The registered users have more options, such as date of birth and gender. Figure 61 demonstrates the registered user's profile tab.

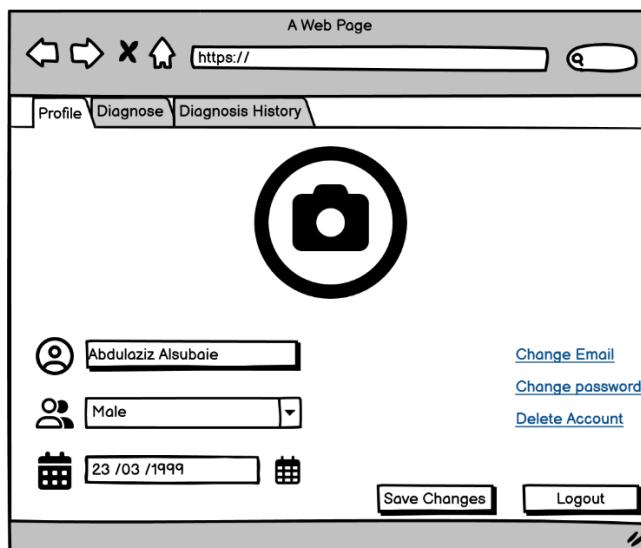


Figure 61 Registered User Profile Interface

#### 5.3.3.5.2 Diagnosis History Tab

This tab displays the diagnosis history of the registered user. Figure 62 demonstrates the "Diagnosis History" tab.

A wireframe screenshot of a web browser window titled 'A Web Page'. The address bar shows 'https://'. The navigation bar contains two tabs: 'profile' (which is active) and 'Diagnosis History'. Below the tabs is a 'Diagnosis History' icon. To the right are search fields for 'MRN' and 'Test ID Date' and a 'Search' button. The main area displays a table of diagnosis history data. The table has columns: MRN, Test ID, Date, Disease, Result, and Result Accuracy. Two rows of data are shown:

MRN	Test ID	Date	Disease	Result	Result Accuracy
2351	22215	20/10/2022	Parkinson	Negative	90.30%
2351	2351	13/10/2022	Liver Cirrhosis	Positive	95.80

*Figure 62 Registered Users' Diagnosis History Tab*

❖ **Diagnosis**

**5.3.3.5.3 Diagnose Diabetes Mellitus Tab**

To diagnose Diabetes Mellitus, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 36 clarifies the diagnosis Diabetes Mellitus tab.

**5.3.3.5.4 Diagnose Chronic Kidney Disease Tab**

To diagnose Chronic Kidney Disease (CKD), the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 37 clarifies the diagnosis CKD tab.

**5.3.3.5.5 Diagnose Coronary Heart Disease Tab**

To diagnose Coronary Heart Disease (CHD), the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 38 clarifies the diagnosis CHD tab.

**5.3.3.5.6 Diagnose Asthma Tab**

To diagnose Asthma, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 39 clarifies the diagnosis Asthma tab.

**5.3.3.5.7 Diagnose Thyroid Cancer Tab**

To diagnose Thyroid Cancer, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 40 clarifies the diagnosis Thyroid Cancer tab.

**5.3.3.5.8 Diagnose Schizophrenia Tab**

To diagnose Schizophrenia, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 41 clarifies the diagnosis Schizophrenia tab.

**5.3.3.5.9 Diagnose Glaucoma Tab**

To diagnose Glaucoma, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 42 clarifies the diagnosis Glaucoma tab.

**5.3.3.5.10 Diagnose Alzheimer's Disease Tab**

To diagnose Alzheimer's Disease, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 43 clarifies the diagnosis Alzheimer's Disease tab.

### **5.3.3.5.11 Diagnose Lung Cancer Tab**

To diagnose Lung Cancer, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 44 clarifies the diagnosis Lung Cancer tab.

### **5.3.3.5.12 Diagnose Rheumatoid Arthritis Tab**

To diagnose Rheumatoid Arthritis, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 45 clarifies the diagnosis Rheumatoid Arthritis tab.

### **5.3.3.5.13 Diagnose Hypothyroidism Tab**

To diagnose Hypothyroidism, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 46 clarifies the diagnosis Hypothyroidism tab.

### **5.3.3.5.14 Diagnose Prostate Cancer Tab**

To diagnose Prostate Cancer, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 47 clarifies the diagnosis Prostate Cancer tab.

### **5.3.3.5.15 Diagnose Cervical Cancer Tab**

To diagnose Cervical Cancer, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 48 clarifies the diagnosis Cervical Cancer tab.

### **5.3.3.5.16 Diagnose Multiple Sclerosis Disease Tab**

To diagnose Multiple Sclerosis, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 49 clarifies the diagnosis Multiple Sclerosis tab.

### **5.3.3.5.17 Diagnose Liver Cirrhosis Tab**

To diagnose Liver Cirrhosis, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 50 clarifies the diagnose Liver Cirrhosis tab.

### **5.3.3.5.18 Diagnose Chronic Obstructive Pulmonary Disease Tab**

To diagnose Chronic Obstructive Pulmonary Disease (COPD), the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 51 clarifies the diagnosis COPD tab.

### **5.3.3.5.19 Diagnose Parkinson’s Disease Tab**

To diagnose Parkinson’s Disease, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 52 clarifies the diagnosis Parkinson’s Disease tab.

### **5.3.3.5.20 Diagnose Hepatitis C Tab**

To diagnose Hepatitis C, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 53 clarifies the diagnosis Hepatitis C tab.

### **5.3.3.5.21 Diagnose Depression Tab**

To diagnose Depression, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 54 clarifies the diagnosis Depression tab.

### **5.3.3.5.22 Diagnose Epileptic Seizure Tab**

To diagnose Epileptic Seizure, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 55 clarifies the diagnosis Epileptic Seizure tab.

### **5.3.3.5.23 Diagnose Osteoporosis Tab**

To diagnose Osteoporosis, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 56 clarifies the diagnosis Osteoporosis tab.

### **5.3.3.5.24 Diagnose Sickle Cell Anemia Tab**

To diagnose Sickle Cell Anemia, the registered user must fill in the required information, then click “Diagnose” to show their diagnostic result. Figure 57 clarifies the diagnosis Sickle Cell Anemia tab.

## **5.3.3.6 Guest Interface**

The interface of the guest users contains three tabs: Diagnose Diabetes Mellitus, Diagnose Chronic Kidney Disease, Diagnose Coronary Heart Disease, Diagnose Asthma, Diagnose Thyroid Cancer, Diagnose Schizophrenia, Diagnose Glaucoma, Diagnose Alzheimer's Disease, Diagnose Lung Cancer, Diagnose Rheumatoid Arthritis, Diagnose Hypothyroidism, Diagnose Prostate Cancer, Diagnose Cervical Cancer, Diagnose Multiple Sclerosis, Diagnose Liver Cirrhosis, Diagnose Chronic Obstructive Pulmonary Disease, Diagnose Parkinson's Disease, Diagnose Hepatitis C, and Diagnose Depression. These tabs have been explained previously in subsections 5.3.3.3.3, 5.3.3.3.4, 5.3.3.3.5, 5.3.3.3.6, 5.3.3.3.7, 5.3.3.3.8, 5.3.3.3.9, 5.3.3.3.10, 5.3.3.3.11, 5.3.3.3.12, 5.3.3.3.13, 5.3.3.3.14, 5.3.3.3.15, 5.3.3.3.16, 5.3.3.3.17, 5.3.3.3.18, 5.3.3.3.19, 5.3.3.3.20, 5.3.3.3.21, 5.3.3.3.22, 5.3.3.3.23, and 5.3.3.3.24.

## **5.3.4 Screen Objects and Actions**

This subsection describes the objects and actions of each interface, as shown in Table 70.

NO.	Object	Type	Action
<i>Common Interfaces</i>			
<i>Login Interface (Figure 24)</i>			

NO.	Object	Type	Action
1.	Login	Button	The user can access their account after entering a valid username and password.
	Forget Password	Link	This link helps the user who forgot their password. It forwards the user to the “Forgot Password” interface to help the user retrieve the password.
	Login as Guest	Link	This link redirects the user directly to the “Diagnosis” interface.
	Create Account	Link	The user is directed to the create account interface.
<i>Retrieve Password Interface (Figure 25)</i>			
5.	Send Reset	Button	This button sends the user an email with a temporary password.
	Back	Button	The user can go back to the “Login” interface.
<i>Create Account Interface (Figure 26)</i>			
7.	Sign up	Button	This button enables the patient to register in the system.
	Back to Login	Button	The button directs the user to the “Login” interface
<i>Profile Interface (Figure 28, 34, 58, 61)</i>			
9.	Change Email	Button	This button allows the user to go to the “Change Email” interface.
	Change Password	Button	This button allows the user to go to the “Change Password” interface.
	Logout	Button	This button allows the user to exit from the system.
<i>Change Email Interface (Figure 28)</i>			
12.	Update Email	Button	This button allows the user to save the new email in the database.
	Back	Icon	This icon allows the user to go back to the “Profile” interface.
<i>Change Password Interface (Figure 29)</i>			
14.	Update Password	Button	This button allows the user to save the new password in the database.
	Back	Icon	This icon allows the user to go back to the “Profile” interface.
<i>Admin interfaces</i>			
<i>Manage Users Tab (Figure 30)</i>			
16.	Add User	Icon	This icon allows the admin to go to the “Add User” interface.
	Remove User	Icon	This icon allows the admin to remove existing users from the users' table.

NO.	Object	Type	Action
18.	Users Table	Table	Allow the admin to select a user.
<i>Add User Tab (Figure 31)</i>			
19.	Add User	Button	This button allows the user to add the new user.
20.	Back	Icon	This icon allows the user to go back to the “Manage User” interface.
<i>Remove User Tab (Figure 32)</i>			
21.	Remove User	Button	This button allows the user to remove users.
22.	Back	Icon	This icon allows the user to go back to the “Manage User” interface.
<i>Rebuild Model Tab (Figure 33)</i>			
23.	Upload Data File	Button	This button allows the user to upload a data file.
24.	Diseases	Drop down list	The user can select one of the diseases.
25.	Generate Model	Button	This button starts building the model. A summary of the model is displayed.
26.	Back	Button	This button allows the user to go back to the “Manage User” interface.
27.	Update Diagnostic Model	Button	This button stores the chosen model with accuracy.
28.	Delete Diagnostic Model	Button	This button Deletes the chosen model with accuracy.
<i>Medical Specialist Interfaces (Figure 34)</i>			
<i>Patient History Interface (Figure 35)</i>			
29.	Search	Search Box	This search box allows the medical specialist to search for the patient’s diagnosis history using the MRN.
30.	Filter	Drop down list	The user can select one of the filters to filter the Diagnose History table
31.	Diagnose History	Table	The table displays the patient’s history.
<i>Diagnose Interface (Figure 36-57)</i>			
32.	Diagnose	Button	This button displays the “Results” interface.
33.	Back	Icon	This icon allows the user to go back to the “Patient History” interface.
<i>Results Interface (Figure 58)</i>			
34.	Print	Button	This button allows the user to print the diagnosis result.
35.	Back	Icon	This icon allows the user to go back to the “Patient History” interface.

NO.	Object	Type	Action
<i>Laboratory Specialist Interfaces (Figure 59)</i>			
<i>Laboratory Test Interface (Figure 60)</i>			
36.	Filter	Drop down list	The laboratory specialist can select one or more of the diseases.
37.	Retrieve	Button	This button checks if the patient exists in the system or not.
38.	Diagnose	Button	This button will save the information and the prediction to the database.
<i>Registered Users Profile Interfaces (Figure 61)</i>			
<i>Registered Users Diagnose History Tab (Figure 62)</i>			
39.	Search	Search box	This search box allows the user to search through the users' diagnosis history.
40.	Filter	Drop down list	The user can select one of the filters to filter the Diagnose History table
41.	Diagnose History	Table	This table shows the diagnosis history of the user.
<i>Diagnose Interface (Figure 36-57)</i>			
42.	Diagnoses	Button	This button displays the "Results" interface.
43.	Back	Icon	This icon allows users to go back to their diagnosis history.
<i>Results Interface (Figure 58)</i>			
44.	Print	Button	This button allows the user to print the diagnosis result.
45.	Back	Icon	This icon allows the user to go back to the "Diagnosis History" interface.

Table 70 Screen Object and Actions

### 5.3.5 Other Interface

To make the system more professional, the system needs to include other interfaces such as error, confirmation, and information messages. Therefore, it will inform the user before implementing any functions or if any error has happened during the process.

#### 5.3.5.1 Error Messages

Error messages give a straightforward depiction of the errors the user has made and how to stay away from these errors. In our system, the error messages show up underneath the invalid information field as opposed to showing as spring-up messages to maintain a strategic distance from an excessive number of snaps. The accompanying Figures (63-76) show samples of error messages in the system.

Figure 63 shows the error message displayed for the invalid username and/or password in the “Login” interface.

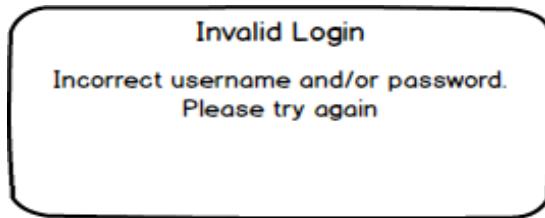


Figure 63 Invalid Login Error message

Figure 64 shows the error message displayed for the third invalid username and/or password in the “Login” interface.

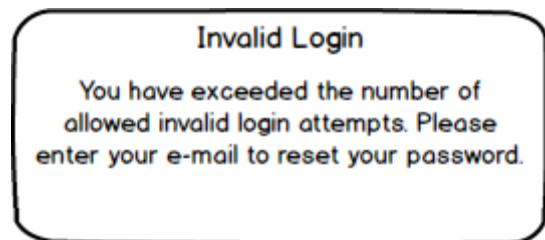


Figure 64 Third Invalid Login Error Message

Figure 65 shows the error message displayed when an expired temporary password is used to Login into the “Login” interface.

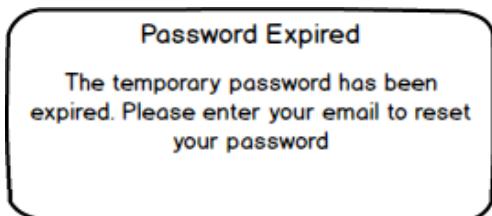


Figure 65 Password Expired Error Message

Figure 66 shows the error message displayed in the “Retrieve Password” interface when the entered email does not exist.

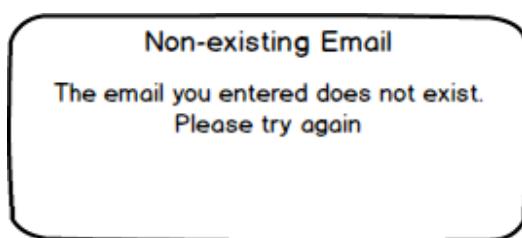


Figure 66 Non-existing Email Error Message

Figure 67 shows the error message displayed for the third invalid email in the “Retrieve Password” interface.

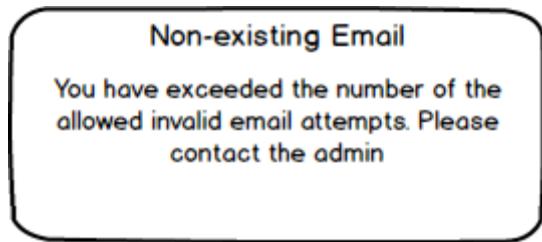


Figure 67 Third Invalid Email Error Message

Figure 68 shows the error message displayed for invalid email format in the “Change Email”, “Add User”, “Create Account”, and “Retrieve Password” interface.

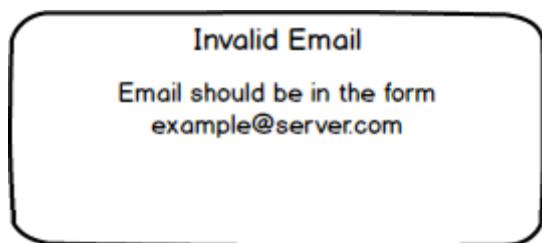


Figure 68 Invalid Email Error Message

Figure 69 shows the error message displayed when the new email and confirmation email do not match in the “Change Email” interface.

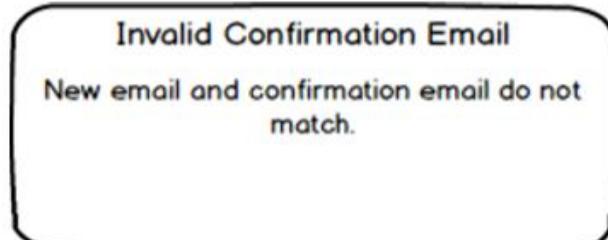


Figure 69 Invalid Confirmation Email Error Message

Figure 70 shows the error message displayed for an invalid password in the “Change Password” and “Create Account” interface.

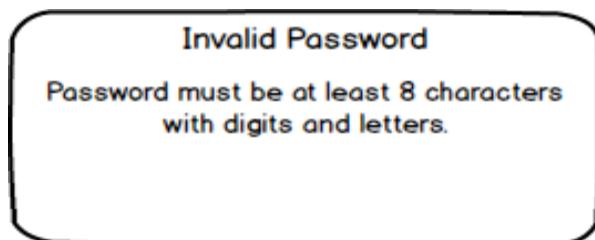


Figure 70 Invalid Password Error Message

Figure 71 shows the error message displayed when the new password and confirmation password do not match in the “Change Password” and “Create Account” interface.

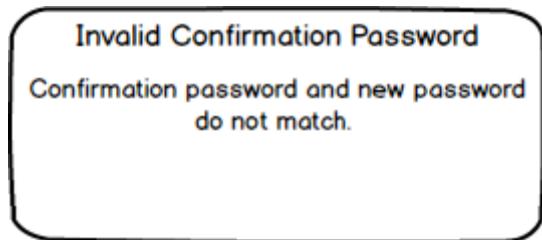


Figure 71 Invalid Confirmation Password Error Message

Figure 72 shows the error message displayed for an invalid old password in the “Change Password” interface.

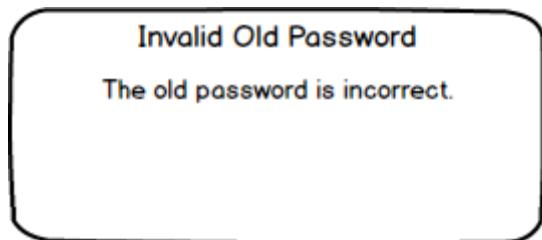


Figure 72 Invalid Old Password Error Message

Figure 73 shows the error message displayed for incomplete required information.

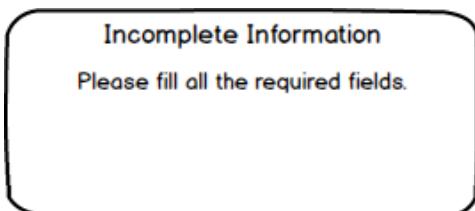


Figure 73 Incomplete Information Error Message

Figure 74 shows the error message displayed for the invalid dataset file extension in the “Rebuild Model” interface.

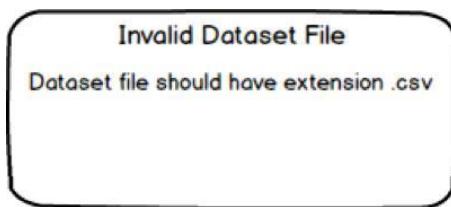


Figure 74 Invalid Dataset Error Message

Figure 75 shows the error message displayed for invalid MRN in the “View History” and “Diagnose” interfaces.

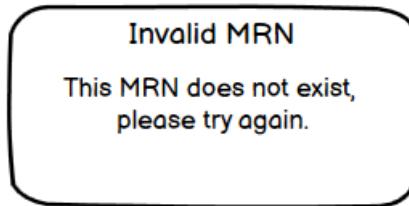


Figure 75 Invalid MRN Error Message

Figure 76 shows the error message displayed for the invalid name in the “Update Profile”, “Diagnosis”, and “Add User” interfaces.

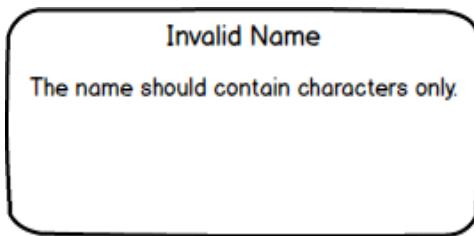


Figure 76 Invalid Name Error Message

#### 5.3.5.2 Confirmation Messages

The following Figures (77 - 80) show multiple samples of the confirmation messages displayed when the user implements major procedures to confirm their action.

Figure 77 shows the confirmation message that asks the user to save the changes in each of “Change Email” and “Change Password” interfaces.

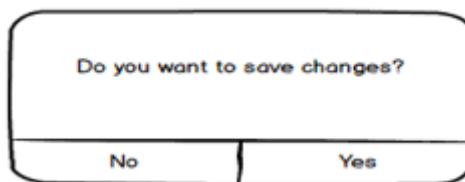


Figure 77 Save Changes Confirmation Message

Figure 78 shows the confirmation message for updating the diagnosis model in the “Rebuild Model” interface.

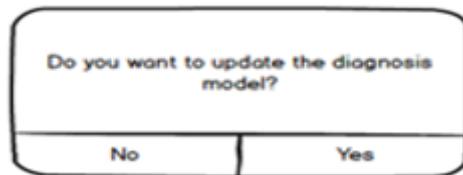


Figure 78 Update Model Confirmation Message

Figure 79 shows the confirmation message for deleting a user in the “Remove User” interface.

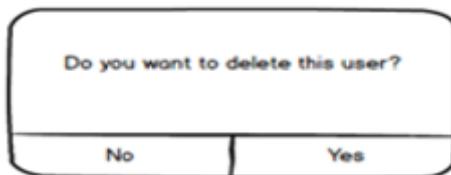


Figure 79 Remove User Confirmation Message

Figure 80 shows the confirmation message for deleting an account in the registered users’ “Profile” interface.

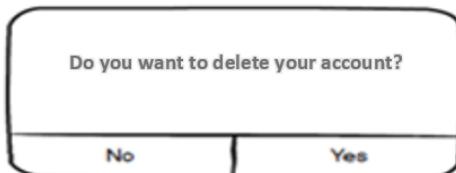


Figure 80 Deleting an Account Confirmation Message

### 5.3.5.3 Information Messages

Figures (81-85) show a sample from the information messages in the system, which will display the result from the major action he/she implements.

Figure 81 shows the information message when the user submits his/her email in the “Retrieve Password” interface.

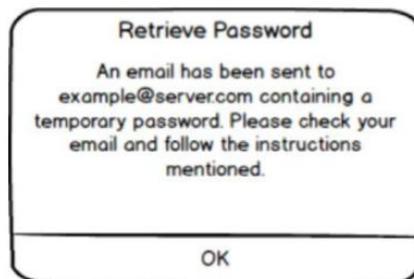


Figure 81 Retrieve Password Information Message

Figure 82 shows the information message when the user updates the diagnosis model in the “Rebuild Model” interface.

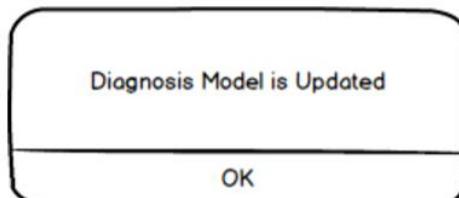


Figure 82 Update Model Information Message

Figure 83 shows the information message when the user adds a new user to the system from the “Add User” interface.

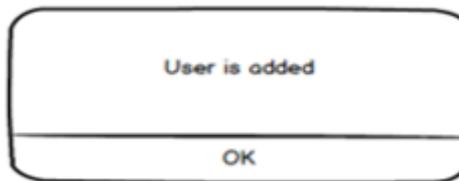


Figure 83 User Added Information Message

Figure 84 shows the information message when the admin deletes the user from the system in the “Delete User” interface.

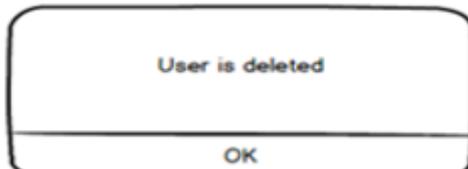


Figure 84 User Deleted Information Message

Figure 85 shows the information message for deleting an account from the system in the registered users’ “Profile” interface.

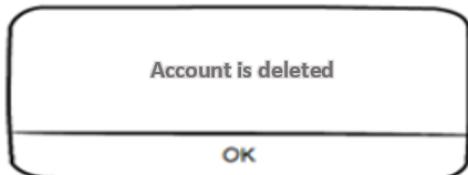


Figure 85 Deleting an Account Information Message

## 5.4 System Architecture

The system uses N-tier layered architecture for building system architecture. The N-tier contains three independent layers as follows:

- **Presentation Layer:** The presentation layer communicates directly with the user interface. This layer can return data from the data layer by communicating with the business layer. The presentation layer transmits the proposed user’s input to the business layer. Then, the business layer will process and manipulate the data from the data layer to provide the output.
- **Business layer:** It is the middle layer that manages the components of the system, which serve as a link between the user and the application menus.
- **Data layer:** This architecture's base layer. This layer's primary concern is everything linked to data functions, such as system data storage and retrieval. The system is often stored on a database server or a file server. It may also be by any other instrument that supports accessing the data with the existence of baring them, but without demonstrating its data storage and retrieval approaches.

The system can gain a benefit from the N-tier architecture in the following aspects:

- **Performance and scalability:** Through N-tier architecture, the system can add resources to each layer without having an impact on other layers while adding. Also, the performance and scalability grow up while using the N-tier architecture in the system.
- **Security and safety:** In the N-tier architecture, the direct way between layers is less, which increases security and safety. Every layer can have an independent secure way, that boosts security and safety.
- **Maintainability:** The N-tier architecture supports maintainability, which refers to updating any processes in the system without having an impact on the system functionality.

- **Flexibility:** The N-tier architecture provides flexibility of the system, which makes it possible to increase each layer if it was needed.

#### 5.4.1 Architectural Design

As previously mentioned, there are three layers of N-tier architecture. Each of these layers contains specific functionalities:

- **Presentation Layer:** This layer requires components such as buttons, tables, combo boxes, etc. to activate the interaction between the users and the system. In this layer, (UI components) will be utilized to build the system's interfaces, each with its components.
- **Business Layer:** For building the interfaces of the system, Python language is needed at this layer. It is like a connection between interfaces and the database. Using the database is for retrieving any wanted data and fetching new data from an interface to store it in the database.
- **Data Layer:** It is the lowest layer where the data is stored and retrieved from the database or a file system.
- 

Figure 86 presents the architectural design diagram for the system

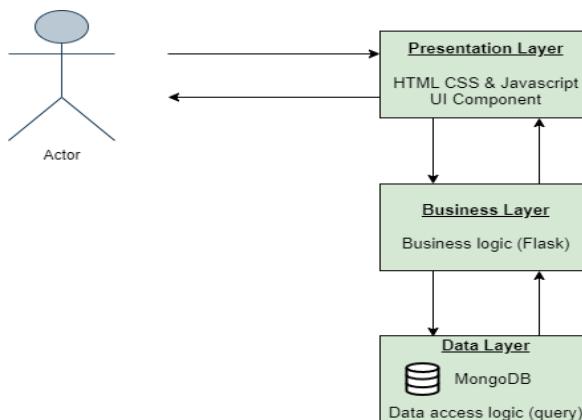


Figure 86 Architectural Design Diagram

#### 5.4.2 Subsystem Architecture

This section provides a decomposition of the subsystems in the architectural design in terms of their functionalities. It determines how the data is being stored in the application and how it is handled. Data Flow Diagrams (DFDs), as represented in this section's figures show the logical architecture to specify the processes, data flows, and data stores at different levels.

##### 5.4.2.1 General View of the System

Figure 87 below, presents a context DFD that shows how both users are exchanging the information with the system. The environment relationships are shown in the context of DFD the application boundaries.

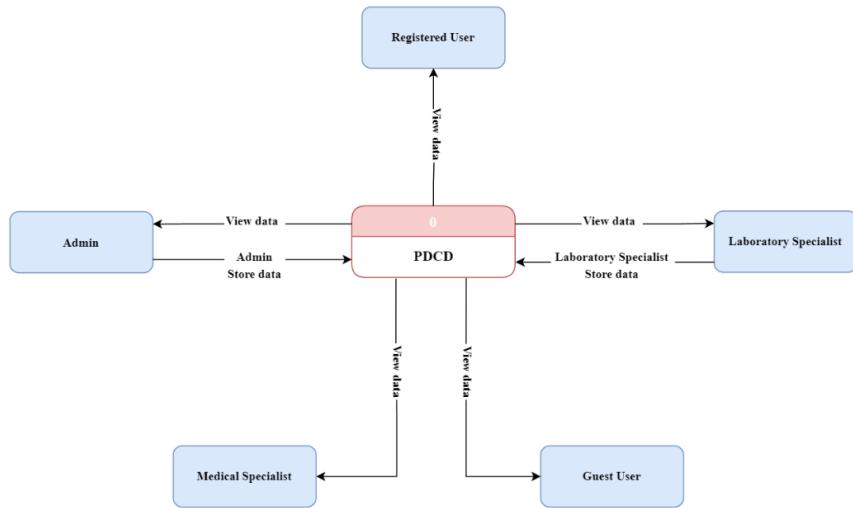


Figure 87 General View of the System

#### 5.4.2.2 All Users' Subsystems

All the users have a common function, like login. Also, forgetting the password is a common function except for the guest user. Figure 88 shows these functions.

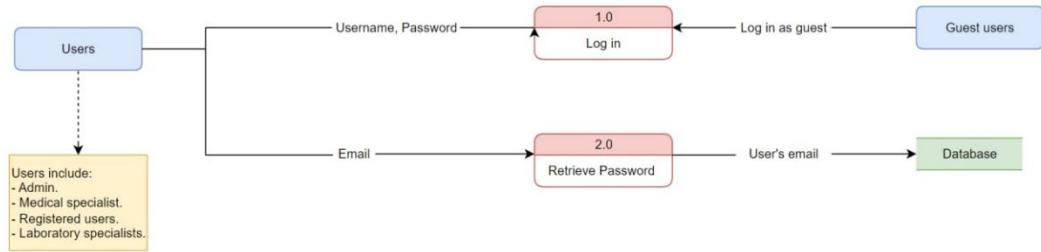


Figure 88 All users' subsystems

##### 5.4.2.2.1 Admin Subsystem

Figure 89 describes the interactions of the admin with the system which were rebuilding the diagnosis model, replacing the diagnosis model, adding the user, deleting the user, and updating the profile where they were mentioned in section 5.3.3.2.

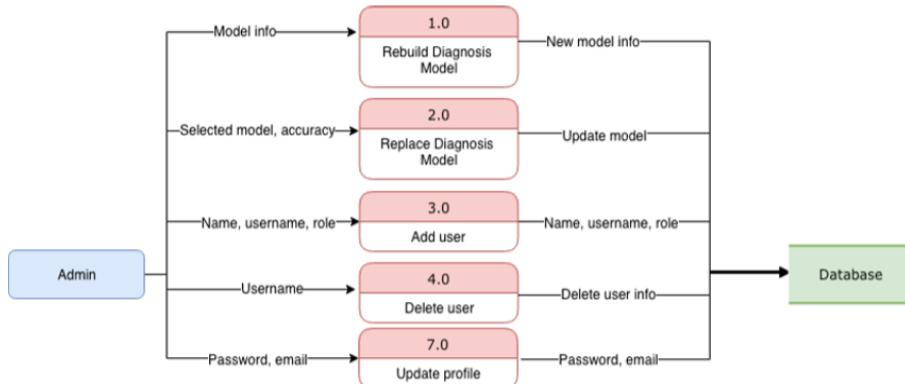


Figure 89 Admin Subsystem

#### 5.4.2.2.2 Medical Specialist Subsystem

Figure 90 describes the interactions of the medical specialist with the system, which were run diagnoses, view diagnosis history, and update profile where they were mentioned in section 5.3.3.3.

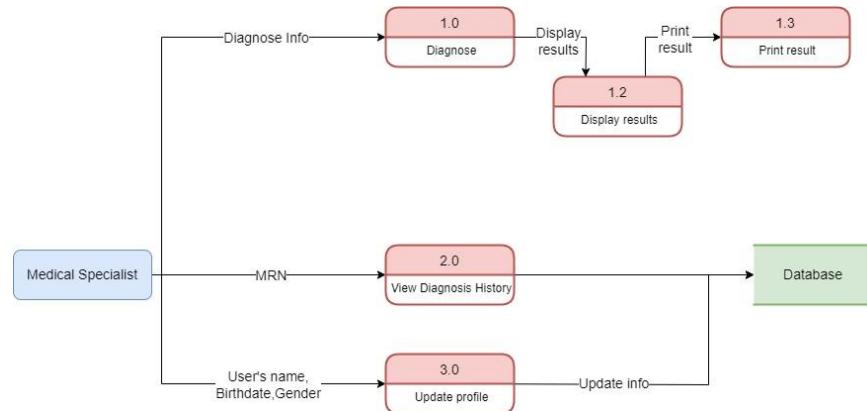


Figure 90 Medical Specialist Subsystem

#### 5.4.2.2.3 Laboratory Specialist Subsystem

Figure 91 describes the interactions of the laboratory specialist with the system, which were run diagnose, and update profile where they were mentioned in section 5.3.3.4.

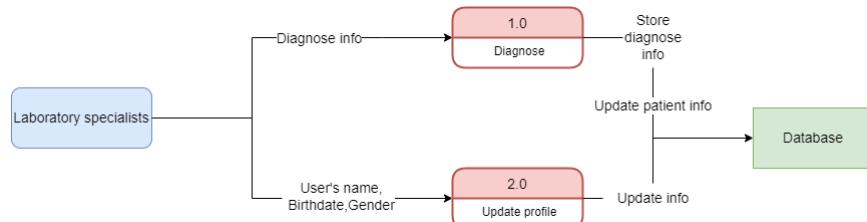


Figure 91 Laboratory Specialist Subsystem

#### 5.4.2.2.4 Registered User Subsystem

Figure 92 describes the interactions of the registered user with the system which were run diagnose, view diagnose history, update profile, create an account and delete account Where they were mentioned in section 5.3.3.5.

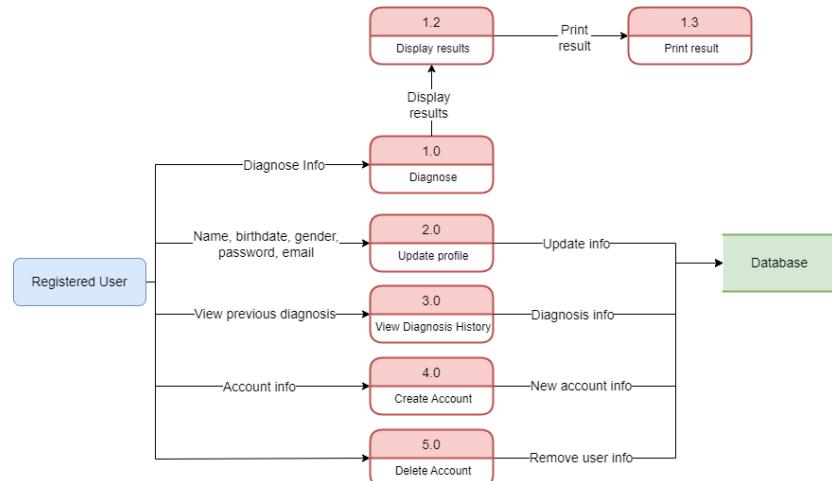


Figure 92 Registered User Subsystem

#### 5.4.2.2.5 Guest User Subsystem

Figure 93 describes the interactions of the guest user with the system which they can only run a diagnose, where they were mentioned in section 5.3.3.6.



Figure 93 Guest User Subsystem

## 5.5 Data Design

This section includes a description of the data, data types, required fields, and a description of the application's database.

### 5.5.1 Data Description

The system contains all the required data to have complete functionality. Every user can access the needed data depending on their role in the system. Table 71 shows the data entities, their fields, type, and constraints.

Entity	Field	Type	Constraints
Account	Username	VARCHAR (40)	Not null, Primary key
	Name	VARCHAR (45)	Not null
	Email	VARCHAR (255)	Not null
	Role	VARCHAR (45)	Not null
	Password	VARCHAR (25)	Not null
	TempPassFlag	ENUM ('0', '1')	Not null
	EmailConfirmed	ENUM ('0', '1')	Default value=0
	EmailConfirmationSentOn	DATETIME	

Patient	MRN	INT (11)	Not null, Primary key, Auto increment
	Username	VARCHAR (40)	Not null, Foreign key
	BirthDate	DATE	Not null
	Gender	ENUM ('Male', 'Female')	Not null
Test	TestID	INT (11)	Not null, Primary key, Auto increment
	Date	DATE	Not null
	MRN	VARCHAR (25)	Not null, Foreign key
	Feature	VARCHAR (25)	Not null
Model	Pred_Disease	Disease	VARCHAR (25)
		Prediction	ENUM ('0', '1')
	ModelID	VARCHAR (25)	Not null, Primary key
	Name	VARCHAR (45)	Not null
TempModel	DiseaseType	VARCHAR (45)	Not null
	Accuracy	DECIMAL (5,2)	Not null
	Active	ENUM ('0','1')	Not null
	TotalInstances	INT (11)	
TempModel	TestInstances	INT (11)	
	TrainingPercentage	DECIMAL (5,2)	Not null
	ModelID	VARCHAR (25)	Not null, Primary key
	ModelName	VARCHAR (25)	Not null
TempModel	Disease Type	VARCHAR (45)	Not null
	Accuracy	DECIMAL (5,2)	Not null
	TotalInstances	INT (11)	
	TestInstances	INT (11)	
TempModel	TrainingPercentage	DECIMAL (5,2)	Not null

Table 71 The Database Entities and Fields

### 5.5.2 Data Dictionary

Table 72 shows the data dictionary of this system list of all the entities, fields, and descriptions.

Entity	Field	Description
Account	Username	Defined user's username. This attribute should be unique for every user.
	Name	User's full name.
	Email	User's email address.
	Role	User's role.
	Password	User's account password.
	EmailConfirmed	A flag that indicates whether the email was confirmed.
	EmailConfirmationSentOn	Date and Time of the sent email.
	TempPassFlag	A flag that indicates if the password in the password field is temporary.
Patient	MRN	Patient MRN, this attribute uniquely identifies the patient within the hospital.

	Username	Defined user's username. This attribute should be unique for every user.
	BirthDate	Patient's date of birth in order to know the age.
	Gender	Patient's gender ("Male", "Female").
	TestID	This attribute should be unique for every diagnosis Result ID result.
	MRN	Patient MRN, this attribute uniquely identifies the patient within the hospital.
Test	Date	The diagnosis result date.
	Feature	Features set of the diseases
Pred_Disease	Disease	Disease name
	Prediction	A binary number is used to indicate whether the patient has the disease or not.
	ModelID	This attribute should be unique for every model.
	ModelName	The model's name with the extension.
	DiseaseType	The disease type that the model was built for.
Model	Accuracy	The diagnosis accuracy of the model.
	Active	A flag that indicates the model used for diagnosis.
	TotalInstances	The number of records in the csv file.
	TestInstances	The number of records used for testing the model.
	TrainingPercentage	The training percentage of the model.
Temp Model	ModelID	This attribute should be unique for every model.
	ModelName	The model's name with the extension.
	Disease Type	Disease name
	TotalInstances	The number of records in the csv file.
	TrainingPercentage	The training percentage of the model.
	Accuracy	The diagnosis accuracy of the model.
	TestInstances	The number of records used for testing the model.

Table 72 Data Dictionary

### 5.5.3 Database Description

In this project, the database within an open source is required. MongoDB database will be used to complete the system process. The system contains five tables, and each table has common characteristics. Figure 94 in the following Entity Relationship Diagram (ERD) shows the entities and the relationship in the system.

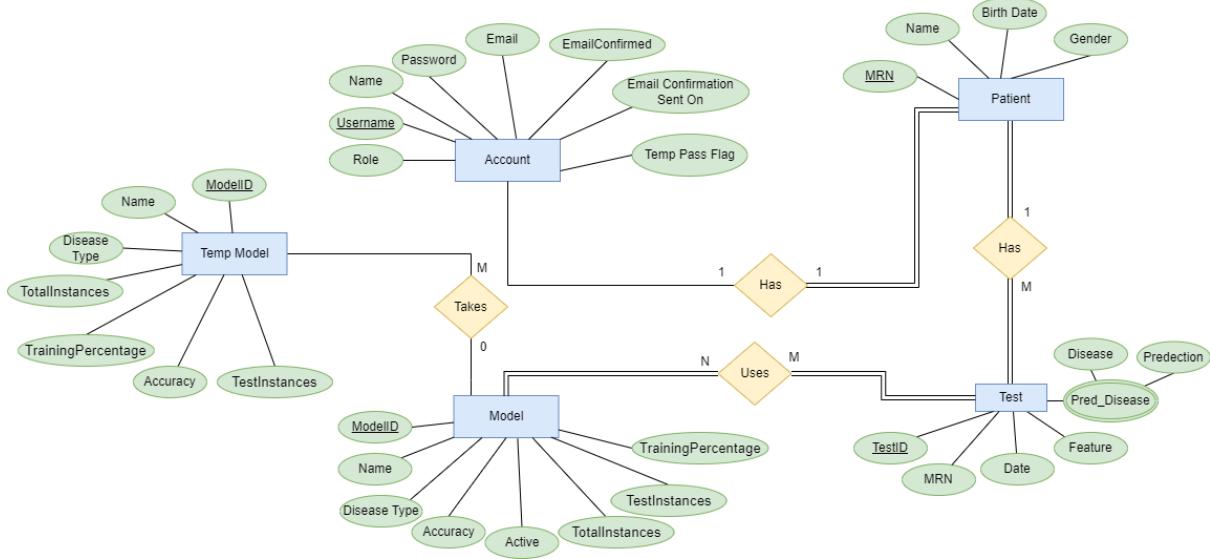


Figure 94 Entity Relationship Diagram

To logically understand the database design. Figure 95 shows the relational schema diagram (ERD Mapping) translated to the (ERD).

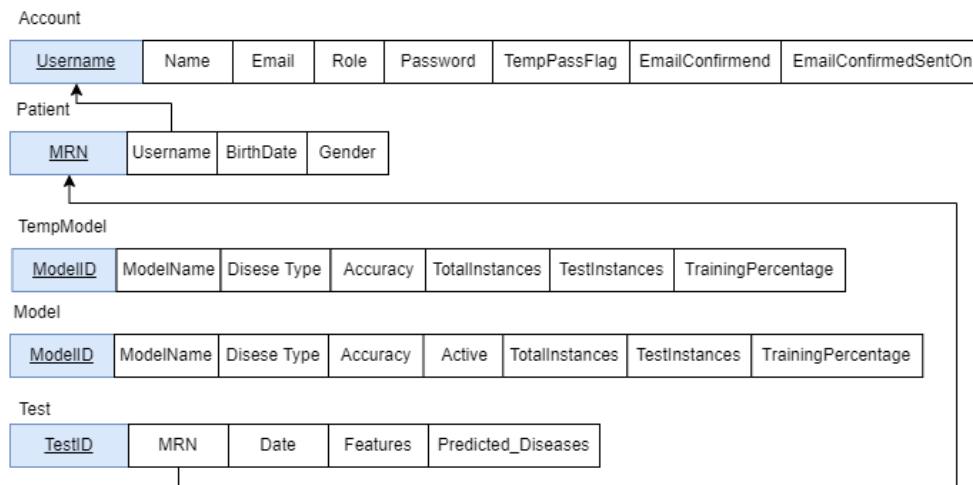


Figure 95 ERD Mapping

## 5.6 Component Design

This section involves the pseudocode of the functional requirements. Pseudocode algorithms are written in informal language to solve problems to help programmers develop algorithms since they are written in English instead of programming language syntax. In the following subsections, pseudocode algorithms will be used for describing the main system's functionalities.

## **5.6.1 Login Function**

### **5.6.1.1 Login\_With\_Credentials ()**

Inputs Username, Password

If username and password match password in database:

If password is temporary:

    If temp password expired:

        Forward to retrieve password interface

    Else:

        Forward to reset password interface

    Else:

        Go to user's homepage

    Else:

        If incorrectlyEnteredPassword counter >3

            Forward to Retrieve Password interface

    Else:

        Display error message

        Increment incorrectly entered password counter

### **5.6.1.2 Login\_As\_Guest ()**

Forward to Diagnosis interface

### **5.6.1.3 Retrieve\_Password ()**

Input Email address

If email format is invalid:

    Display error message

Else if email address exists in the database:

    System sends an email with random password to user's email address

Else if email address does not exist and number of times email is entered >3:

    Display error message

## **5.6.2 Account Functions**

### **5.6.2.1 Create\_Account ()**

Input username, name, email, birthdate, gender, password, and confirmPassword

If at least one of the fields empty:

    Display error message

Else if username is not unique:

    Display message Re-enter username

Else if name contains digits:

    Display error message

Else if email not in the correct format:

    Display error message

Else if password is less than 8 characters or does not contain digits or letters:

    Display error message

Else if password and confirm password do not match:

    Display error message

Else:

    Add new account record and new Patient record

    Create account

#### **5.6.2.2 Change\_Password ()**

Input old password, new password, confirm password

If at least one of the fields empty:

    Display error message

Else if old password is invalid:

    Display error message

Else if new password is less than 8 characters or does not contain digits or letters:

    Display error message

Else if new password and confirm password do not match:

    Display error message

Else:

    Update password

    Send an email to confirm the changes

#### **5.6.2.3 Change\_Email ()**

Input current email, new email, confirm email

If at least one of the fields is empty:

    Display error message

Else if current email does not match user's email:

    Display error message

Else if new email is not in the correct format:

    Display error message

Else if new email and confirm email do not match:

    Display error message

Else:

    Update email

    Send an email to confirm the changes

#### **5.6.2.4 Change\_Personal\_Info ()**

Input new name, birthdate, or change gender

If name contains digits:

    Display error message

Else:

    Update changes

### **5.6.3 Admin Functions**

#### **5.6.3.1 Add\_User ()**

Input name, username, intialPassword, email, and select a user role

If name or username field is empty:  
    Display error message  
Else if the username was taken:  
    Display message Re-enter username  
Else if name field has digit numbers:  
    Display error message Re-enter name  
Else if new user role is not selected:  
    Display error message  
Else if email format is invalid:  
    Display error message  
Else:  
    Save the added information to the User table in the database

#### **5.6.3.2 Remove\_User ()**

Input username  
If the Admin typed in the username field:  
    Populate matching users from the user table Admin selects user from table to remove

#### **5.6.3.3 Rebuild\_Model (Dataset, trainingPercent, diseaseType)**

If Dataset = empty or not type .csv:  
    Display error message  
If trainingPercent = empty or not valid value:  
    Display error message  
If diseaseType not selected:  
    Display error message  
Else:  
    Split Dataset to trainingSet and testingSet using trainingPercent  
    Train machine learning model using trainingSet  
    Extract the trained model and store it to the PC with a modelName derived from the time  
        data                               of its creation.  
    accuracy = validate trained model with testingSet  
    Insert modelName, trainingPercent, diseaseType, and accuracy to the Temp Model table  
        into the database

#### **5.6.3.4 Replace\_Model (modelID)**

Retrieve modelName, diseaseType, and accuracy from the Model table in the database  
uploadDate = get today's date  
Update Model database table with Active = 1 to Active = 0 with same disease  
Insert modelID, modelName, diseaseType, uploadDate, and accuracy and set Active = 1 to  
Model table in the database  
Delete the previous record where modelID's = modelID

## **5.6.4 Diagnostic Functions**

### **5.6.4.1 View\_History ()**

Input MRN

If the MRN is empty:

    Display error message

Else if the MRN exists in the database:

    Retrieve TestID, Date, DiseaseType, Result and ResultAccuracy automatically from the database.

Else:

    Display error message

### **5.6.4.2 Diagnosis ()**

Input diagnosticInformation

If at least one of the diagnostic information is not in its correct format:

    Display error message

Result = Apply diagnosis model on diagnosticInformation

Display Result, diseaseType and accuracy in the result interface

### **5.6.4.3 Print\_Result ()**

Input MRN and diseaseType

Display MRN, testResult, diseaseType and modelAccuracy in a PDF format

Send the print job to the operating system

### **5.6.4.4 Overall\_Diagnosis ()**

Select from the diseases

Input MRN, demographical data, and diagnosticInformation

If at least one of the diagnostic information is not in its correct format:

    Display error message

Else if MRN does not exist:

    Display error message

Generate a TestID

Result= Apply all the selected disease diagnosis model on diagnosticInformation

Store TestID, MRN, and diagnosticInformation in the database

Display Result, diseaseType and accuracy in the result interface

## **5.7 Detailed System Design**

This section provides the classification and definition of each system component with their constraints and collaboration with other components. Furthermore, it provides the application resources, uses/interaction, processing, interface/exports, and detailed subsystems design. Sequence and Unified Modeling Language (UML) diagrams are used to clarify the processes and interactions between subsystems.

### 5.7.1 Classification, Definition, and Responsibilities

This section includes each system component's classification, definition, and responsibilities as shown in Table 73.

Component	Classification	Definition and Responsibility
<i>Login</i>	Function	This function gives different users, except for the guest users, the ability to log into the system. The users should provide a valid username and password to successfully access the system.
<i>Common Functions between Admin, Laboratory Specialist, Medical Specialist, and Registered User Subsystem</i>		
<i>Retrieve Password</i>	Function	This function allows the users to retrieve a password by sending a temporary password via email.
<i>Update Profile</i>	Function	This function allows the users to update their profiles. The information that can be changed is the email and password. An email is sent to confirm the changes. Also, for the registered user, the name, birthdate, and gender can be changed.
<i>Common function between Medical Specialist and Registered User Subsystem</i>		
<i>View Diagnosis History</i>	Function	This function gives the users the ability to view previous diagnosis results if they existed.
<i>Common Function between Laboratory Specialist, Medical Specialist, Registered User, and Guest User Subsystem</i>		
<i>Diagnose</i>	Function	This function gives users the ability to perform the diagnosis process by providing the required medical information.
<i>Common Function between Medical Specialist, Registered User and Guest Subsystem</i>		
<i>Print Result</i>	Function	This function allows users to print their diagnosis results by converting the result into PDF format then sending them to the printer.
<i>Admin Subsystem</i>		
<i>Rebuild Diagnosis Model</i>	Function	This function allows the admin to periodically rebuild a diagnosis model using the most recent dataset to train the model.
<i>Replace Diagnosis Model</i>	Function	This function allows the admin to upload the model with the desired accuracy in the diagnosis process.
<i>Add User</i>	Function	This function allows the admin to add a new medical specialist, laboratory specialist, patient or another admin by providing their name, email, username, password and role. The username and password for the new user are sent via email.

<i>Remove User</i>	Function	This function allows the admin to remove a user from the system.
<i>Registered User Subsystem</i>		
<i>Create Account</i>	Function	This function allows the user to create an account to be a registered user in the system.
<i>Laboratory Specialist Subsystem</i>		
<i>Overall Diagnosis</i>	Function	This function allows the Laboratory to make different disease diagnosis and store it in the database.

Table 73 Component, Classification, Definition and Responsibility

### 5.7.2 Constraints and Composition

This section provides the constraints for each system component. Constraints include pre-conditions and post-conditions for the system component, as shown in Table 74.

<b>Component</b>	<b>Limitation</b>	<b>Pre-condition</b>	<b>Post-condition</b>
<i>Login</i>	The user must be already registered in the system except for the guest user.	Provide a valid username and password. For guest users, they should login by clicking “login as guest”.	The system authorizes the user to successfully login.
<i>Common Functions between Admin, Laboratory Specialist, Medical Specialist, and Registered User Subsystem</i>			
<i>Retrieve Password</i>	The user must be already registered in the system.	Provide valid registered email.	The system sends a temporary password via the registered email.
<i>Update Profile</i>	The user must be logged into the system.	Provide valid new information.	Information is updated.
<i>Common Function between Medical Specialist and Registered User Subsystem</i>			
<i>View Diagnosis History</i>	The user must have previous diagnosis records.	The medical specialist should provide a valid MRN.	Previous diagnosis results are displayed.
<i>Common Function between Medical Specialist, Laboratory Specialist, Registered User, and Guest User Subsystem</i>			
<i>Diagnose</i>	All required information must be complete and valid.	Provide the required information.	The diagnosis result is displayed.
<i>Common Function between Medical Specialist, Registered User and Guest Subsystem</i>			
<i>Print Result</i>	Diagnosis is performed.	Click “Print Result” button.	The diagnosis results will be converted to PDF format and sent to the printer.

Admin Subsystem			
<i>Rebuild Diagnosis Model</i> <i>Update Diagnosis Model</i> <i>Add User</i> <i>Remove User</i>	The dataset must be available in a suitable format.	Provide the dataset and training percentage.	A diagnosis model is generated.
	At least one diagnosis model was generated.	Provide diagnosis model.	Provide diagnosis model.
	The user to be added must be a laboratory specialist, medical specialist, patient or the admin.	Provide the name, email, username, password and role of the new user.	- The user is added. - Username and password are sent via email.
	The user to be deleted must be registered in the system.	Provide a valid username.	The user is deleted.
Registered User Subsystem			
<i>Create Account</i>	The user must fill all the required information.	Provide name, email, gender, date of birth, and password.	An account is created.
<i>Delete Account</i>	The user must click on a button to delete his/her account.	Click a button and confirm account deletion.	An account is deleted.
Laboratory Specialist Subsystem			
<i>Overall Diagnosis</i>	The user must be already registered in the system.	Fill out all the required information.	The diagnostic and result information will be stored in the database.

Table 74 Component, Limitation Pre-condition, and post-condition

### 5.7.3 Uses/Interactions

This section demonstrates how system components interact with each other using sequence diagrams.

#### 5.7.3.1 Login

Figure 96 shows the “Login” sequence diagram.

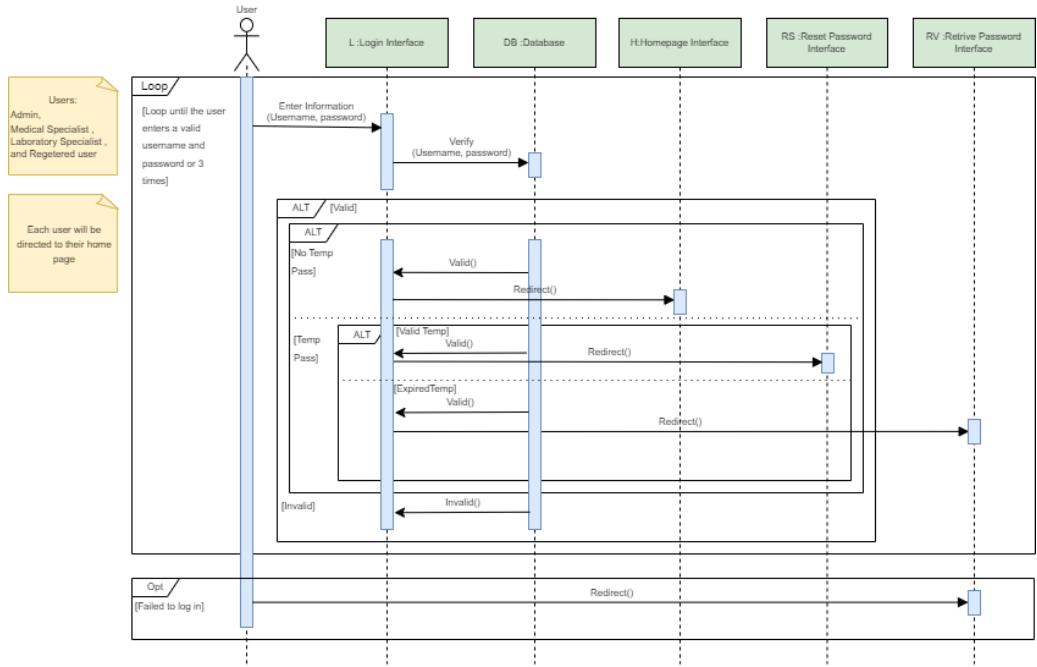


Figure 96 The “Login” sequence diagram

### 5.7.3.2 Retrieve Password

Figure 97 shows the “Retrieve Password” sequence diagram.

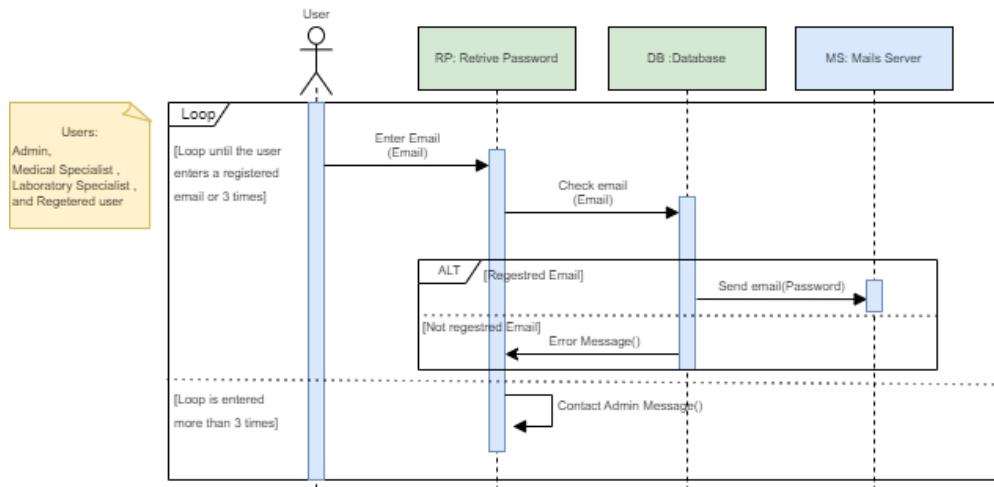


Figure 97 The “Retrieve Password” sequence diagram.

### 5.7.3.3 Update Profile

The “Update Profile” functionality is divided into three sub functionalities: “Change Email”, “Change Password” and “Personal Information”. Figure 98 shows the “Change Email” sequence diagram.

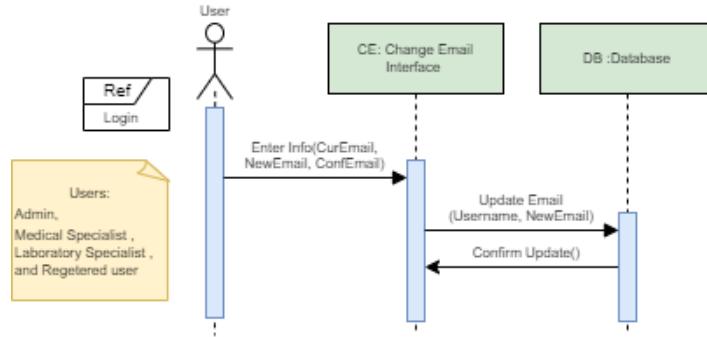


Figure 98 The “Change Email” sequence diagram

Figure 99 shows the “Change Password” sequence diagram.

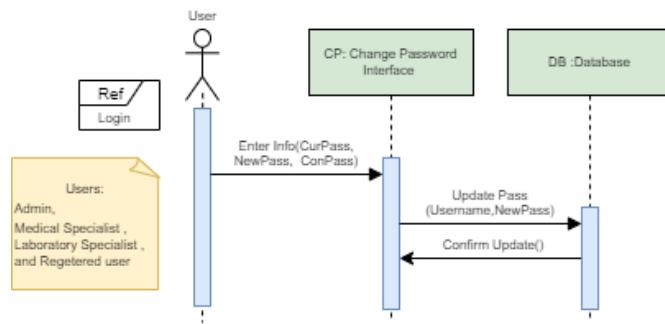


Figure 99 The “Change Password” sequence diagram

Figure 100 shows the “Change Personal Information” sequence diagram.

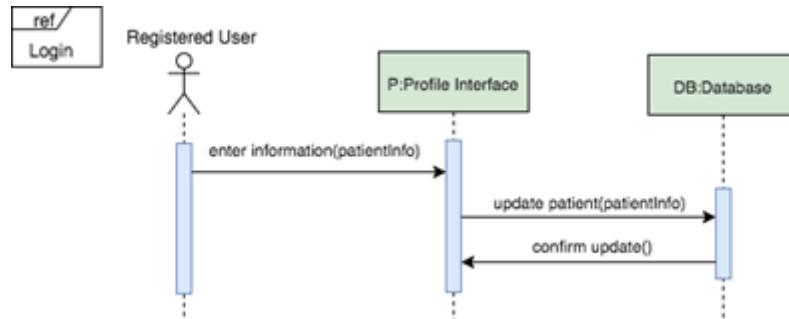


Figure 100 Change Personal Information

#### 5.7.3.4 View Diagnosis History

Figure 101 shows the “View Diagnosis History” sequence diagram.

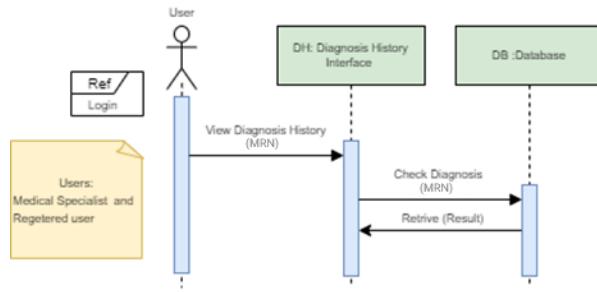


Figure 101 The “View Diagnosis History” sequence diagram

### 5.7.3.5 Diagnose

Figure 102 shows the “Diagnose” sequence diagram.

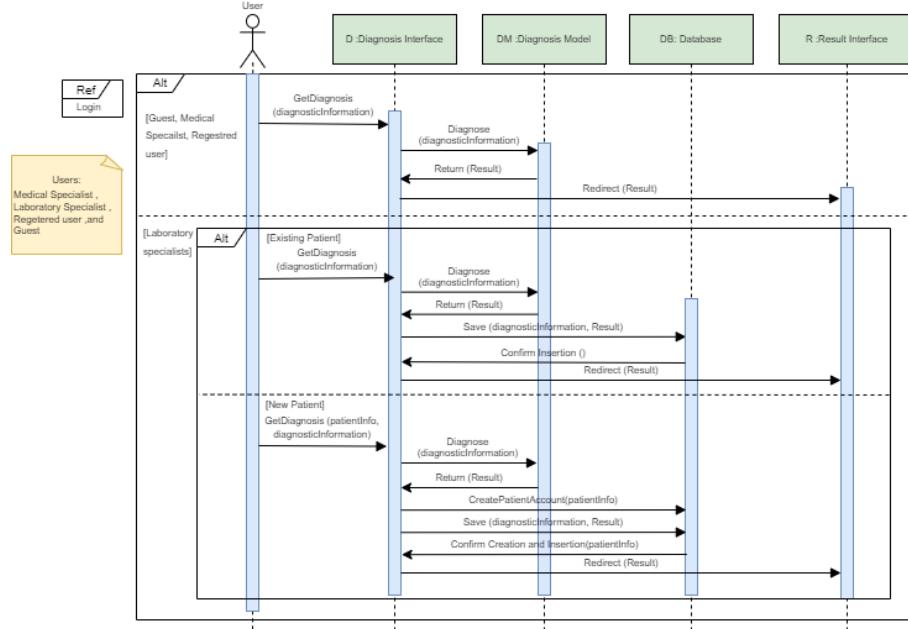


Figure 102 The “Diagnose” sequence diagram.

### 5.7.3.6 Print Result

Figure 103 shows the “Print Result” sequence diagram.

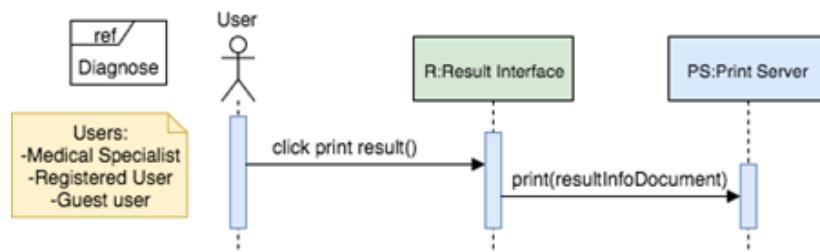


Figure 103 The “Print Result” sequence diagram

### 5.7.3.7 Rebuild Diagnosis Model

Figure 104 shows the “Rebuild Diagnosis Model” sequence diagram.

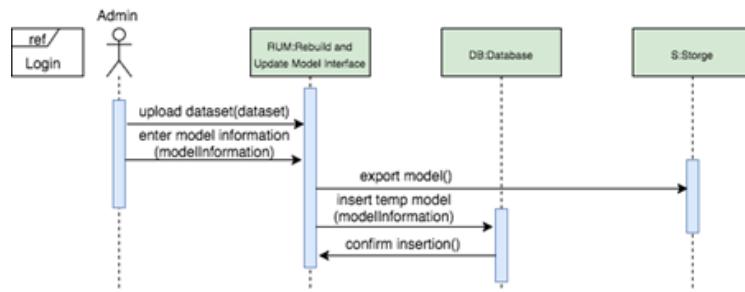


Figure 104 The “Rebuild Diagnosis Model” sequence diagram

### 5.7.3.8 Update Diagnosis Model

Figure 105 shows the “Update Diagnosis Model” sequence diagram.

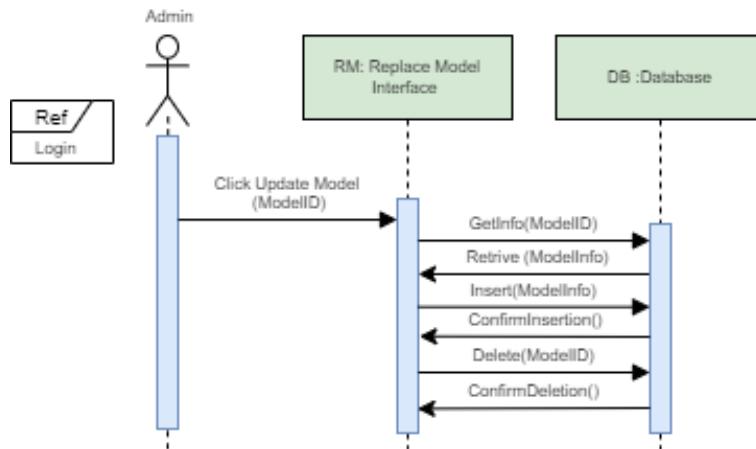


Figure 105 The “Update Diagnosis Model” sequence diagram

### 5.7.3.9 Add User

Figure 106 shows the “Add User” sequence diagram.

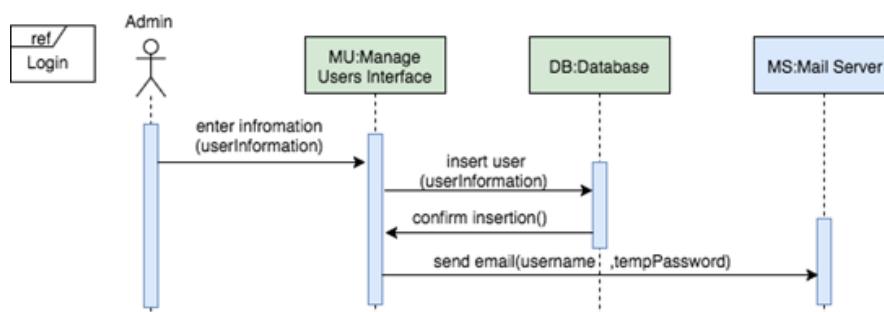


Figure 106 The “Add User” sequence diagram

### 5.7.3.10 Remove User

Figure 107 shows the “Remove User” sequence diagram.

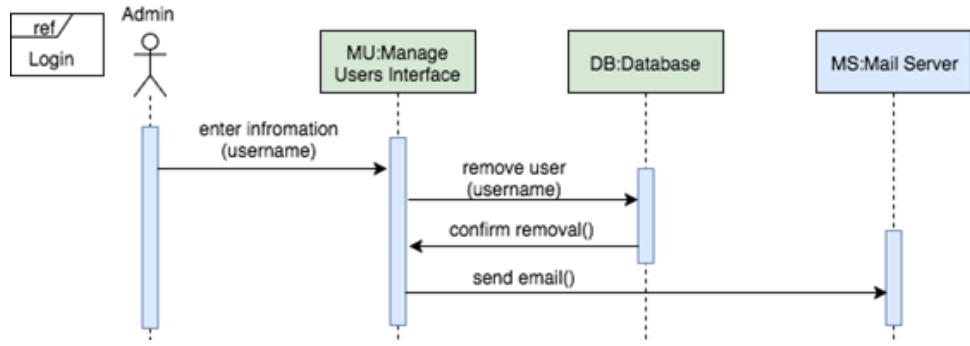


Figure 107 The “Remove User” sequence diagram

### 5.7.3.11 Create Account

Figure 108 shows the “Create Account” sequence diagram.

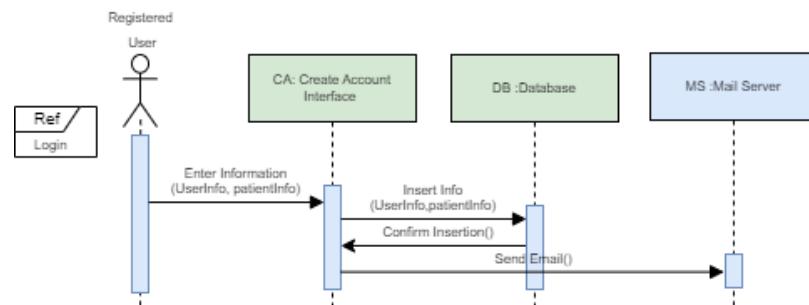


Figure 108 The “Create Account” sequence diagram

### 5.7.3.12 Overall Diagnosis

Figure 109 shows the “Overall Diagnosis” sequence diagram.

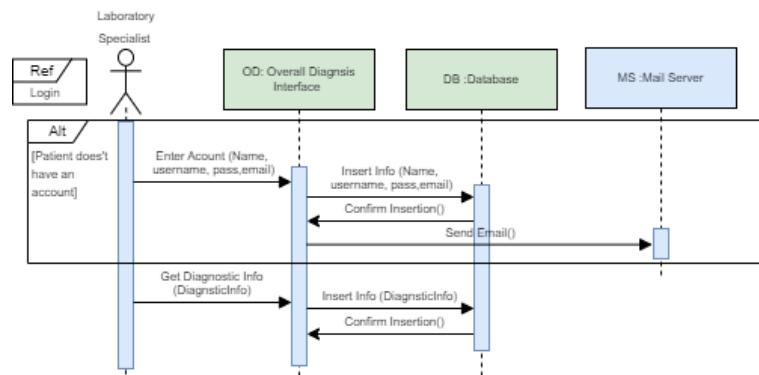


Figure 109 The “Overall Diagnosis” sequence diagram

## 5.7.4 Processing

Each component has inputs, outputs, and descriptions. The following subsections provide a description of how each component performs its functionalities.

#### 5.7.4.1 Login

Table 75 presents the description, inputs, outputs, and constraints of the “Login” component.

<b>Description</b>	<ul style="list-style-type: none"> <li>• A user should be able to log into the system by providing their username and password.</li> <li>• The system checks the validity of the entered information against the data stored in the database.</li> <li>• A user should be redirected to the homepage after successful verification.</li> <li>• A user logs in with a temporary password if they are first registered to the system or if they forgot their password.</li> <li>• If the user logs in using a temporary password before the password is expired, they should be redirected to the “Change Password” interface. However, if the temporary password is expired, the user is redirected to the “Retrieve Password” interface to generate a new temporary password.</li> <li>• For the third invalid login, the system should forward the user to the “Retrieve Password” interface.</li> <li>• For the third invalid entry of the email, the user should contact the admin.</li> </ul>
<b>Input</b>	<ul style="list-style-type: none"> <li>• Username.</li> <li>• Password.</li> </ul>
<b>Output</b>	<p>Admin, Medical specialist, Laboratory specialist, and Registered user:</p> <ul style="list-style-type: none"> <li>• Access Homepage.</li> </ul> <p>Guest:</p> <ul style="list-style-type: none"> <li>• Access the “Diagnosis” interface.</li> </ul>
<b>Constraint</b>	The username and password should be valid.

Table 75 Description of “Login” Component

#### 5.7.4.2 Retrieve Password

Table 76 presents the description, inputs, outputs, and constraints of the “Retrieve Password” component.

<b>Description</b>	<ul style="list-style-type: none"> <li>• If the user forgot their password, they should be able to reset their password by entering their registered email. Then, an email containing a temporary random password is sent to that email if it existed.</li> <li>• The user should be able to access the system using the temporary password before it is expired.</li> </ul>
<b>Input</b>	Email.
<b>Output</b>	A temporary password is stored in the database and sent by email to the user.
<b>Constraints</b>	The entered email should be valid and in the correct format.

Table 76 Description of “Retrieve Password” Component

#### 5.7.4.3 Update Profile

Table 77 presents the description inputs, outputs, and constraints of the “Update Profile” component.

Description	<ul style="list-style-type: none"><li>The users should be able to update their profiles by changing their passwords or emails. Registered users can also change their name, birthdate, and gender.</li><li>The system checks the validity of new information.</li></ul>
Input	Password, email, name, birth date, or gender.
Output	<ul style="list-style-type: none"><li>Information updated.</li><li>Notification email is sent for updating email/password.</li></ul>
Constraints	<ul style="list-style-type: none"><li>The email format should be valid.</li><li>The new email and confirmation email should match.</li><li>The old password should be valid.</li><li>The new password should be at least 8 alphanumeric characters.</li><li>The new password and confirm password should match.</li><li>The name should not include numeric digits.</li></ul>

Table 77 Description of “Update Profile” Component

#### 5.7.4.4 View Diagnosis History

Table 78 presents the description inputs, outputs, and constraints of “View Diagnosis History” component.

Description	<ul style="list-style-type: none"><li>The medical specialist can view the history of the patients if it exists.</li><li>The registered user could view their previous diagnosis results if they existed.</li></ul>
Input	MRN is required only from the medical specialist.
Output	Display diagnosis history.
Constraints	The patient needs to have an account to have MRN.

Table 78 Description of “View Diagnosis History” Component

#### 5.7.4.5 Diagnose

Table 79 presents the description, inputs, outputs, and constraints of the “Diagnose” component.

Description	<ul style="list-style-type: none"><li>The laboratory and medical specialist must be able to diagnose patients as having Diabetes Mellitus, Chronic Kidney Disease, Coronary Heart Disease, Asthma, Thyroid Cancer, Schizophrenia, Glaucoma, Alzheimer's Disease, Lung Cancer, Rheumatoid Arthritis, Hypothyroidism, Prostate Cancer, Cervical Cancer, Multiple Sclerosis, Liver Cirrhosis, Chronic Obstructive Pulmonary Disease, Parkinson's</li></ul>
-------------	---

	<p>Disease, Hepatitis C, Depression, Epileptic Seizure, Osteoporosis and Sickle Cell Anemia.</p> <ul style="list-style-type: none"> <li>Registered and guest users must be able to perform the diagnosis process.</li> </ul>
<b>Input</b>	<p>For diagnosing Diabetes Mellitus:</p> <ul style="list-style-type: none"> <li>Sugar Level</li> <li>Hematocrit Level</li> <li>Mean Platelet Volume (MPV)</li> </ul> <p>For diagnosing CKD:</p> <ul style="list-style-type: none"> <li>Blood Urea Nitrogen</li> <li>Creatinine</li> </ul> <p>For diagnosing CHD:</p> <ul style="list-style-type: none"> <li>Gender</li> <li>Age</li> <li>Mean Corpuscular Hemoglobin (MCH)</li> <li>Mean Corpuscular Hemoglobin Concentration (MCHC)</li> <li>Red Cell Distribution Width (RDW)</li> <li>Platelet Count</li> <li>MPV</li> <li>Hemoglobin</li> <li>Neutrophil Granulocyte Instrument</li> <li>Basophil Instrument %</li> <li>Basophil Instrument Absolute</li> <li>Neutrophil Granulocyte Instrument Absolute</li> <li>Mononucleosis Absolute</li> <li>Potassium</li> <li>Anion Gap</li> <li>Gamma-glutamyl transpeptidase (GGTP)</li> <li>Serum glutamic-oxaloacetic transaminase (SGOT)</li> <li>Serum glutamic-pyruvic transaminase (SGPT)</li> </ul> <p>For diagnosing Asthma Disease:</p> <ul style="list-style-type: none"> <li>Gender</li> <li>Age</li> <li>Basophil Instrument %</li> <li>Hematocrit</li> <li>Hemoglobin</li> <li>MCH</li> <li>MCHC</li> <li>MPV</li> <li>White Blood Cell count</li> </ul>

For diagnosing Thyroid Cancer:

- Gender
- Age
- Hematocrit
- MCHC
- MPV
- Red Blood Cell count
- White Blood Cell count

For diagnosing Schizophrenia:

- Age
- White Blood Cell
- Hemoglobin
- Hematocrit
- Mean Corpuscular Volume (MCV)
- MCH
- MCHC
- Platelet
- MPV
- Aspartate Aminotransferase (AST)
- Total Protein
- Gamma Glutamyl transferase (GGT)

For diagnosing Glaucoma:

- At
- Ean
- Mhci
- Vasi
- Varg
- Vars
- Tmi

For diagnosing Alzheimer's Disease:

- Gender
- Age
- Pulse Ox
- Respiratory Rate
- BP - Diastolic
- White Blood Cell count
- Red Blood Cell count
- Hemoglobin
- Hematocrit
- MCV

- MCH
- RDW
- MPV

For diagnosing Lung Cancer:

- Gender
- Age
- Smoking
- Yellow Fingers
- Anxiety
- Wheezing
- Peer Pressure
- Chronic Disease
- Fatigue
- Allergy
- Coughing
- Alcohol
- Shortness Of Breath
- Swallowing Difficulty
- Chest Pain

For diagnosing Rheumatoid Arthritis:

- Gender
- Age
- Albumin
- Alkaline phosphatase
- Blood Urea Nitrogen
- Chloride
- Carbon dioxide
- Creatinine
- Direct Bilirubin
- GGTP
- Hemoglobin
- Hematocrit Level
- Potassium
- Lactic Acid Dehydrogenase
- MCH
- MPV
- MCHC
- MCV
- Sodium
- Platelet
- Red Blood Cell Count

- RDW
- SGOT
- SGPT
- Total Bilirubin
- Total Protein

For diagnosing Hypothyroidism:

- Age
- BP - Systolic
- Respiratory Rate
- MCV
- Pulse Ox

For diagnosing Prostate Cancer:

- Perimeter
- Area
- Smoothness
- Compactness

For diagnosing Cervical Cancer:

- STDs: Number of Diagnosis
- STDs Condylomatosis
- STDs Syphilis
- STDs HIV
- STDs HPV
- Dx
- Dx: CIN
- Dx: HPV

For diagnosing Multiple Sclerosis:

- Age
- Alanine Transaminase (ALT)
- Lactate Dehydrogenase (LDH)
- Creatinine
- Blood Urea Nitrogen
- Total Bilirubin
- GGT
- Alkaline Phosphatase
- AST
- Platelet
- BP – Systolic

For diagnosing Liver Cirrhosis:

- Gender

- Age
- N\_Days
- Hepatomegaly
- Spiders
- Edema
- Cholesterol
- Copper
- SGOT
- Platelet
- Prothrombin
- Ascites
- Serum Bilirubin
- Albumin
- Alkaline phosphatase
- Triglycerides
- Drug
- Status

For diagnosing Chronic Obstructive Pulmonary Disease:

- Gender
- Age
- Smoking
- Imagery part minimum
- Imagery part average
- Real part minimum
- Real part average

For diagnosing Parkinson's Disease:

- Gender
- Age
- Anion Gap
- ALT
- LDH
- White Blood Cells
- Red Blood Cells
- Hemoglobin
- Hematocrit
- Sodium
- Potassium
- Chloride
- Carbon Dioxide
- Creatinine

- Total Protein
- Albumin
- Blood Urea Nitrogen
- Total Bilirubin
- Direct Bilirubin
- GGT
- MCV
- MCH
- MCHC
- Alkaline Phosphatase
- RDW
- AST

For diagnosing Hepatitis C:

- Age
- Total Protein
- Total Bilirubin
- Direct Bilirubin
- GGT
- Alkaline Phosphatase
- Lymphocyte - Instrument %
- Neutrophil Granulocyte - Instrument Absolute
- Platelet
- Basophil - Instrument %
- BP – Systolic
- Fall Risk - Morse
- Body Mass Index
- International Normalized Ratio

For diagnosing Depression:

- Age
- Household Size
- Education Level
- Value of livestock
- Value of durable goods
- Value of savings
- Land owned
- Consumed Alcohol
- Consumed Tobacco
- Education expenditure
- Non-ag business flow expenses, monthly
- Livestock sales and meat revenue, monthly

- Total expenses, monthly
- Whole days without food
- Non-durable Investments
- Amount received using M-Pesa
- Marital status
- Children
- hh\_children
- Non-agricultural business owner
- Saved money using M-Pesa
- Early Survey

For diagnosing Epileptic Seizure:

- NumberOfNonPsych Comorbidities
- NumberOfPrior AEDs
- Asthma
- Migraine
- Chronic Pain
- Diabetes
- non metastatic cancer
- NumberOfNonSeizureNonPsych
- Medication
- NumberOfCurrent AEDs
- Baseline
- MedianDurationOfSeizures
- NumberOfSeizureTypes
- InjuryWithSeizure
- Catamenial
- Trigger of sleep deprivation
- Aura
- IctalEyeClosure
- IctalHallucinations
- Oralautomatisms
- Incontinence
- LimbAutomatisms
- IctalTonic-clonic
- MuscleTwitching
- HipThrusting
- Post-ictalFatigue
- HeadInjury
- PsychTraumaticEvents
- Concussionw/oLOC
- Concussionw/LOC

	<ul style="list-style-type: none"> <li>• SevereTBI</li> <li>• Opioids</li> <li>• SexAbuse</li> <li>• PhysicalAbuse</li> <li>• Rape</li> </ul> <p>For diagnosing Osteoporosis:</p> <ul style="list-style-type: none"> <li>• Gender</li> <li>• Age</li> <li>• Weight</li> <li>• Height</li> <li>• Diabetes</li> <li>• Hypothyroidism</li> <li>• SeizureDisorder</li> <li>• Alcohol</li> <li>• Smoking</li> <li>• EstrogenUse</li> <li>• JointPain</li> <li>• HistoryOfFracture</li> <li>• Dialysis</li> <li>• Family History of Osteoporosis</li> <li>• Maximum Walking distance</li> <li>• Daily Eating habits</li> <li>• BMI</li> <li>• Site</li> <li>• Obesity</li> </ul> <p>For diagnosing Sickle Cell Anemia:</p> <ul style="list-style-type: none"> <li>• Sex</li> <li>• Tribe</li> <li>• HB</li> <li>• MCV</li> <li>• MCH</li> <li>• MCHC</li> <li>• RBC</li> <li>• TWBC</li> <li>• PLT</li> <li>• PCV</li> </ul> <p>For laboratory specialists to diagnose a patient, there are additional data:</p> <ul style="list-style-type: none"> <li>• MRN.</li> </ul>
<b>Output</b>	<ul style="list-style-type: none"> <li>• Diagnosis results and accuracy are displayed.</li> </ul>

	<ul style="list-style-type: none"> <li>The diagnosis by the laboratory specialist is saved in the database.</li> <li>If a patient needs to be diagnosed by a laboratory specialist and the account does not exist, the laboratory needs to create an initial account for the patient.</li> </ul>
<b>Constraints</b>	Diagnostic information must be complete and valid

*Table 79 Description of “Diagnose” Component*

#### 5.7.4.6 Print Result

Table 80 presents the description, inputs, outputs, and constraints of “Print Result” component.

<b>Description</b>	For printing diagnosis results, it will be converted to PDF format and sent to the printer.
<b>Input</b>	-
<b>Output</b>	Printed diagnosis results.
<b>Constraints</b>	The printer should be configured with the system.

*Table 80 Description of “Print Result” Component*

#### 5.7.4.7 Rebuild Diagnosis Model

Table 81 presents the description, inputs, outputs, and constraints of the “Rebuild Diagnosis Model” component.

<b>Description</b>	The admin must be able to reconstruct the diagnosis model periodically by uploading the most recent patients’ dataset that is used to train the model.
<b>Input</b>	<ul style="list-style-type: none"> <li>Dataset file.</li> <li>Training percentage.</li> <li>Disease type.</li> </ul>
<b>Output</b>	<ul style="list-style-type: none"> <li>Diagnosis model is generated, and its accuracy is produced.</li> <li>New diagnosis model’s information is stored.</li> </ul>
<b>Constraints</b>	Dataset file should have a valid extension.

*Table 81 Description of “Rebuild Diagnosis Model” Component*

#### 5.7.4.8 Update Diagnosis Model

Table 82 present the description, inputs, outputs, and constraints of the “Update Diagnosis Model” component.

<b>Description</b>	The admin should be able to upload the model with the desired accuracy to be used in the diagnosis process.
<b>Input</b>	<ul style="list-style-type: none"> <li>Diagnosis model.</li> </ul>
<b>Output</b>	The future diagnosis processes will be based on the updated model.
<b>Constraints</b>	At least one diagnosis model was generated.

*Table 82 Description of “Update Diagnosis Model” Component*

#### 5.7.4.9 Add User

Table 83 presents the description, inputs, outputs, and constraints of the “Add User” element.

<b>Description</b>	<ul style="list-style-type: none"> <li>The admin can add new laboratory specialists, medical specialists, and other admins to the system.</li> </ul>
<b>Input</b>	<ul style="list-style-type: none"> <li>User’s name.</li> <li>Username.</li> <li>Initial Password.</li> <li>Email.</li> <li>Role.</li> </ul>
<b>Output</b>	<ul style="list-style-type: none"> <li>The username and temporary password are sent to the new user via email.</li> <li>The user is added to the database.</li> </ul>
<b>Constraints</b>	The email must have a proper format.

*Table 83 Description of “Add User” Component*

#### 5.7.4.10 Remove User

Table 84 presents the description, inputs, outputs, and constraints of the “Remove User” element.

<b>Description</b>	The admin can delete users from the system.
<b>Input</b>	User’s username.
<b>Output</b>	A confirmation message will appear, and the user will be deleted from the database.
<b>Constraints</b>	The username must be valid.

*Table 84 Description of “Remove User” Component*

#### 5.7.4.11 Create Account

Table 85 presents the description, inputs, outputs, and constraints of the “Create Account” element.

<b>Description</b>	<p><b>46.</b> A user can create an account to be registered in the system.</p> <p><b>47.</b> The needed information to create an account includes their full name, username, email, birthdate, gender, and password.</p>
<b>Input</b>	<ul style="list-style-type: none"> <li>User’s full name.</li> <li>Username.</li> <li>Email.</li> <li>Birthdate.</li> <li>Gender.</li> <li>Password.</li> <li>Password Confirmation.</li> </ul>

<b>Output</b>	<ul style="list-style-type: none"> <li>The registered user is a patient by default, their information will be added to the patient table.</li> <li>An account is created.</li> </ul>
<b>Constraints</b>	<ul style="list-style-type: none"> <li>The name must contain characters only.</li> <li>The email must have a proper format.</li> <li>The password and confirm password must be similar.</li> <li>The password must satisfy the password criteria (at least 8 alphanumeric characters).</li> </ul>

Table 85 Description of “Create Account” Component

#### 5.7.4.12 Overall Diagnosis

Table 86 presents the description, inputs, outputs, and constraints of the “Overall Diagnosis” element.

<b>Description</b>	<b>48.</b> The laboratory specialist Can diagnose a patient with selected diseases using one interface.
<b>Input</b>	<ul style="list-style-type: none"> <li>MRN.</li> <li>Diseases to be diagnosed.</li> <li>Demographical data.</li> <li>Blood tests.</li> <li>Symptoms.</li> <li>Other tests.</li> </ul>
<b>Output</b>	<ul style="list-style-type: none"> <li>A diagnosis of the selected diseases will be issued and saved in the database.</li> <li>If the result is positive, the medical specialist will get an alert.</li> <li>An account is created if the patient doesn't have an account.</li> </ul>
<b>Constraints</b>	<ul style="list-style-type: none"> <li>MRN must exist in the database.</li> <li>All the required fields need to fill with the appropriate data type.</li> </ul>

Table 86 Description of “Overall Diagnosis” Component

#### 5.7.5 Interface/Exports

Developers must understand all subsystem’s functionalities to confirm that the system functions serve as expected. They must be informed about the system's user input, pre-conditions, required processes, post-conditions, and constraints. The limitations, pre-conditions, and post-conditions of the system’s functionalities are discussed in section 5.7.2. For the input, output, and functional requirements’ description specifications refer to section 5.7.4. For the activity diagrams, refer to the functional requirements in the SRS section 4.2.2.

#### 5.7.6 Detailed Subsystem Design

This section presents a detailed description of this application component’s behavior and control flow. Assign to the SRS functional requirements section 4.2 for the activity diagrams.

## 5.8 Other Design Features

All the design features are mentioned in the previous sections.

## 5.9 Detailed System Design

Table 87 illustrate the requirement matrix.

Associated ID in SRS	Technical Assumption	Functional Requirement	User	Associated ID in SDS	System Component
4.2.3.2.1 .1	<ul style="list-style-type: none"> <li>The admin should register medical specialists to log into the system.</li> <li>Visitors to the clinic can log into the system as registered users by creating accounts or as guests.</li> </ul>	Login	<ul style="list-style-type: none"> <li>Admin.</li> <li>Medical Specialist.</li> <li>Laboratory Specialist.</li> <li>Registered User.</li> <li>Guest.</li> </ul>	5.6.1.1, 5.6.1.2	Login_With_Credentials (), Login_As_Guest ()
4.2.3.2.1 .2	<ul style="list-style-type: none"> <li>Issued if a user forgets their password or enters an invalid password three times or more.</li> </ul>	Retrieve Password	<ul style="list-style-type: none"> <li>Admin.</li> <li>Medical Specialist.</li> <li>Laboratory Specialist.</li> <li>Registered User.</li> </ul>	5.6.1.3	Retrieve_Password ()
4.2.3.2.1 .3	<ul style="list-style-type: none"> <li>In the case of changing the email address, the user enters the current email address and the new email address twice.</li> <li>In the case of changing the password, the user enters the current password and the new password twice.</li> <li>In the case of changing the personal information of a registered user, the user needs to enter the updated information.</li> </ul>	Update Profile	<ul style="list-style-type: none"> <li>Admin.</li> <li>Medical Specialist.</li> <li>Laboratory Specialist.</li> <li>Registered User.</li> </ul>	5.6.2.2, 5.6.2.3, 5.6.2.4	Change_Password (), Change_Email (), Change_Personal_Info ()
4.2.3.2.1 .4	<ul style="list-style-type: none"> <li>Displays previous diagnosis results.</li> </ul>	View Diagnosis History	<ul style="list-style-type: none"> <li>Medical Specialist.</li> <li>Registered User.</li> </ul>	5.6.4.1	View_History ()

4.2.3.2.1 .5	<ul style="list-style-type: none"> <li>The user should be able to fill the fields needed for medical diagnosis.</li> </ul>	Diagnose	<ul style="list-style-type: none"> <li>Medical Specialist.</li> <li>Laboratory Specialist.</li> <li>Registered User.</li> <li>Guest.</li> </ul>	5.6.4.2	Diagnosis ()
4.2.3.2.1 .6	<ul style="list-style-type: none"> <li>Enables the user to print the diagnosis result.</li> </ul>	Print Result	<ul style="list-style-type: none"> <li>Medical Specialist.</li> <li>Registered User.</li> <li>Guest.</li> </ul>	5.6.4.3	Print_Result ()
4.2.3.2.2 .1	<ul style="list-style-type: none"> <li>The admin should be able to upload data files.</li> <li>Try different training percentages.</li> </ul>	Rebuild Diagnosis Model	<ul style="list-style-type: none"> <li>Admin.</li> </ul>	5.6.3.3	Rebuild_Model (Dataset, trainingPercent, diseaseType)
4.2.3.2.2 .2	<ul style="list-style-type: none"> <li>Upload the diagnostic model.</li> </ul>	Replace Diagnosis Model	<ul style="list-style-type: none"> <li>Admin.</li> </ul>	5.6.3.4	Replace_Model (modelID)
4.2.3.2.2 .3	<ul style="list-style-type: none"> <li>The admin should be able to add users.</li> </ul>	Add User	<ul style="list-style-type: none"> <li>Admin.</li> </ul>	5.6.3.1	Add_User ()
4.2.3.2.2 .4	<ul style="list-style-type: none"> <li>The admin should be able to delete users.</li> </ul>	Remove User	<ul style="list-style-type: none"> <li>Admin.</li> </ul>	5.6.3.2	Remove_User ()
4.2.3.2.5 .1	<ul style="list-style-type: none"> <li>Visitors should be able to register to keep track of their diagnostic records.</li> </ul>	Create Account	<ul style="list-style-type: none"> <li>Registered User.</li> </ul>	5.6.2.1	Create_Account ()
4.2.3.2.4 .1	<ul style="list-style-type: none"> <li>Laboratory Specialist should be able to perform a diagnosis for several diseases using only one page.</li> </ul>	Overall Diagnosis	<ul style="list-style-type: none"> <li>Laboratory Specialist.</li> </ul>	5.6.4.4	Overall_Diagnosis ()

Table 87 Requirement Tractability Matrix

## Chapter 6: Implementation Process

This chapter is divided into four sections: Osteoporosis Empirical Study, Epileptic Seizures Empirical Study, Sickle Cell Anemia Empirical Study, and the website implementation.

### 6.1 Changes from the proposal phase and justifications

A number of changes have been made to improve our final senior project's performance. Firstly, Epilepsy and osteoporosis datasets were replaced with more effective datasets to improve model performance and to increase accuracy. Furthermore, five new approaches have been incorporated to enhance the project's feature selecting capabilities, LogisticRegression, SVM\_Gaussian, GaussianNB, Decision Tree, and AdaBoost. These techniques are computationally efficient and interpretable, also they capture complex non-linear relationships between features, making it effective in high-dimensional spaces. Only MRN is being utilized for patient identification, and the System ID has been eliminated to simplify and expedite the process. Furthermore, the choice has been made to move from SQL to MongoDB, a NoSQL database management system that is capable of handling intricate queries and substantial amounts of unstructured data. This choice was decided to allow the laboratory specialist to manage the project's 167 distinct features, which would be difficult to maintain in a typical SQL database because of their high dimensionality. The success of the project depends on the NoSQL database's ability to enable high-dimensional data management. Lastly, adding patients has been restricted for laboratory specialists with restricted user credentials in order to maintain data integrity and avoid errors in patient identification.

### 6.2 Empirical Study on Osteoporosis Dataset

#### 6.2.1 Data Description

The dataset utilized in this research originates from a health camp organized by the Unani and Panchkarma Hospital in Srinagar, Jammu & Kashmir, India, held from December 21 to December 31, 2019 [69]. The dataset consists of 240 entries, each corresponding to an individual patient and encompassing 26 laboratory biomarkers and demographic features. Noteworthy features include Joint Pain, Gender, Age, Menopause Age, Height (Meter), Weight (KG), Smoking and Alcohol habits, Diabetes, Hypothyroidism, Number of Pregnancies, Seizure Disorder, Estrogen Use, Occupation, History of Fracture, Dialysis, Family History of Osteoporosis, Maximum Walking Distance (km), Daily Eating Habits, Medical History, BMI, Site of examination, Obesity status, and Diagnosis, considering each participant's T-score and Z-score data from their knee X-ray and Quantitative Ultrasound System. All features are illustrated in Table 88 along with their respective types. This dataset provides a comprehensive snapshot of patients' health conditions, facilitating in-depth exploration and analysis for research purposes.

Feature	Type
Joint Pain	Categorical
Gender	Categorical
Age	Integer
Menopause Age	Integer

Height (Meter)	Float
Weight (KG)	Integer
Smoker	Categorical
Alcoholic	Categorical
Diabetes	Categorical
Hypothyroidism	Categorical
Number of Pregnancies	Integer
Seizure Disorder	Categorical
Estrogen Use	Categorical
Occupation	Categorical
History of Fracture	Categorical
Dialysis	Categorical
Family History of Osteoporosis	Categorical
Maximum Walking Distance (km)	Float
Daily Eating Habits	Categorical
Medical History	Categorical
T-score Value	Float
Z-Score Value	Float
BMI	Float
Site of examination	Categorical
Obesity status	Categorical
Diagnosis	Categorical

Table 88 Features' description.

### 6.2.2 Statistical Analysis of the Dataset

This section delves into a comprehensive statistical analysis pivotal for uncovering intrinsic data patterns and determining requisite preprocessing techniques for optimal model preparation. The dataset under scrutiny presents a mix of numerical and categorical features, each bearing significance in the subsequent modeling phase.

The numerical attributes underwent thorough scrutiny through various statistical metrics. These metrics provided profound insights into the inherent characteristics of the data. A detailed statistical breakdown showcasing the properties and attributes of the numerical features is presented in Table 89, offering a profound understanding of the dataset's numeric aspects.

Feature	Mean	Standard deviation	Min	25 <sup>th</sup> quartile	50 <sup>th</sup> quartile	75 <sup>th</sup> quartile	Max	Missing value counts
Age	51.13	13.23	17.00	44.50	50.00	60.00	107.00	0
weight	69.05	9.88	39.00	63.00	69.00	74.50	98.00	0
height	1.58	0.09	1.37	1.52	1.57	1.65	1.82	0
Maximum walking distance (km)	1.94	1.98	0.10	0.50	1.00	3.00	10.00	3
BMI	27.58	4.05	16.13	24.94	27.26	30.21	42.75	3

Table 89 Statistical analysis of numerical features

The categorical features as shown in Figure 110 across the dataset shed light on various patient attributes. The distribution of the target feature indicates that among 240 patients, 154 have osteopenia, 49 have osteoporosis, and 37 are considered normal. Gender distribution showcases 132 female and 108 male patients. Alcohol consumption shows that all the patients do not consume Alcohol. Smoking habits vary, with 199 non-smokers and 41 smokers. 12 patients are diabetic and 228 do not suffer from diabetes. Estrogen intake showcases 229 patients not taking supplements and 11 who do. Joint pain involves 2 patients who do not have any joint pain and 238 patients who suffer from joint pain. 34 patients have Hypothyroidism and 206 do not. Family history of osteoporosis comprises 174 patients with no family history and 66 with a positive history. 8 patients have Seizer Disorder and 232 do not. One patient is on dialysis and 339 are not. Menopause Age, Number of Pregnancies, Occupation, History of Fracture, Maximum Walking distance, Daily Eating habits, Medical History, Site, and Obesity complete the categorical feature analysis, providing a comprehensive view of patient demographics and health-related behaviors.

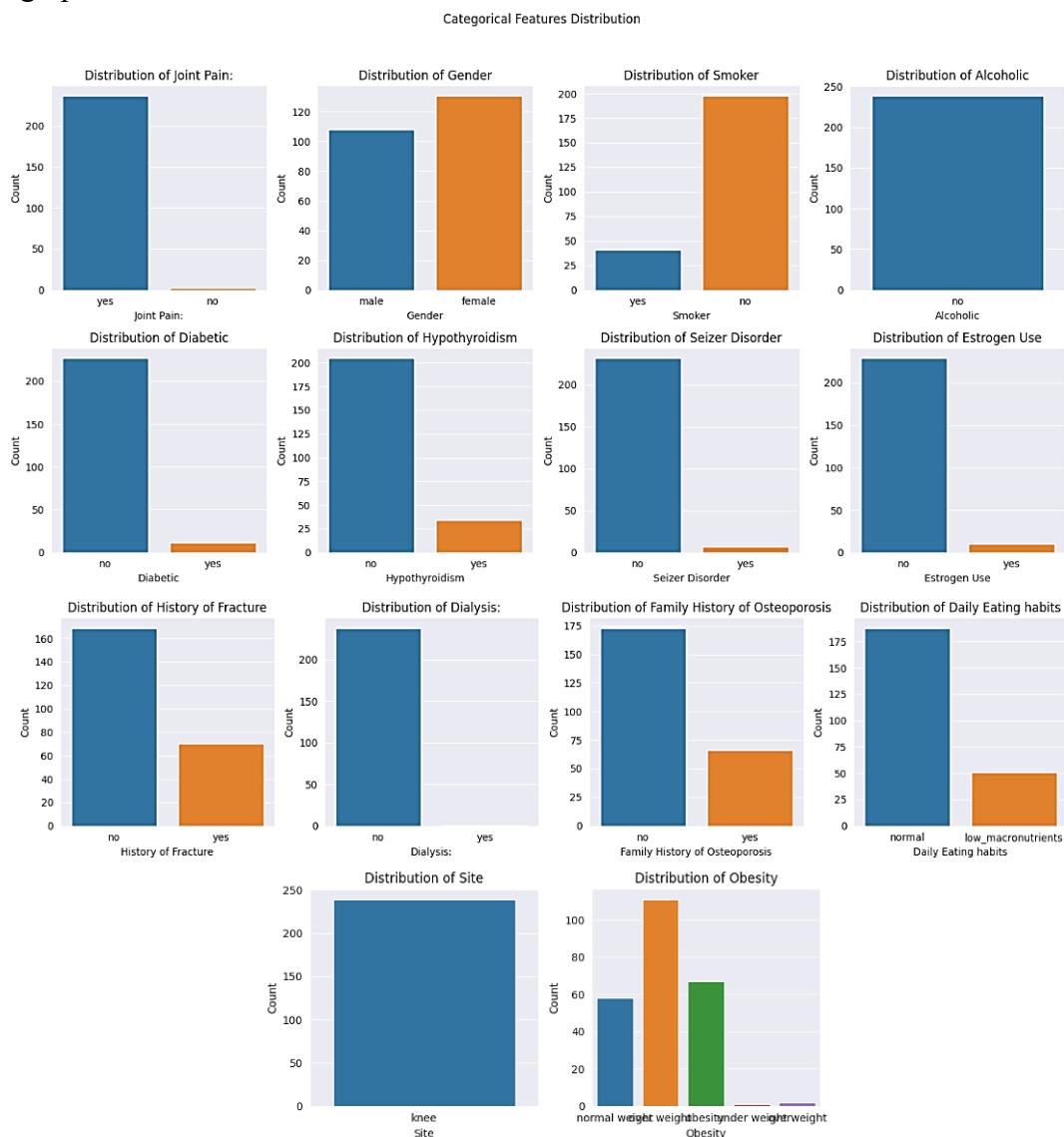


Figure 110 description of the categorical features

### 6.2.3 Experimental Setup

This study developed a pre-emptive model for diagnosing osteoporosis and osteopenia using Python programming language. All operations were conducted with a fixed seed value of 0 to ensure repeatability of results and consistency across runs. As seen in Figure 1, the dataset went through a number of crucial pre-processing steps before modeling to ensure data integrity and feature quality. These steps included handling outliers, checking for duplication, and checking for missing values across all features. After that, LabelEncoder was used to convert the values of categorical columns into numerical representations for better model interpretability. Moreover, to address the class imbalance, the SMOTETomek technique was used—a hybrid sampling method that combines the advantages of SMOTE (Synthetic Minority Over-sampling Technique) and Tomek links. With the use of Tomek links and SMOTE, it attempts to rectify class imbalance by undersampling the majority class and oversampling the minority class. The dataset was then divided into training and testing sets, using 80% for training and 20% for testing, and standardized using StandardScaler for optimal model performance. This method assumes that data is normally distributed and will scale them such that the distribution is centered around zero with a standard deviation of one. Feature selection was conducted using Sequential Forward Feature Selection (SFFS), a method that sequentially chooses the most suitable features for a machine learning model. Five different classifiers—Random Forest, SVM, KNN, Gradient Boosting, and XGBoost—were utilized. For each model, GridSearchCV was employed to tune hyperparameters via cross-validation using 10 folds. The best hyperparameters for each model were identified, and their performance was evaluated on the test set using accuracy, precision, recall, F1-score, and AUC. Following the selection of the model that performed the best, eXplainable Artificial Intelligence (XAI) techniques like SHAP and LIME were used to examine the model's decision-making process in more detail. This entire process is illustrated in Figure 111.

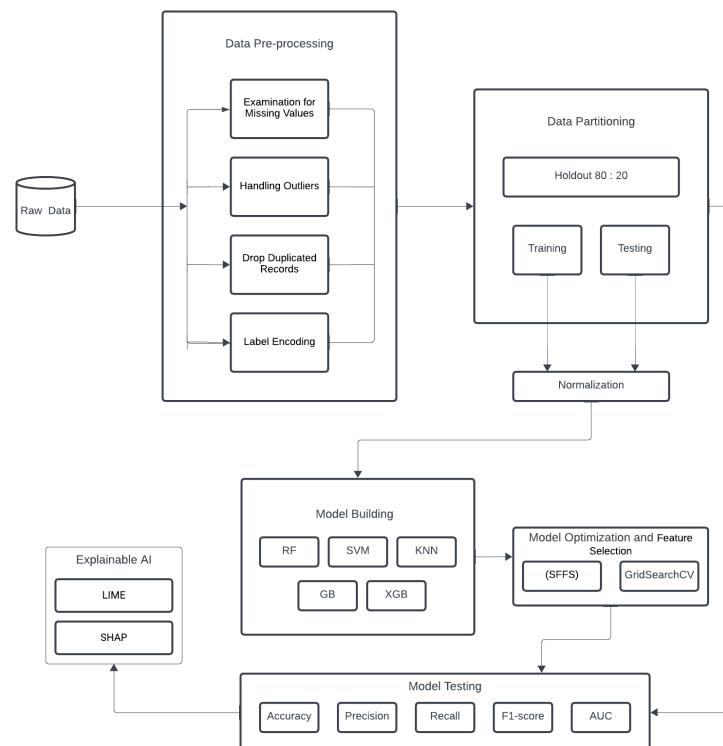


Figure 111 The proposed framework for the pre-emptive diagnosis of Osteoporosis and osteopenia

## **6.2.4 Description of proposed Techniques**

The sections below represent a theoretical background of the classifiers utilized to preemptively predict the possible incidence of osteoporosis disease.

### **6.2.4.1 Support Vector Machine (SVM)**

Support Vector Machine (SVM) was proposed by Cortes and Vapnik in 1990 and has since gained popularity within the machine learning community [76]. Operating as a supervised learning algorithm, SVM addresses both classification and regression problems, with a primary focus on binary classification [77]. SVM establishes a hyperplane in the feature space to separate different classes, ensuring a maximum margin between them [104]. During testing, data points are mapped onto the feature space and categorized based on their position relative to the margin, showcasing SVM's effectiveness in creating robust decision boundaries.

### **6.2.4.2 Random Forest (RF)**

Random Forest, introduced by Leo Breiman in 2001, revolutionized ensemble learning through the concept of bagging or "bootstrap aggregation [105]." Comprising multiple Decision Trees (DTs), the RF classifier avoids overfitting by aggregating predictions through a majority vote mechanism. Instead of relying on a single DT, RF leverages the collective insights from various trees, enhancing predictive accuracy. This approach, rooted in diversity, makes Random Forest a powerful tool for classification tasks [106].

### **6.2.4.3 k-Nearest Neighbors (KNN)**

k-Nearest Neighbors (KNN) is a straightforward yet effective supervised machine learning algorithm that emerged in the field of pattern recognition. The KNN algorithm was introduced by Evelyn Fix and Joseph Hodges in 1951 as a non-parametric method for pattern recognition. It gained further prominence and formalization in the 1960s and 1970s [107]. Operating on the principle of proximity in feature space, KNN classifies new data points based on the majority vote of their k-nearest neighbors. While intuitive and easy to understand, KNN may face challenges in high-dimensional spaces due to the "curse of dimensionality," where the effectiveness of the algorithm decreases as the number of features increases. Despite this limitation, its strength lies in its adaptability to various datasets and simplicity in implementation, making it a valuable tool in many machine-learning applications [108].

### **6.2.4.4 Extreme Gradient Boosting (XGBoost)**

Extreme Gradient Boosting (XGBoost) emerged in 2011, introduced by Carlos Guestrin and Tianqi Chen. Continuously optimized for modern data science challenges, XGBoost is a boosting tree-based learning framework renowned for its scalability, parallelizability, and superior performance. By combining multiple models, XGBoost creates a robust ensemble that frequently outperforms competing algorithms [109]. The regularization techniques employed in XGBoost effectively control overfitting, contributing to its high level of performance [110].

#### 6.2.4.5 Gradient Boosting (GBoost)

Gradient Boosting, a powerful ensemble learning technique, was introduced by Jerome Friedman in 1999. The specific algorithm, known as the Gradient Boosting Machine (GBM), was later developed by Friedman, Trevor Hastie, and Robert Tibshirani. GBM constructs a predictive model through an ensemble of weak learners, typically decision trees. The algorithm sequentially adds trees to correct errors made by the existing ensemble, resulting in a strong predictive model. Gradient Boosting excels in capturing complex relationships within the data, making it suitable for regression and classification tasks. Its ability to handle diverse datasets and improve predictive accuracy over iterations has contributed to its widespread use in various machine-learning applications [111].

#### 6.2.5 Optimization strategy

To get the best possible performance for the models, hyperparameters were utilized. Proper hyperparameter tuning is often a critical step in achieving optimal model performance. To get the suitable hyperparameter for each algorithm the GridSearchCV method was employed. The method was given a set of different values for each hyperparameter, it then generates the best combination of these values by using the training set. Table 90 illustrates the algorithms and their best hyperparameter with the original data.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	10
	Min_samples_leaf	1
	Min_samples_split	5
	N_estimators	100
SVM	Kernel	linear
	C	1
	gamma	scale
K-NN	N_neighbors	7
	algorithm	auto
	weights	uniform
Gboost	Learning_rate	0.01
	Max_depth	3
	N_estimators	200
XGboost	Gamma	0.1
	Learning_rate	0.01
	Max_depth	3
	N_estimators	100

Table 90 the optimal hyperparameter for each classifier in the original data

Table 91 represents the algorithms with their best hyperparameters in oversampled data and using all the features.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	none
	Min_samples_leaf	1

	Min_samples_split	5
	N_estimator	150
SVM	Kernel	rbf
	C	10
	gamma	scale
	N_neighbors	5
K-NN	algorithm	auto
	weights	distance
	Learning_rate	0.1
Gboost	Max_depth	5
	N_estimator	150
	Gamma	0
XGboost	Learning_rate	0.1
	Max_depth	3
	N_estimator	200

Table 91 the optimal hyperparameter for each classifier in over-sampled data with all the features selected

## 6.2.6 Results and discussion

Within this section, we present the outcomes derived from the developed models after the implementation of GridSearchCV on both the original dataset and the sampled data using SMOTETomek, a method used to balance the class distribution within the dataset by oversampling the minority class and under sampling the majority class. Following the acquisition of optimal hyperparameters and model training through stratified 10-fold cross-validation. Utilizing all the 20 features. The ensuing section delineates a comprehensive examination of the results obtained, as presented in Table 92.

Classifier	Dataset	Testing Accuracy	Precision	Recall	F1-Score
Random Forest	Original	70.83%	71.00%	71.00%	70.00%
	Using SMOTETomek	91.11%	92%	91%	91%
Gradient Boosting	Original	73.61%	67.00%	69.00%	68.00%
	Using SMOTETomek	87.77%	88.00%	88.00%	88.00%
SVM	Original	73.61%	76.00%	74.00%	74.00%
	Using SMOTETomek	87.77%	88.00%	88.00%	87.00%
XGBoost	Original	68.05%	74.00%	68.00%	67.00%
	Using SMOTETomek	86.66%	87.00%	87.00%	86.00%
K-NN	Original	69.44%	67.00%	67.00%	68.00%
	Using SMOTETomek	71.11%	71.00%	71.00%	68.00%

Table 92 The results of the proposed models before and after sampling was applied

The original results highlighted the need to address the class imbalance, as demonstrated by the enhanced performance of all models following SMOTETomek's implementation. However, substantial improvements were observed in all algorithms after using SMOTETomek. Testing accuracies soared, demonstrating a notable improvement in predictive power. Moreover, the use of hyperparameter tuning and feature selection contributed to more robust models. Random Forest emerged as the algorithm with the highest testing accuracy, achieving 91.11%, precision, recall and F1 values are 92%, 91%, and 91% respectively. Gradient Boosting and SVM both followed with a testing accuracy of 87.77%. Gradient Boosting with precision, recall and F1 values are all 88%. SVM with precision, recall and F1 values are 88%, 88%, and 87% respectively. These models have shown significant improvements in predictive accuracy, demonstrating the usefulness of SMOTETomek in managing class imbalance and boosting the models' capacity to predict osteoporosis and osteopenia, which may facilitate early detection and intervention. Figure 112 shows the number of samples of each class before and after applying SMOTETomek.

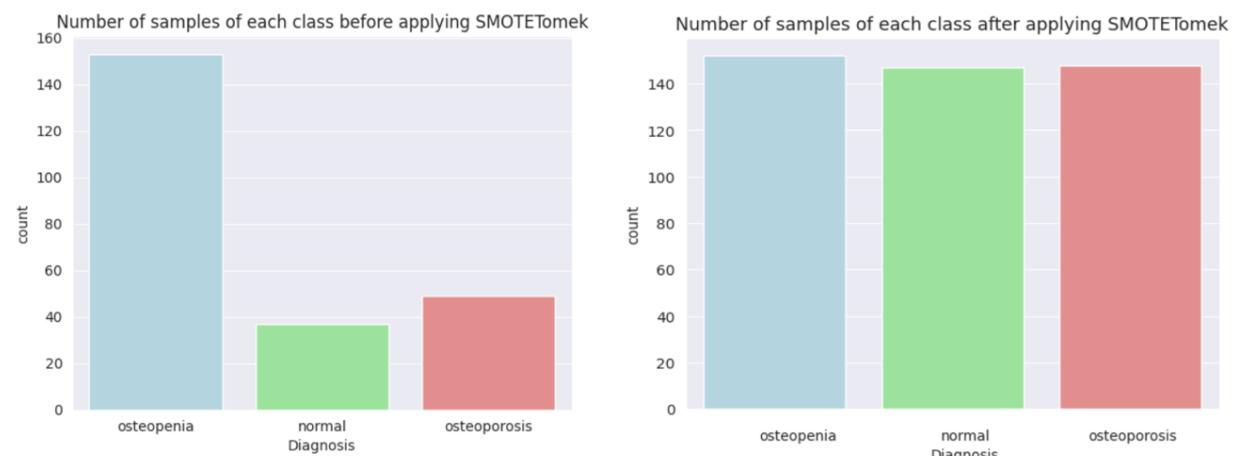


Figure 112 Number of samples of each class before and after applying SMOTETomek

### 6.2.6.1 Results with Feature Selection

Sequential Forward Feature Selection (SFFS) represents a strategic methodology employed to optimize the performance of machine learning models by iteratively selecting informative subsets of features. In this study, the SFFS technique was applied to a dataset comprised of 20 features extracted from clinical data gathered during the BMD camp conducted by the Unani and Panchkarma Hospital in Srinagar, J&K, India [69]. The objective was to ascertain the impact of varying feature subsets on model accuracy across different machine-learning algorithms. Sequential Forward Feature Selection (SFFS) is an iterative technique used in feature selection, initiates with an empty set, and systematically enriches subsets of features by adding one feature at a time, guided by their impact on boosting the model's predictive accuracy [112]. Using a selected metric, SFFS assesses various feature combinations at each iteration, retaining the combination that exhibits the greatest improvement. This process keeps going until it reaches a predetermined end point, like reaching a certain number of features or reaching the point where performance improvement is no longer possible. By choosing a subset that maximizes prediction accuracy, SFFS seeks to discover and incorporate the most valuable features, improving the model's predictive power and reducing processing requirements. Table 93 presents several feature subsets identified through SFFS. These subsets are examined to

gauge their influence on model performance when paired with optimal hyperparameters and the SMOTETomek technique.

Features	RF	SVM	KNN	GB	XGBoost	Features selected
<b>8</b>	87.77 %	86.66%	85.55 %	88.88 %	87.77%	Joint Pain, Gender, Age, Diabetic, Seizer Disorder, Dialysis, Site, age_group
<b>10</b>	87.77 %	88.88%	85.55 %	87.77 %	87.77%	Joint Pain, Gender, Age, Alcoholic, Diabetic, Seizer Disorder, Estrogen Use, Dialysis, Site, age_group
<b>15</b>	87.77 %	88.88%	81.11 %	85.55 %	88.88%	Joint Pain, Gender, Age, Weight, Smoker, Alcoholic, Diabetic, Hypothyroidism, Seizer Disorder, Estrogen Use, Dialysis, Daily Eating habits, Site, Obesity,age_group
<b>20</b>	91.11 %	87.77%	71.11 %	87.77 %	86.66%	Joint Pain, Gender, Age, Height,Weight, SmokeAlcoholic, Diabetic, Hypothyroidism, Seizer Disorder, Estrogen Use, History of Fracture, Maximum Walking distance, Dialysis, Family History of Osteoporosis, Daily Eating habits, BMI, Site, Obesity,age_group

Table 93 Testing Accuracy for Different Feature Subsets

The results showed varying performances of all the five machine learning algorithms, with different feature subsets. Random Forest achieve its highest accuracy using all 20 features with a testing accuracy of 91.11%, SVM, Gradient Boosting, and XGBoost maintains stable accuracy across different feature subsets. Additionally, K-NN shows a decrease in accuracy as the number of features increases.

#### 6.2.7.2 Further Discussion of the Results

RandomForest's confusion matrix showcases a robust performance across all classes as shown in Table (a), with 30 instances correctly predicted as osteopenia, 24 instances as normal, and 28 instances as osteoporosis, demonstrating fewer misclassifications compared to other models. Following that, SVM in Table (c) demonstrates a strong performance with 30 instances correctly predicted as osteopenia, 22 instances as normal, and 27 instances as osteoporosis, displaying fewer misclassifications overall, particularly between class normal and class osteoporosis. Table (b) with GradientBoosting exhibits a reasonably balanced performance, with 29 instances correctly predicted as osteopenia, 24 instances as normal, and 26 instances as osteoporosis. XGBoost in Table (e) displays a similar prediction to GradientBoosting, correctly predicting 29 instances as osteopenia, 23 instances as normal, and 26 instances as osteoporosis. However, KNN demonstrates comparatively more misclassifications across all classes as shown in Table (d), correctly predicting 26 instances as osteopenia, 11 instances as normal, and 27 instances as osteoporosis, highlighting its relatively weaker performance among the models considered.

RandomForest		Predicted			GradientBoosting		Predicted		
		osteopenia	normal	osteoporo			osteopenia	normal	osteoporosis
Actual	osteopenia	30 (TP)	0 (FN)	0 (FN)	Actual	osteopenia	29 (TP)	1 (FN)	0 (FN)
	normal	3 (FP)	24 (TN)	4 (FN)		normal	4 (FP)	24 (TN)	3 (FN)
	osteoporosis	0 (FP)	1 (FP)	28 (TN)		osteoporosis	0 (FP)	3 (FP)	26 (TN)

(a)

(b)

SVM		Predicted			KNN		Predicted		
		osteopenia	normal	osteoporo			osteopenia	normal	osteoporosis
Actual	osteopenia	30 (TP)	0 (FN)	0 (FN)	Actual	osteopenia	26 (TP)	4 (FN)	0 (FN)
	normal	4 (FP)	22 (TN)	5 (FN)		normal	26 (TP)	4 (FN)	0 (FN)

	osteoporosis	0 (FP)	2 (FP)	27 (TN)		normal	7 (FP)	11 (TN)	13 (FN)
						osteoporosis	1 (FP)	1 (FP)	27 (TN)

(c)

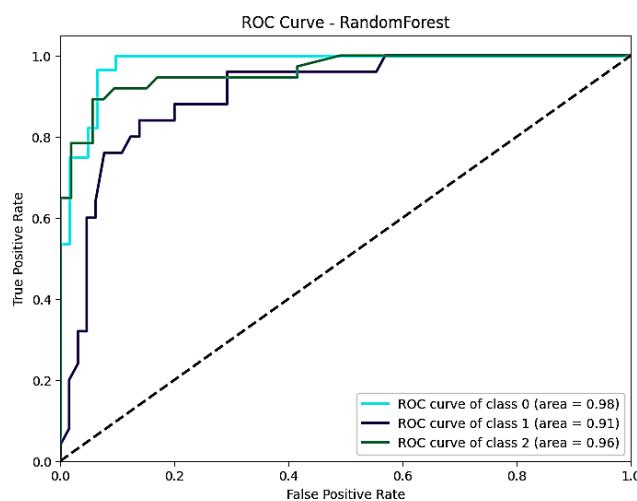
(d)

XGBoost		Predicted		
		osteopenia	normal	osteoporosis
Actual	osteopenia	29 (TP)	1 (FN)	0 (FN)
	normal	3 (FP)	25 (TN)	5 (FN)
	osteoporosis	0 (FP)	3 (FP)	26 (TN)

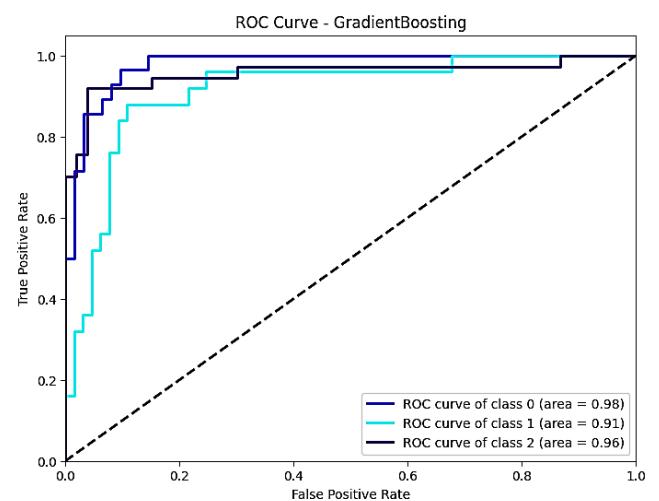
(e)

Table 94 Confusion Matrix for (a) RF, (b) GBoost, (c) SVM, (d) KNN, and (e) XGBoost

Figure 112 shows the Receiver Operating Characteristics (ROC) curve analysis for RF, GradientBoosting, SVM, KNN and XGBoost, which is a valuable tool for evaluating the performance of classification models across different classes. It illustrates the trade-off between a true positive rate (sensitivity) and a false positive rate (1-specificity) across different thresholds. RandomForest and GradientBoosting both indicate strong discrimination ability across all classes, with Class 0, Class 1, and Class 2 showcasing AUC values of 0.98, 0.91, and 0.96 respectively. These results underline the effective discrimination capabilities of both GradientBoosting and RandomForest models across different classes, showcasing their aptitude in classification tasks. SVM showcases strong discrimination for Class 0 (0.97) and Class 2 (0.94) but slightly lower for Class 1 (0.86). KNN demonstrates moderate discrimination, with Class 0 (0.94) having the highest AUC, followed by Class 2 (0.89) and Class 1 (0.71). Moreover, XGBoost displays high discrimination ability across all classes (Class 0: 0.97, Class 1: 0.90, Class 2: 0.95). These results outline the varying discrimination capacities of each model across different classes in classification tasks.



(a)



(b)

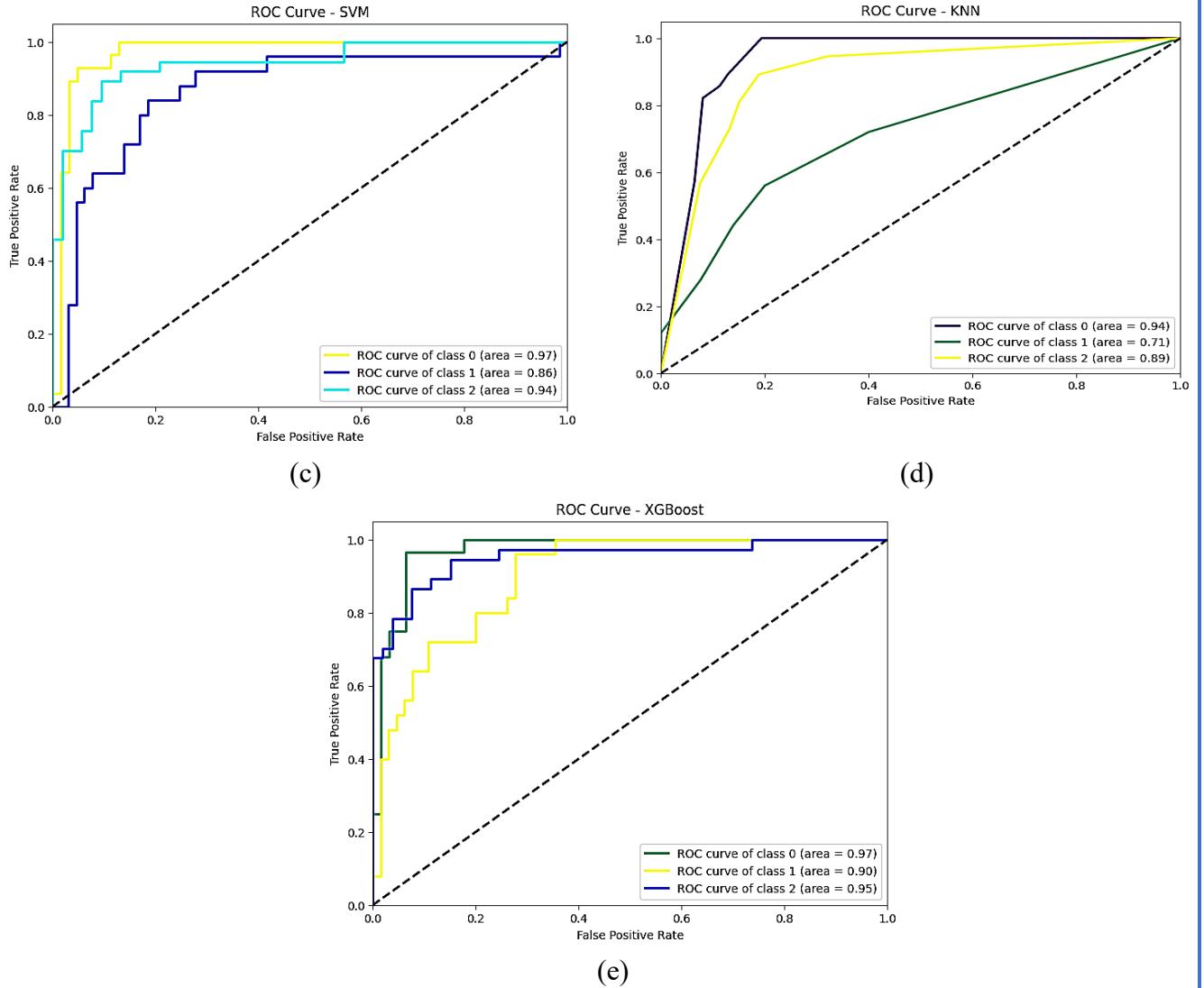


Figure 113 (a) RF, (b) GradientBoosting, (c) SVM, (d) KNN, and (e) XGBoost ROC Curve

## 6.3 Empirical Study on Epileptic Seizures Disease Dataset

### 6.3.1 Data Description

The dataset comprised individuals sourced from a continuous stream of referrals to both the Functional Neurological Disorders Clinic and the University of Colorado (CU) Epilepsy Monitoring Unit (EMU) during the period spanning January 1, 2017, to May 15, 2019 [68]. These patients were mostly identified by a confirmed video-electroencephalography (vEEG) diagnosis of epilepsy or a history of seizures (DS) which is referred to dissociative seizures or non-epileptic seizures, with vEEG evaluations performed both inside and outside the medical system. Patients were randomly chosen to form two groups, comprising individuals with dissociative seizures (DS) and epileptic seizures (ES). Additionally, patients with both DS and ES diagnoses were included due to their clinical importance and limited data availability.

Numerous demographic, clinical, and diagnostic variables are included in the dataset. These include gender, seizure characteristics, clinical and medical history, records of traumas or traumatic occurrences, and other relevant information. Table 95 lists all the features and their

corresponding types. For research purposes, this dataset offers a thorough overview of the health status of the patients, enabling in-depth examination and analysis.

Feature	Type
RRID	int64
Diagnosis	object
Sex	object
#non psych comorbidities	int64
# Prior AEDs	int64
Asthma	int64
Migraine	int64
Chronic Pain	int64
Diabetes	int64
non metastatic cancer	int64
Number of non-seizure non psych medication	int64
# Current AEDs	int64
Date Onset	object
Date Admit	object
Baseline (sz freq)	float64
Median duration of seizures	float64
# of seizure types	Int64
Injury with seizure	object
Catamenial	object
Trigger of sleep deprivation	object
Aura	object
Ictal Eye Closure	object
Ictal hallucinations	object
Oral automatisms	object
Incontinence	object
Limb automatisms	Object
Ictal tonic-clonic	object
Muscle twitching	Object
Hip thrusting	Object
Post-ictal fatigue	Object
Any head injury	Object
Psych traumatic events	Object
Concussion w/o LOC	Object
Concussion w/LOC	Object
Severe TBI (LOC>30min)	Object
Opioids? Yes/no	Int64
Sex abuse	Object
Physical Abuse	Object
Rape	Object

Table 95 Features' description

### 6.3.2 Statistical Analysis of the Dataset

This section illustrates the characteristics of each attribute in the dataset. This step is essential to determine the most suitable pre-processing techniques for optimal modeling methods. Both categorical and numerical attributes will be explored as both hold a significant value for forthcoming modeling stages.

Various statistical matrices were used for numerical attributes. These metrics offer deep insights into the inherent characteristics of the data. Table 96 represents a detailed statistic for numerical features.

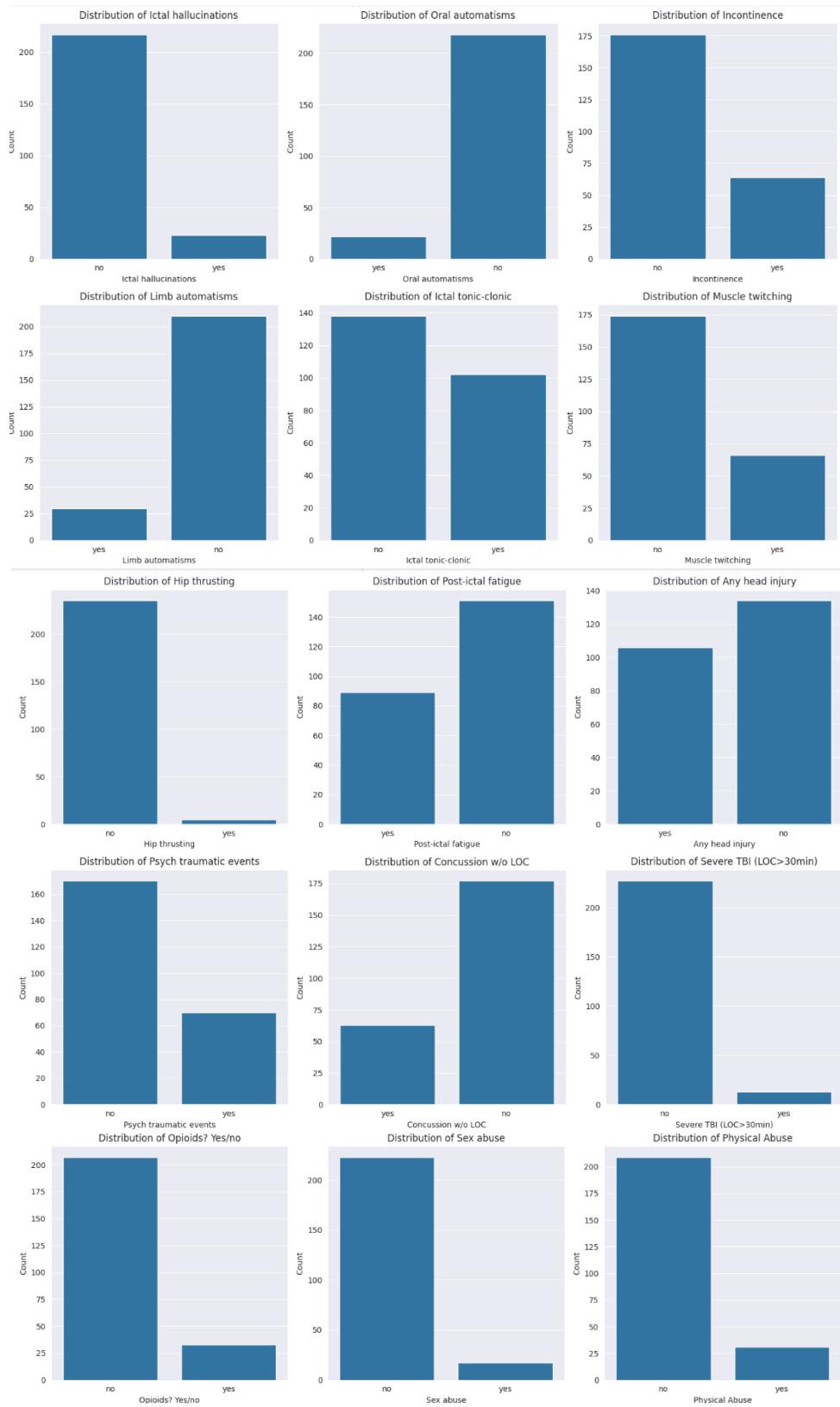
Feature	Mean	Standard deviation	Min	25 <sup>th</sup> quartile	50 <sup>th</sup> quartile	75 <sup>th</sup> quartile	Max	Missing value
Number of non-psych comorbidities	4.58	4.98	0.00	1.00	3.00	6.00	31.00	0
Number of Prior AEDs	1.71	2.32	0.00	0.00	1.00	3.00	14.00	0
Number of non-seizure non psych medication	5.41	6.20	0.00	1.00	3.00	8.00	44.00	0
Number of Current AEDs	1.77	1.34	0.00	1.00	2.00	3.00	6.00	0
Baseline (sz freq)	24.95	41.16	0.00	3.00	10.00	30.00	308.00	5
Median duration of seizures	5.17	10.50	0.05	1.00	2.00	5.00	65.00	12
Number of seizure types	2.19	1.35	1.00	1.00	2.00	3.00	10.00	0

Table 96 Statistical analysis of numerical features

Figure 114 shows the distribution of various categorical features of patients. The distribution of the class named diagnosis which is the target class shows that 134 patients have non-epileptic seizures, 75 have epileptic seizures, and 31 have been diagnosed with both. The dataset contains 160 female patients and 80 male patients. 45 patients have asthma while 195 are not infected with it. A total of 73 patients suffered from migraine pain and 167 did not. 37 patients experienced chronic pain and 203 did not. The patients who have diabetes are 14 and 226 have not been diagnosed with it. Patients who have non metastatic cancer are 19 while 221 do not have this type of cancer. The seizures cause injury for 55 patients, on the other hand, 182 did not face such a problem. A majority of 230 patients haven't experience catamenial epilepsy only 9 patients have experience it. 213 have no sleep deprivation while 25 have trigger sleep deprivation. The patients who feel the aura before the seizure start are 143 but 97 have not feel the aura. 195 patients haven't close their eyes during the seizure, 45 have closed their eyes. 23 patients who experienced seizures did suffer from hallucinations during the seizure while 217 have not experienced it. Through the seizure 22 have oral automatisms and 218 have not. A total of 64 patients has Incontinence and 176 have not. 30 have their limb automatisms while 210 have normal limb. During the seizure 102 experience tonic-colonic phase and 138

did not. 66 patients have their muscle twitching and 174 have not. The majority of the patients did not have their hip thrusting with total of 235 while 5 did have. After the seizure 89 felt fatigue while 151 did not feel fatigue. 106 have an injury on their head while 134 have not. 170 patients have not experienced any Psychological traumatic event while 70 have experienced. A total of 48 patients has experience concussion with loss of consciousness and 192 have not. 13 patients have lost their consciousness for more than 30 minutes while 227 have not experienced such a condition. 33 patients have used Opioids while 207 have not. 17 and 31 patients have experienced sex or physical abuse respectfully while 223 and 209 have not. 7 patients have been raped and 209 did not. the patients who have experienced concussion without losing their consciousness are 63 while 177 have loss it.





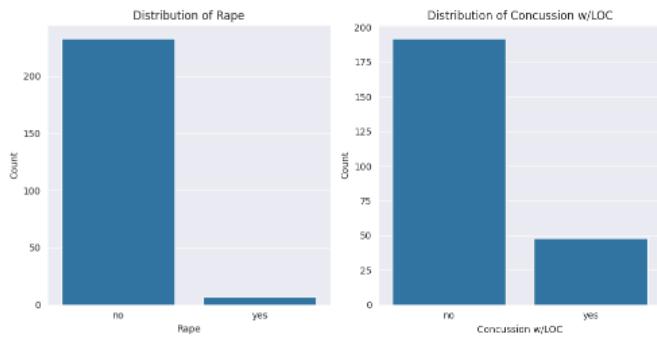


Figure 114 distribution of categorical features

### 6.3.3 Experimental Setup

This study developed a pre-emptive model for diagnosing Epilepsy by detecting epileptic seizures using Python programming language. The materials and methods utilized in this study involved several key steps. The dataset was first imported using the Pandas package, and then its original contents and structure were analyzed. Preprocessing steps included handling duplicate entries and assessing missing values. Visualization techniques such as heatmaps were employed to visualize the distribution of missing values. Furthermore, specific features were explored using histograms and box plots to identify outliers and understand their distribution. For features exhibiting outliers, the median was selected for imputation to mitigate the influence of outliers. Categorical features with missing values were imputed using the most frequent strategy. Additionally, certain categorical features underwent data cleaning to standardize values. Following preprocessing, label encoding was applied to convert categorical features into numerical representations. The dataset was split into 70% training and 30% testing sets, and feature scaling was performed using standardization. Eight machine learning methods are used: K-Nearest Neighbors, Gradient Boosting, eXtreme Gradient Boosting, Random Forest, Support Vector Machine, Logistic Regression, Decision Tree, and AdaBoost, were implemented using GridSearchCV for hyperparameter tuning. The model performance was evaluated using accuracy scores, confusion matrices, and classification reports to assess predictive performance on both training and testing datasets. Once the top-performing model was chosen, eXplainable Artificial Intelligence (XAI) methods such as Feature Importance and LIME were employed to investigate the model's decision-making mechanism in greater depth. Figure 115 shows this complete procedure.

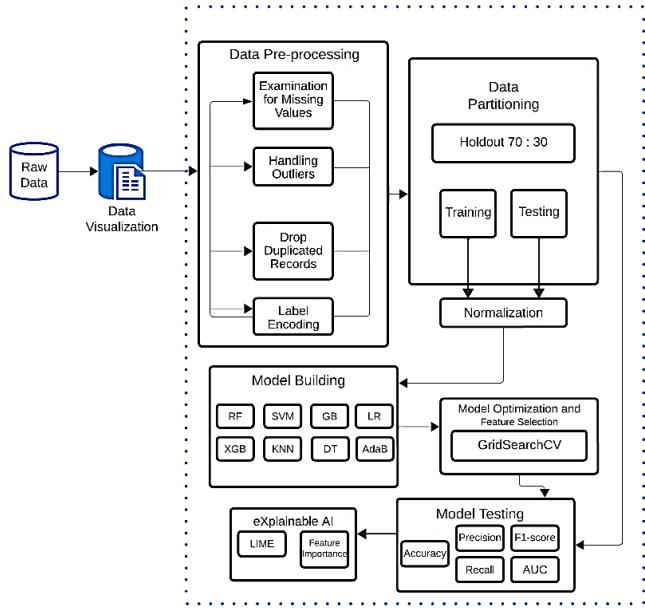


Figure 115 The proposed framework for the pre-emptive diagnosis of Epileptic Seizures

### 6.3.4 Description of the Proposed Techniques

#### 6.3.4.1 K-Nearest Neighbors (KNN):

KNN is a simple, instance-based learning algorithm used for classification and regression tasks. It classifies data points based on the majority class of their k-nearest neighbors in the feature space. It's non-parametric and effective for small datasets [113].

#### 6.3.4.2 Gradient Boosting:

Gradient Boosting is an ensemble learning technique that builds decision trees sequentially, each correcting the errors of its predecessor. It minimizes a loss function by greedily adding weak learners. It's widely used in both regression and classification problems due to its high predictive accuracy [114].

#### 6.3.4.3 eXtreme Gradient Boosting (XGBoost):

XGBoost is an optimized version of gradient boosting, known for its scalability and speed. It employs a more regularized model formalization to control overfitting and provides better performance. It's highly efficient for large datasets and has become a popular choice in various machine learning competitions [92].

#### 6.3.4.4 Random Forest:

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of individual trees. It offers robustness to overfitting and high accuracy, making it suitable for a wide range of tasks [115].

#### 6.3.4.5 Support Vector Machine (SVM)

SVM is a supervised learning algorithm that analyzes data and recognizes patterns, used for classification and regression analysis. It works by finding the hyperplane that best separates

different classes in the feature space. SVM is effective in high-dimensional spaces and in cases where the number of dimensions exceeds the number of samples [100].

#### **6.3.4.6 Logistic Regression:**

Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. It's primarily used for binary classification problems and estimates the probability that a given input belongs to a particular class. Despite its name, it's a classification algorithm rather than a regression algorithm [97].

#### **6.3.4.7 Decision Tree:**

Decision Tree is a supervised learning algorithm used for classification and regression tasks. It partitions the data into subsets based on the value of the features and makes predictions by following the tree from the root to the leaf nodes. It's intuitive, easy to understand, and capable of handling both numerical and categorical data [116].

#### **6.3.4.8 AdaBoost:**

AdaBoost, short for Adaptive Boosting, is an ensemble learning method that combines multiple weak learners to create a strong classifier. It assigns higher weights to misclassified data points, allowing subsequent weak learners to focus more on difficult cases. It's particularly effective in improving the performance of classifiers in situations where simple models perform poorly [99].

#### **6.3.5 Optimization strategy**

Selecting the best hyperparameter for each algorithm has a significant impact on the algorithm's performance. Proper hyperparameter leads to the best possible performance and accuracy. To achieve this goal GridSearchCV method was utilized, this method works by giving a set of various values. The method will try and combine all the values until it finds the best combination. Table 97 below shows the best hyperparameter for each algorithm with all the features.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	None
	Min_samples_leaf	1
	Min_samples_split	2
	N_estimators	100
SVM	C	1
	Gamma	Scale
	kernel	Rbf
K-NN	algorithm	Auto
	N_neighbors	7
	weights	Distance
Gboost	Learning_rate	0.01
	Max_depth	5
	N_estimators	100

<b>XGboost</b>	Gamma	0
	Learning_rate	0.1
	Max_depth	7
	N_estimators	100
<b>Logistic Regression</b>	C	0.1
	Solver	Liblinear
<b>Decision Tree</b>	Max_depth	5
	Min_samplesleaf	2
	Min_samples_split	2
<b>AdaBoost</b>	Learning rate	0.01
	N_estimators	150

Table 97 the best hyperparameter selected on the original data

### 6.3.6 Results and discussion

Within this section, we present the outcomes derived from our developed models after the implementation of GridSearchCV on both the original dataset and the sampled data using SMOTETomek, which is a method used to balance the class distribution within the dataset by oversampling the minority class and under sampling the majority class.[117] Following the acquisition of optimal hyperparameters and model training through stratified 10-fold cross-validation and utilizing all the 36 features, the following section delineates a comprehensive examination of the results obtained, as presented in Table 98.

Classifier	Dataset	Testing Accuracy	Precision	Recall	F1-Score
<b>Logistic Regression</b>	Original	87.67%	88.00%	88.00%	88.00%
	Using SMOTETomek	79.45%	85.00%	79.00%	81.00%
<b>Random Forest</b>	Original	80.82%	75.00%	81.00%	78.00%
	Using SMOTETomek	78.08%	79.00%	78.00%	79.00%
<b>SVM</b>	Original	86.30%	87.00%	86.00%	85.00%
	Using SMOTETomek	78.08%	78.00%	78.00%	78.00%

<b>KNN</b>	Original	80.82%	82.00%	81.00%	80.00%
	Using SMOTETomek	60.27%	70.00%	60.00%	62.00%
<b>Gradient Boosting</b>	Original	82.19%	82.00%	82.00%	82.00%
	Using SMOTETomek	72.60%	77.00%	73.00%	75.00%
<b>XGBoost</b>	Original	79.45%	80.00%	79.00%	80.00%
	Using SMOTETomek	78.08%	83.00%	78.00%	80.00%
<b>Decision Tree</b>	Original	76.71%	78.00%	77.00%	77.00%
	Using SMOTETomek	57.53%	69.00%	58.00%	61.00%
<b>AdaBoost</b>	Original	76.71%	76.00%	77.00%	76.00%
	Using SMOTETomek	78.08%	83.00%	78.00%	80.00%

Table 98 The results of the proposed models before and after sampling was applied.

The results show that, although the SMOTETomek oversampling strategy was applied in an attempt to improve class imbalance, classifier performance was not always improved by this method. SMOTETomek reduced the testing accuracy of several classifiers (Random Forest, SVM, KNN, Logistic Regression, Gradient Boosting, XGBoost, and Decision Tree) when compared to the original dataset. This finding indicates that, for this dataset, the original, non-oversampled results are considered more reliable. With the best testing accuracy of 87.67%, a precision of 88.00%, recall of 88.00%, and F1-score of 88.00%. Among the original results, Logistic Regression was the best option for modeling this dataset without using the SMOTETomek technique. Following that, SVM obtained 86.30% accuracy, 87.00% precision, 86.00% recall, and 85.00% F1-score. This analysis emphasizes how important it is to carefully assess the effects of the SMOTETomek technique and choose the best classifier when working with unbalanced datasets. Figure 116 shows the number of samples of each class of the target feature before and after applying SMOTETomek. (Class 0 represents NES, class 1 represents NES, class 2 represents NES and ES.)

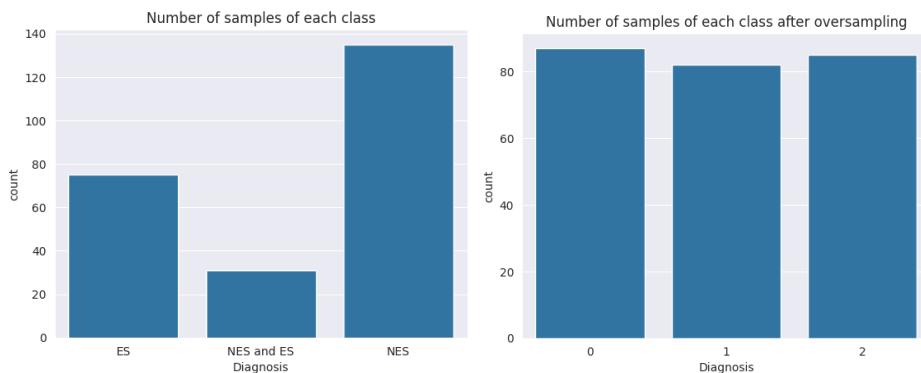


Figure 116 Number of samples of each class of the target feature before and after applying SMOTETomek.

### 6.3.6.1 Results with Feature Selection

The feature selection utilized is based on the statistical test chi-squared to check if two categorical variables have significant relationship [118]. In this research, a dataset of 36 attributes that were taken from an ongoing stream of referrals to the FND Clinic and CU EMU was tested with the chi-squared test [68]. Comparing observed and expected frequencies inside a contingency table is the basis of the test. By examining the correlation between the components, the chi-square test addresses feature selection issues. The purpose of this analysis is to ascertain whether the correlation between two sample categorical variables accurately reflects the correlation between them in the population [118]. The following is the chi-squared equation 23, where O is the observed value and E is the expected value.

$$\text{Equation (23): } X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

SelectKBest is the methodology used in this study to select the best K number of features from the chi-squared test result. The chi-squared calculates the relationship between the target variable ‘Diagnosis’ and each feature in the dataset. The result is then sorted, so the SelectKBest takes the highest K features in the selection process. Table 99 shows the result of the ML algorithms with feature selection.

Feature Subsets	10	20	30	All features (36)
LoR	80.82%	86.30%	86.30%	87.67%
RF	80.82%	83.56%	83.56%	80.82%
SVM	78.08%	84.93%	86.30%	86.30%
KNN	80.82%	79.45%	76.71%	80.82%
GB	84.93%	80.82%	82.19%	76.71%
XGBoost	80.82%	83.56%	82.19%	79.45%
DT	80.82%	71.23%	75.34%	71.23%
AdaBoost	76.71%	76.71%	76.71%	76.71%

Table 99 result with feature selection

### 6.3.6.2 Further Discussion of the Results

The confusion matrices offer valuable insights into the performance of various machine learning models in classifying instances into three distinct categories: ES (epileptic seizure), NES (non-epileptic seizures), and a combined category of ES & NES. Beginning with RandomForest Table (a), it demonstrates a relatively balanced performance, accurately classifying 19 instances of ES, 40 instances of NES, and 0 instance in the ES & NES category. However, there are instances where it struggles to distinguish between ES and ES & NES, as evidenced by the misclassifications present. Moving to SVM Table (b), it shows a strong performance, correctly classifying 18 instances of ES, 43 instances of NES, and 2 instances in the ES & NES category. This model exhibits fewer misclassifications compared to RandomForest, particularly in the ES category, indicating its robustness in accurately identifying epileptic seizure instances. KNN Table (c) displays a reasonably balanced performance, accurately classifying 19 instances of ES, 38 instances of NES, and 2 instances

in the ES & NES category. However, it shows higher misclassification rates in distinguishing between ES and NES, suggesting potential challenges in generalizing across these categories. GradientBoosting Table (d) showcases competitive performance, accurately classifying 16 instances of ES, 42 instances of NES, and 2 instances in the ES & NES category. Similar to RandomForest and SVM, it faces challenges in distinguishing between ES and NES, leading to some misclassifications.

Similarly, XGBoost Table (e) provides results akin to GradientBoosting, accurately classifying 17 instances of ES, 42 instances of NES, and 2 instances in the ES & NES category. Yet, like GradientBoosting, it struggles to differentiate between ES and NES, resulting in misclassifications. LogisticRegression Table (f) exhibits a balanced performance, accurately classifying 19 instances of ES, 42 instances of NES, and 3 instances in the ES & NES category. However, it shows slightly higher misclassification rates compared to SVM and RandomForest, particularly in distinguishing between NES and ES & NES. Moving on to DecisionTree Table (g), it demonstrates varying performance across classes, correctly classifying 20 instances of ES, 35 instances of NES, and 1 instance in the ES & NES category. It shows higher misclassification rates, especially in the NES category, indicating potential limitations in accurately identifying non-epileptic seizure instances. Finally, AdaBoost Table (h) provides relatively balanced results, accurately classifying 15 instances of ES, 40 instances of NES, and 3 instances in the ES & NES category. However, it faces challenges in distinguishing between ES and ES & NES, leading to some misclassifications.

RandomForest		Predicted		
		ES	NES	ES And NES
Actual	ES	19 (TP)	4 (FN)	0 (FN)
	NES	5 (FP)	40 (TN)	0 (FN)
	ES And NES	2 (FP)	3 (FP)	0 (TN)

(a)

SVM		Predicted		
		ES	NES	ES And NES
Actual	ES	18 (TP)	5 (FN)	0 (FN)
	NES	2 (FP)	43 (TN)	0 (FN)
	ES And NES	1 (FP)	2 (FP)	2 (TN)

(b)

KNN		Predicted		
		ES	NES	ES And NES
Actual	ES	19 (TP)	4 (FN)	0 (FN)
	NES	7 (FP)	38 (TN)	0 (FN)
	ES And NES	0 (FP)	3 (FP)	2 (TN)

(c)

Gradient Boosting		Predicted		
		ES	NES	ES And NES
Actual	ES	16 (TP)	4 (FN)	3 (FN)
	NES	3 (FP)	42 (TN)	0 (FN)
	ES And NES	2 (FP)	1 (FP)	2 (TN)

(d)

XGBoost		Predicted		
		ES	NES	ES And NES
Actual	ES	17 (TP)	4 (FN)	2 (FN)
	NES	5 (FP)	39 (TN)	1 (FN)
	ES And NES	2 (FP)	1 (FP)	2 (TN)

(e)

Logistic Regression		Predicted		
		ES	NES	ES And NES
Actual	ES	19 (TP)	2 (FN)	2 (FN)
	NES	3 (FP)	42 (TN)	0 (FN)
	ES And NES	1 (FP)	1 (FP)	3 (TN)

(f)

DecisionTree		Predicted		
		ES	NES	ES And NES
Actual	ES	20 (TP)	3 (FN)	0 (FN)

AdaBoost		Predicted		
		ES	NES	ES And NES
Actual	ES	15 (TP)	6 (FN)	2 (FN)

	<b>NES</b>	<b>7 (FP)</b>	<b>35 (TN)</b>	<b>3 (FN)</b>		<b>NES</b>	<b>4 (FP)</b>	<b>40 (TN)</b>	<b>1 (FN)</b>
<b>ES And NES</b>	<b>3 (FP)</b>	<b>1 (FP)</b>	<b>1 (TN)</b>		<b>ES And NES</b>	<b>2 (FP)</b>	<b>2 (FP)</b>	<b>1 (TN)</b>	

(g)

(h)

Table 100 Confusion Matrix of (a) RandomForest, (b) SVM, (c) KNN, (d) Gboost, (e) XGBoost, (f) LoR, (g) DT, (h) AdaBoost,

The displayed ROC curves in figure 117 represent the performance of various machine learning models applied in a study for the pre-emptive diagnosis of epileptic seizures using clinical data. Each curve plots the true positive rate against the false positive rate for three classes, labeled 0, 1, and 2, at various discrimination thresholds. A model with a curve closer to the top-left corner indicates a higher true positive rate and a lower false positive rate, showcasing better diagnostic accuracy. The area under the curve (AUC) serves as a summary metric, with a value of 1 representing perfect classification and 0.5 representing a random guess.

In these ROC curves, we see that the Support Vector Machine (SVM) model exhibits the most robust discrimination ability among the machine learning models evaluated. The SVM demonstrates high Area Under the Curve (AUC) values for all classes, 0.95 for Class 0, 0.96 for Class 1, and 0.96 for Class 2, implying a strong predictive performance with high true positive rates and low false positive rates. Logistic Regression follows closely, presenting AUCs of 0.95 for Class 0, 0.96 for Class 1, and 0.95 for Class 2, which suggests it is also highly effective at distinguishing between the classes. Random Forest ranks next with AUC values of 0.95, 0.93, and 0.93 for Classes 0, 1, and 2, respectively, showing its considerable predictive capabilities. The XGBoost model, while slightly lower, still shows impressive AUC measures of 0.91, 0.92, and 0.90 for the three classes, indicating a strong classification performance. Other models like Gradient Boosting and AdaBoost also show good performance but with slightly lower AUC values, while the Decision Tree and KNN models show moderate to lower AUCs, indicating more variability in their class distinction capability. These insights facilitate a clear comparison of model efficacy, with SVM and Logistic Regression leading in this context, thereby providing a valuable guide for researchers in selecting optimal models for epileptic seizure prediction.

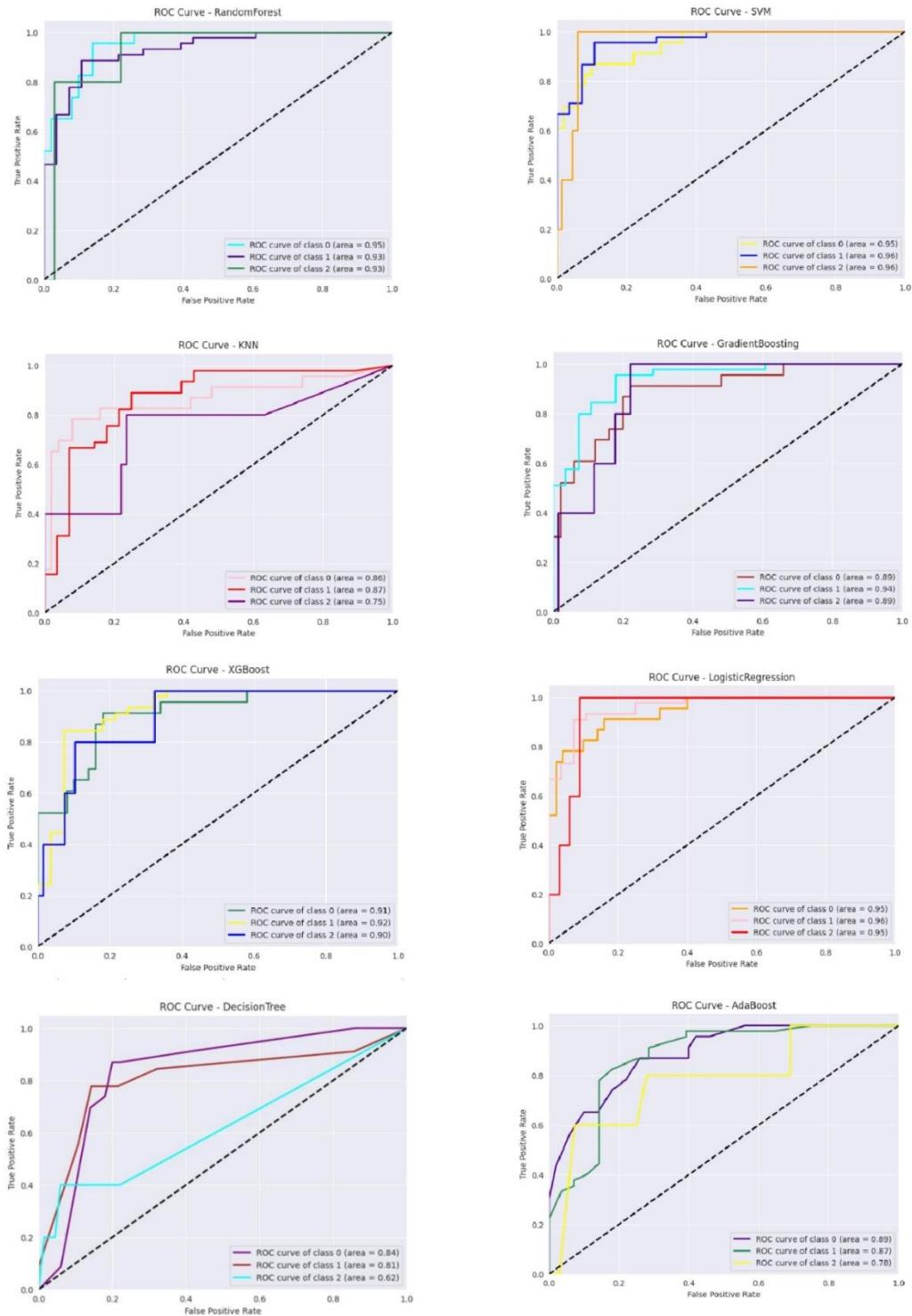


Figure 117 ROC curves of the eight models applied.

## 6.4 Empirical Study on Sickle Cell Anemia Disease Dataset

### 6.4.1 Data Description

The dataset for this study was derived from a descriptive cross-sectional analysis conducted at Al Fashir Teaching Hospital in Sudan, spanning from December 2017 to August 2018 [70]. This study included 400 pediatric patients who were admitted to the hospital during the specified period. These patients, aged between 0 and 18 years, were approached to participate in the study, with their parents providing informed consent. The selection process employed random sampling, ensuring equal opportunity for any child admitted to the pediatric ward to be included. To gather sociodemographic information, including age, gender, and tribal affiliation, a structured questionnaire was used. 400 parents gave their consent, indicating a moderate percentage of refusal. Blood samples, amounting to 5 ml each, were collected via peripheral venipuncture under strictly aseptic conditions. The samples were promptly transferred to the laboratory within 5 minutes of collection for further analysis. Upon arrival at the laboratory, the blood samples underwent immediate testing using advanced diagnostic equipment. Complete blood counts were performed using an automated haematological analyzer, Sysmex Kx 21N, while hemoglobin electrophoresis was conducted utilizing the MINICAP HEMOGLOBIN capillary zone electrophoresis (CZE) system by Sebia, France.

For the haemoglobin electrophoresis procedure, blood samples were initially cooled to 2–8 °C to facilitate sediment formation over several hours. Following centrifugation at 5000 rpm for 5 minutes, the plasma was removed, and the sediment was vortexed briefly. Reagent cups were prepared for analysis, ensuring proper calibration of haemoglobin buffers and waste disposal. Each sample, along with a normal Hb A2 control, was labeled with specific barcodes and placed into haemolysing tubes, which were then positioned within the carousel for automated analysis. Table 101 shows each feature with its corresponding type.

Feature	Type
Hb Electro	object
Platelets	Integer
Total White Blood Cells	float64
Mean Cell Hemoglobin Concentration	float64
Mean Cell Hemoglobin	float64
Mean Cell Volume	float64
Red Blood Cells	float64
Packed Cell Volume	float64
Hemoglobin	float64
Tribe	Integer
Age/ Years	object
Sex	object
No	Integer

Table 101 Features' description

#### 6.4.2 Statistical Analysis of the Dataset

Statistical analysis is an essential process to explore and understand the characteristics of both categorical and numerical data. In addition, determine any further needed preprocessing technique is needed. This will ensure that the model will perform smoothly with the highest possible accuracy.

Numerical attributes were explored through various statistical matrices to ensure the understanding of these attribute and its characteristics. Table 102 shows the statistical matrices for the numerical attribute.

Feature	Mean	Standard deviation	Min	25 <sup>th</sup> quartile	50 <sup>th</sup> quartile	75 <sup>th</sup> quartile	Max	Missing value counts
PLTs	284.59	134.52	21.00	195.25	272.00	351.00	780.00	0
TWBCs	7.95	6.27	1.00	4.60	6.65	9.40	61.70	0
MCHC	32.67	1.80	25.80	31.60	32.80	33.80	40.20	0
MCH	27.06	2.99	15.90	24.52	27.30	28.90	39.50	0
MCV	82.57	8.61	34.00	77.90	83.05	88.00	107.70	0
PCV	35.48	6.77	11.20	31.82	36.20	39.87	53.00	0
RBCs	4.47	1.53	1.14	4.07	4.49	4.83	22.10	0
Hb	11.61	2.23	3.50	10.50	11.90	13.00	17.70	0
Tribe	11.06	13.09	1.00	2.00	4.00	15.75	59.00	0

Table 102 Statistical analysis for numerical attribute

The figure below shows the distribution of the only categories attribute which is the gender. The dataset contains 226 female patient and 176 male patients.

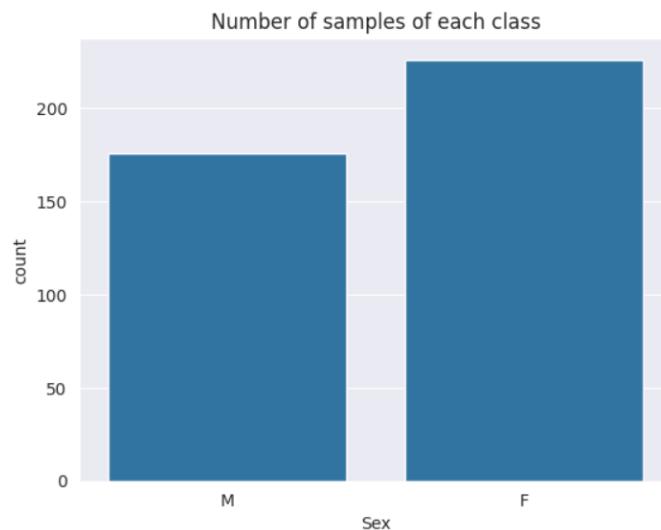


Figure 118 The distribution of gender attribute

#### 6.4.3 Experimental Setup

This study aimed to develop a preemptive model for diagnosing sickle cell anemia utilizing the Python programming language. The materials and methods employed in this research encompassed a series of procedural steps designed to effectively handle the dataset and

implement machine learning algorithms. Initially, the dataset was imported into the analysis environment using the Pandas package. Preprocessing steps were then executed to ensure data quality and consistency. Duplicate entries were identified and removed, while missing values were handled through a combination of visualization techniques and imputation strategies. Visualization methods, including heatmaps, histograms, and count plots, were employed to assess the distribution of missing values, wrong values, and identify outliers. Categorical features with missing values were imputed using the most frequent strategy, and data cleaning procedures were conducted to standardize categorical values. Subsequently, categorical features were encoded into numerical representations using label encoding, facilitating compatibility with machine learning algorithms. The dataset was then partitioned into training (70%) and testing (30%) sets to facilitate model evaluation. Feature scaling was performed using standardization to ensure uniformity in feature magnitudes across the dataset. Ten machine learning algorithms, including RandomForest, SVM, KNN, GradientBoosting, XGBoost, LogisticRegression, DecisionTree, AdaBoost, GaussianNB, and SVM\_Gaussian, were implemented for predictive modeling. Hyperparameter tuning was conducted using GridSearchCV to optimize algorithm performance. Evaluation of model performance involved assessing accuracy scores, confusion matrices, and classification reports on both training and testing datasets. Once the top-performing model was identified, eXplainable Artificial Intelligence (XAI) techniques, such as Feature Importance and LIME, were employed. The complete procedural workflow is illustrated in Figure 119.

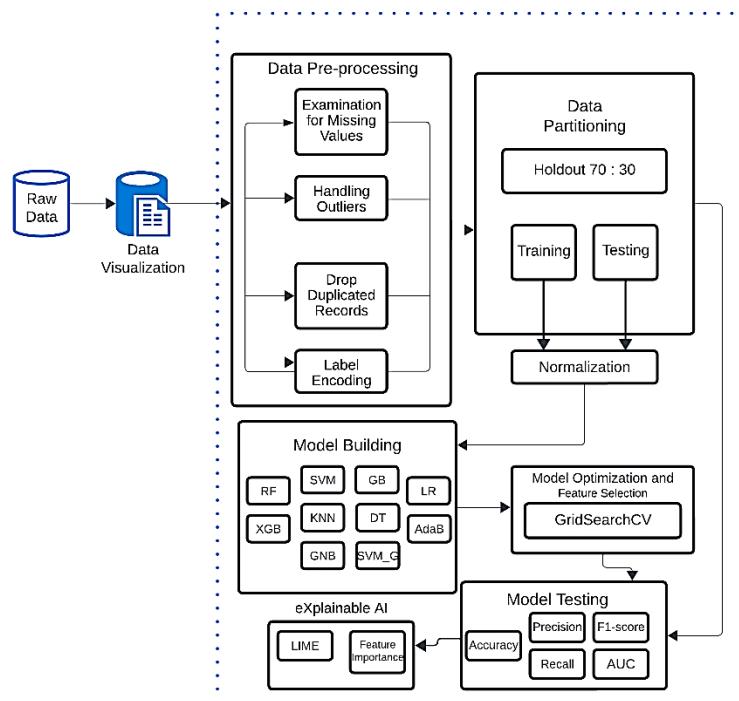


Figure 119 The proposed framework for the pre-emptive diagnosis of SCA

## **6.4.4 Description of the Proposed Techniques**

### **6.4.4.1 Support Vector Machine (SVM):**

SVM is a powerful classifier that works by finding the hyperplane that best divides a dataset into classes. It is effective in high-dimensional spaces and versatile as different Kernel functions can be specified for the decision function. Robust against overfitting, especially in high-dimensional space, SVM is widely used for pattern recognition [100].

### **6.4.4.2 AdaBoost (Adaptive Boosting):**

AdaBoost is an ensemble technique that combines multiple weak classifiers to form a strong classifier. By reweighting the training samples, it focuses on the hard-to-classify instances in subsequent models, enhancing classification accuracy. AdaBoost is particularly effective for binary classification tasks and is less susceptible to overfitting compared to other algorithms [119].

### **6.4.4.3 K-Nearest Neighbors (KNN):**

KNN is a simple, instance-based learning algorithm where the class of a sample is determined by the majority of the classes of its nearest neighbors. It is highly adaptable and easy to implement, making it suitable for solving both classification and regression problems. However, it becomes significantly slower as the size of the data increases [120].

### **6.4.4.4 XGBoost (Extreme Gradient Boosting):**

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. It is renowned for its performance and speed in classification tasks. XGBoost provides a scalable and efficient solution, often winning many Kaggle competitions [92].

### **6.4.4.5 Logistic Regression:**

Logistic Regression is used for binary classification problems, where predictions are mapped to probabilities via the logistic function. It is robust to noise and efficient for linearly separable classes. This model is often used as a baseline for binary classification problems [97].

### **6.4.4.6 Gradient Boosting:**

Gradient Boosting constructs an additive model in a forward stage-wise fashion, allowing optimization of arbitrary differentiable loss functions. Known for its effectiveness in handling heterogeneous features and variable interactions, it is widely used in both classification and regression tasks [114].

### **6.4.4.7 Decision Tree:**

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision Trees are easy to interpret and visualize [98].

### **6.4.4.8 Random Forest:**

Random Forest is an ensemble of Decision Trees, typically trained via the bagging method. The individual trees operate as weak learners, while their collective decisions, taken by

majority voting, provide robust predictions against overfitting. Random Forest performs well on large datasets with high dimensionality [71].

#### **6.4.4.9 Gaussian Naive Bayes (GaussianNB):**

GaussianNB applies Bayes' theorem with the assumption of independence among predictors. GaussianNB is particularly suited when features are continuous and normally distributed. It is simple and effective, especially for large datasets [121].

#### **6.4.4.10 SVM with Gaussian Kernel (SVM\_Gaussian):**

SVM with a Gaussian kernel allows the model to create more complex boundaries around classes. The kernel transforms the data into a higher-dimensional space where a hyperplane can effectively separate classes that are non-linearly separable in the original space. This variant is useful for datasets with complex patterns [122].

#### **6.4.5 Optimization strategy**

Hyperparameters are the parameters that need to be set before the training starts. The impact of selecting the optimal parameter is significant on the model's performance. GridSearchCV () was employed for the aim of finding the best hyperparameter for each model. Table 103 illustrate the hyperparameter that have been selecting by GridSearch CV () method with the original dataset.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	None
	Min_samples_leaf	1
	Min_samples_split	2
	N_estimator	100
SVM	C	10
	Gamma	Scale
	kernel	Rbf
K-NN	algorithm	Auto
	N_neighbors	5
	weights	Distance
Gboost	Learning_rate	0.1
	Max_depth	5
	N_estimator	100
XGboost	Gamma	0.1
	Learning_rate	0.01
	Max_depth	3
	N_estimator	150
Logistic Regression	C	10
	penalty	12
Decision Tree	Max_depth	5
	Min_samplesleaf	1

	Min_samples_split	2
GaussianNB	-	-
AdaBoost	Learning rate	0.1
	N estimators	100
SVM_Gaussian	C	10
	gamma	scale

Table 103 Best hyperparameter with original dataset

The table below shows the best hyperparameter with original dataset and features selection.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	None
	Min_samples_leaf	2
	Min_samples_split	5
	N_estimator	200
SVM	C	10
	Gamma	Scale
	kernel	Rbf
K-NN	algorithm	Auto
	N_neighbors	5
	weights	Distance
Gboost	Learning_rate	0.01
	Max_depth	3
	N_estimator	150
XGboost	Gamma	0
	Learning_rate	0.01
	Max_depth	5
	N_estimator	150
Logistic Regression	C	0.1
	Penalty	12
Decision Tree	Max_depth	5
	Min_samplesleaf	2
	Min_samples_split	2
GaussianNB	-	-
AdaBoost	Learning rate	0.1
	N estimators	100
SVM_Gaussian	C	10
	gamma	scale

Table 104 Best hyperparameters with original dataset and features selection

Table 105 represent the selected hyperparameters with oversampled dataset.

algorithm	Hyperparameter	Best Hyperparameter
-----------	----------------	---------------------

<b>RF</b>	<code>Max_depth</code>	None
	<code>Min_samples_leaf</code>	1
	<code>Min_samples_split</code>	2
	<code>N_estimators</code>	200
<b>SVM</b>	<code>C</code>	10
	<code>Gamma</code>	auto
	<code>kernel</code>	Rbf
<b>K-NN</b>	<code>algorithm</code>	Auto
	<code>N_neighbors</code>	3
	<code>weights</code>	Distance
<b>Gboost</b>	<code>Learning_rate</code>	0.1
	<code>Max_depth</code>	5
	<code>N_estimators</code>	100
<b>XGboost</b>	<code>Gamma</code>	0
	<code>Learning_rate</code>	0.01
	<code>Max_depth</code>	7
	<code>N_estimators</code>	200
<b>Logistic Regression</b>	<code>C</code>	10
	<code>Penalty</code>	12
<b>Decision Tree</b>	<code>Max_depth</code>	10
	<code>Min_samplesleaf</code>	1
	<code>Min_samples_split</code>	2
<b>GaussianNB</b>	-	-
<b>AdaBoost</b>	<code>Learning rate</code>	0.5
	<code>N estimators</code>	150
<b>SVM_Gaussian</b>	<code>C</code>	10
	<code>gamma</code>	auto

Table 105 5 Best hyperparameters with oversampled dataset

#### 6.4.6 Results and Discussion

In this section, we report the results of our developed models after the application of GridSearchCV on the sampled data obtained from SMOTETomek, a technique that balances the distribution of classes in the dataset by undersampling the majority class and oversampling the minority class [117]. Following the process of identifying the optimal hyperparameters, training the model with all ten features through stratified 10-fold cross-validation, and displaying the results in Table 105, the next section offers a detailed analysis of the findings.

Classifier	Dataset	Testing	Precision	Recall	F1-Score
		Accuracy			
<b>Random Forest</b>	Original	85.95%	86.00%	86.00%	83.00%
	Using SMOTETomek	75.15%	82.00%	75.00%	78.00%
<b>SVM</b>	Original	83.47%	80.00%	83.00%	79.00%
	Using SMOTETomek	61.49%	81.00%	61.00%	67.00%

	Original	85.95%	86.00%	86.00%	83.00%
<b>KNN</b>	Using SMOTETomek	62.11%	82.00%	62.00%	68.00%
	Original	84.29%	83.00%	84.00%	80.00%
<b>Gradient Boosting</b>	Using SMOTETomek	73.91%	83.00%	74.00%	77.00%
	Original	84.29%	87.00%	84.00%	78.00%
<b>XGBoost</b>	Using SMOTETomek	60.24%	84.00%	60.00%	66.00%
	Original	80.99%	73.00%	81.00%	75.00%
<b>Logistic Regression</b>	Using SMOTETomek	49.68%	75.00%	50.00%	57.00%
	Original	84.29%	82.00%	84.00%	80.00%
<b>Decision Tree</b>	Using SMOTETomek	70.80%	81.00%	71.00%	75.00%
	Original	81.81%	76.00%	82.00%	77.00%
<b>AdaBoost</b>	Using SMOTETomek	60.86%	83.00%	61.00%	67.00%
	Original	80.99%	78.00%	81.00%	79.00%
<b>GaussianNB</b>	Using SMOTETomek	67.70%	81.00%	68.00%	72.00%
	Original	83.47%	80.00%	83.00%	79.00%
<b>SVM_Gaussian</b>	Using SMOTETomek	61.49%	81.00%	61.00%	67.00%

Table 106 The results of the proposed models before and after sampling was applied.

The findings indicate that while the SMOTETomek oversampling strategy was implemented to reduce class imbalance, classifier performance was not consistently enhanced by this technique. All classifiers' testing accuracy was decreased by SMOTETomek in comparison to the original dataset. This result suggests that the original results will be considered as more reliable for this dataset. Random Forest proved to be the most effective method for modeling this dataset without the need for the SMOTETomek methodology, with the best testing accuracy of 85.95%. This investigation highlights how crucial it is to evaluate the SMOTETomek technique's impacts carefully and select the optimal classifier when dealing with unbalanced datasets. Thus, we conclude that using the original dataset without oversampling produced better results for our sickle cell anemia prediction task, with RandomForest appearing as the most successful classifier with an accuracy of 85.95%, a precision of 86.00%, recall of 86.00%, and F1-score of 83.00%. The number of samples for each class of the target feature, both before and after using SMOTETomek, is displayed in Figure 119. (Class 0 stands for N, and class 1 for SS.)

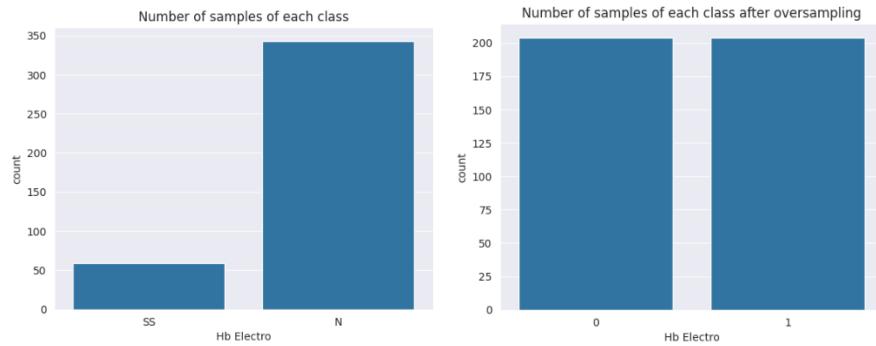


Figure 120 Number of samples of each class of the target feature before and after applying SMOTETomek.

#### 6.4.6.1 Results with Feature Selection

The feature selection utilized is the tree based embedded method to select the features with the highest importance to the chosen classifier [123]. In this research, a dataset of 400 patients at the pediatric ward were taken from Al Fashir Teaching Hospital, Sudan [70]. In this methodology the feature selection is embedded in the classifier learning process [123]. The ‘RandomForestClassifier’ is utilized in ‘SelectFromModel’ method for computing the feature importance scores during the training process. For each classifier the feature selection method chooses a subset of all the features and uses ‘RandomForestClassifier’ to calculate the feature importance score, then the feature subset with the highest score is chosen for the final modeling. Figure 121 demonstrate the process of feature selection.

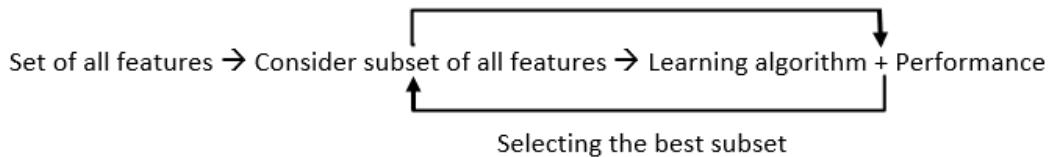


Figure 121 Feature selection process

The RF classifier with 5 features achieved the highest performance with an accuracy rate of 86.77%, a recall of 89%, a precision of 87% and 83% F1-score. Table 107 shows the result of the ML algorithms with feature selection with their best feature subsets, compared to the results of the classifiers with all features.

ML classifiers	All features (10)	With feature selection	Number of features	Feature subset
<b>RF</b>	85.95%	86.77%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
<b>SVM</b>	83.47%	82.64%	6	['PLTs', 'MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']

<b>KNN</b>	<b>85.95%</b>	82.64%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
<b>GB</b>	84.29%	83.47%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
<b>XGBoost</b>	84.29%	85.12%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
<b>LoR</b>	80.99%	82.64%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
<b>DT</b>	84.29%	84.29%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
<b>AdaBoost</b>	81.81%	82.64%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
<b>GaussianNB</b>	80.99%	81.81%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
<b>SVMGaussian</b>	83.47%	82.64%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']

Table 107 Results with feature selection

#### 6.4.6.2 Further Discussion of the Results

The confusion matrices provided offer insightful observations into the performance of various machine learning models in classifying instances into two categories: SS (sickle cell trait and sickle cell disease) and N (no sickle cell). Each model shows unique strengths and weaknesses in handling this binary classification problem.

Starting with RandomForest Table 108(a), we see a perfect classification for the SS category (100 correctly classified), but the model struggles slightly with the N category, misclassifying 5 cases as SS out of 121(). This suggests a high sensitivity but potential overfitting to the SS category, which might lead to false positives in practical scenarios.

Transitioning to SVM Table 108(b), the model demonstrates robust performance with a high accuracy rate for both categories, accurately classifying 96 SS and 17 N cases. However, like RandomForest, it shows a tendency to misclassify 4 instances of N as SS. This indicates that SVM is generally effective but might also be slightly biased towards predicting SS.

KNN Table 108(c) exhibits a slightly lower performance compared to SVM, with more misclassifications: 93 SS correctly identified but 7 misclassified, and 14 N correctly identified with 7 misclassified. This could imply that KNN may require tuning of parameters such as the number of neighbors to improve its precision, especially in distinguishing SS from N effectively.

GradientBoosting Table 108(d) shows competitive accuracy, particularly in minimizing false negatives for the SS category with 99 correct predictions and 1 misclassification. However, it still misclassifies 2 out of 21 N instances as SS, similar to the previous models, suggesting a common challenge among these algorithms in avoiding type I errors in this dataset.

Similarly, XGBoost Table 108(e) aligns closely with GradientBoosting in performance, achieving high accuracy in the SS category with 100 correct predictions and no SS misclassifications but encountering some difficulty with 3 out of 21 N category misclassifications. Both models display a strong capability to identify SS correctly but need to improve in specificity to reduce the false positive rate.

LogisticRegression Table 108(f) provides a balanced approach with a strong classification record for both categories, accurately identifying 99 SS and 20 N cases but misclassifying 1 SS and 1 N. It presents a slightly higher rate of misclassifications compared to SVM and RandomForest, but its overall performance is solid, indicating good generalizability across different data distributions.

DecisionTree Table 108(g) shows variability in its performance, with a near-perfect score in classifying SS (99 correct, 1 misclassified) but a higher misclassification rate for N, correctly identifying 18 and misclassifying 3. This could reflect DecisionTree's sensitivity to the training dataset's nuances, potentially leading to overfitting and less robust predictions.

AdaBoost Table 108(h) offers balanced results, with decent accuracy in both categories: 98 SS correctly identified with 2 misclassified, and 19 N correctly identified with 2 misclassified. AdaBoost's performance suggests that while it is capable of handling diverse data, like many ensemble methods, it might also benefit from further tuning to enhance its discriminative ability.

GaussianNB Table 108(i) shows moderate sensitivity with 94 SS cases correctly identified and 6 misclassified. For the N category, it again misclassifies 5 out of 21 cases as SS, reinforcing the common challenge of managing false positives among these models.

Finally, in Table 108(j) SVM with a Gaussian kernel mirrors the performance of the standard SVM with 96 SS correctly classified and 4 misclassified. In the N category, it also misclassifies 4 cases as SS, indicating that the kernel choice does not significantly affect performance in this dataset but still highlights issues with potential overfitting to SS.

Overall, each model presents unique strengths and areas for improvement in classifying sickle cell anemia disease, highlighting the importance of model selection and parameter optimization in predictive healthcare analytics.

RandomForest		Predicted	
		SS	N
Actual	SS	100	0
	N	16	5

(a)

SVM		Predicted	
		SS	N
Actual	SS	96	4
	N	17	4

(b)

KNN		Predicted	
		SS	N
Actual	SS	93	7
	N	14	7

GradientBoosting		Predicted	
		SS	N
Actual	SS	99	1
	N	19	2

(c)

XGBoost		Predicted	
		SS	N
Actual	SS	100	0
	N	18	3

(d)

LogisticRegression		Predicted	
		SS	N
Actual	SS	99	1
	N	20	1

(e)

DecisionTree		Predicted	
		SS	N
Actual	SS	99	1
	N	18	3

(f)

AdaBoost		Predicted	
		SS	N
Actual	SS	98	2
	N	19	2

(g)

GaussianNB		Predicted	
		SS	N
Actual	SS	94	6
	N	16	5

(h)

SVM_Gaussian		Predicted	
		SS	N
Actual	SS	96	4
	N	17	4

(i)

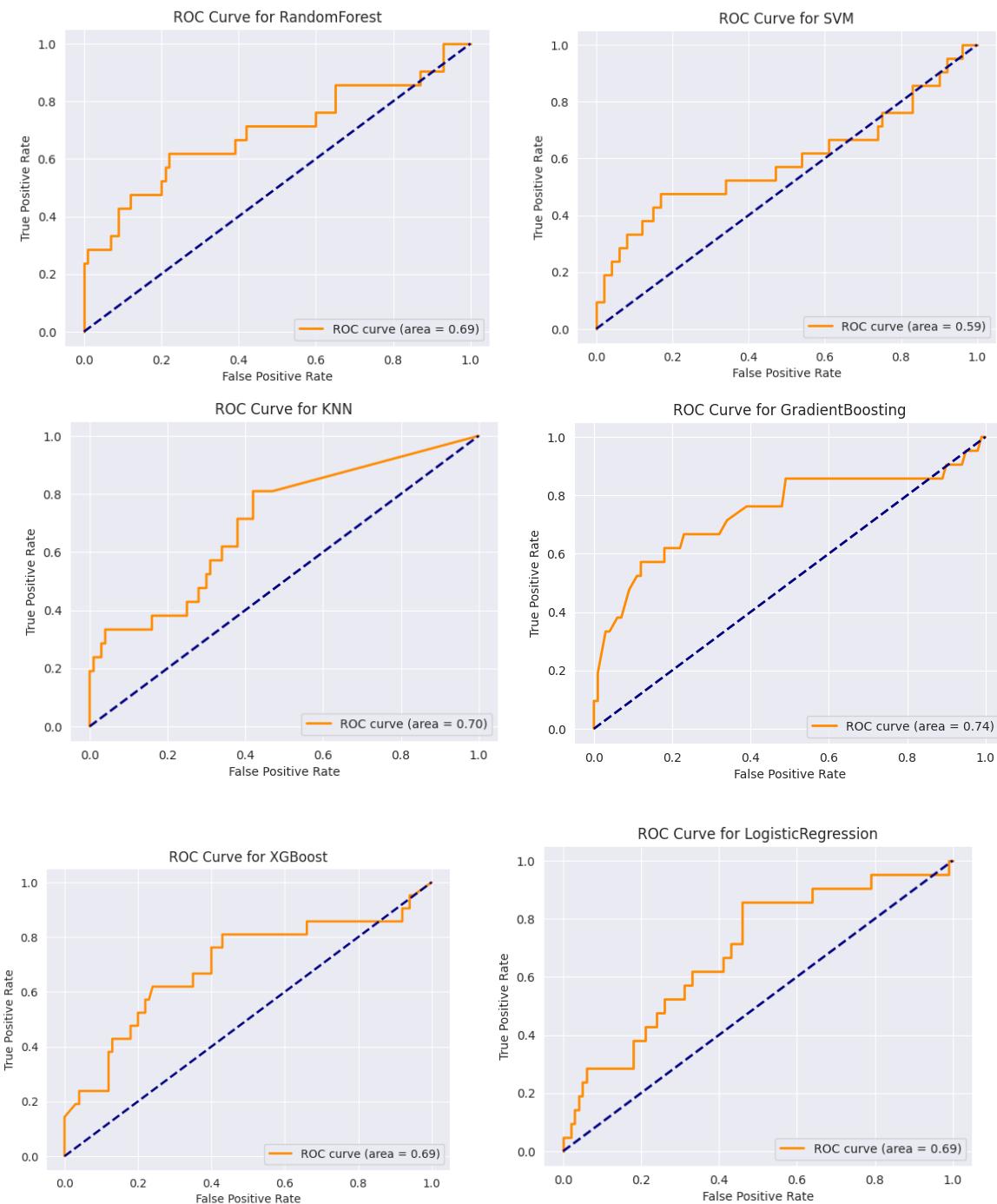
Table 108 Confusion Matrix of (a) RandomForest, (b) SVM, (c) KNN, (d) Gboost, (e) XGBoost, (f) LoR, (g) DT, (h) AdaBoost, (i) GaussianNB and (j) SVM\_Gaussian

The ROC curves displayed across these figures measure the performance of various machine learning models utilized for predicting specific clinical outcomes. Each curve illustrates the true positive rate against the false positive rate for different thresholds, where a higher Area Under the Curve (AUC) indicates more effective model performance. The AUC serves as a summary metric, with a value of 1.0 depicting perfect prediction and 0.5 a random guess.

In this analysis, the Gradient Boosting model showcases the highest effectiveness with an AUC of 0.74, indicating a strong capability to distinguish between outcomes accurately. Following closely is the AdaBoost model, with an AUC of 0.72, demonstrating robust predictive power. The K-Nearest Neighbors (KNN) algorithm also performs well, achieving an AUC of 0.70, suggesting its usefulness in this specific predictive task. Both the Random Forest and XGBoost models report an AUC of 0.69, showing good, albeit not optimal, predictive accuracies.

On the other hand, Logistic Regression and the Decision Tree models display moderate effectiveness with AUCs of 0.69 and 0.67, respectively, indicating reasonable capabilities but

with potential limitations in more complex scenarios. The Support Vector Machine (SVM) with a Gaussian kernel presents an AUC of 0.65, and the Gaussian Naive Bayes model has an AUC of 0.63, both reflecting fair performance. The standard SVM model trails with an AUC of 0.59, suggesting it might be less suited for this particular prediction task compared to the others. These insights underline the importance of model selection in clinical data analysis, where Gradient Boosting and AdaBoost emerge as particularly effective for high-stakes predictions. This comparative analysis provides a valuable benchmark, guiding researchers towards selecting models that not only maximize the true positive rate but also effectively minimize false positives in clinical predictions.



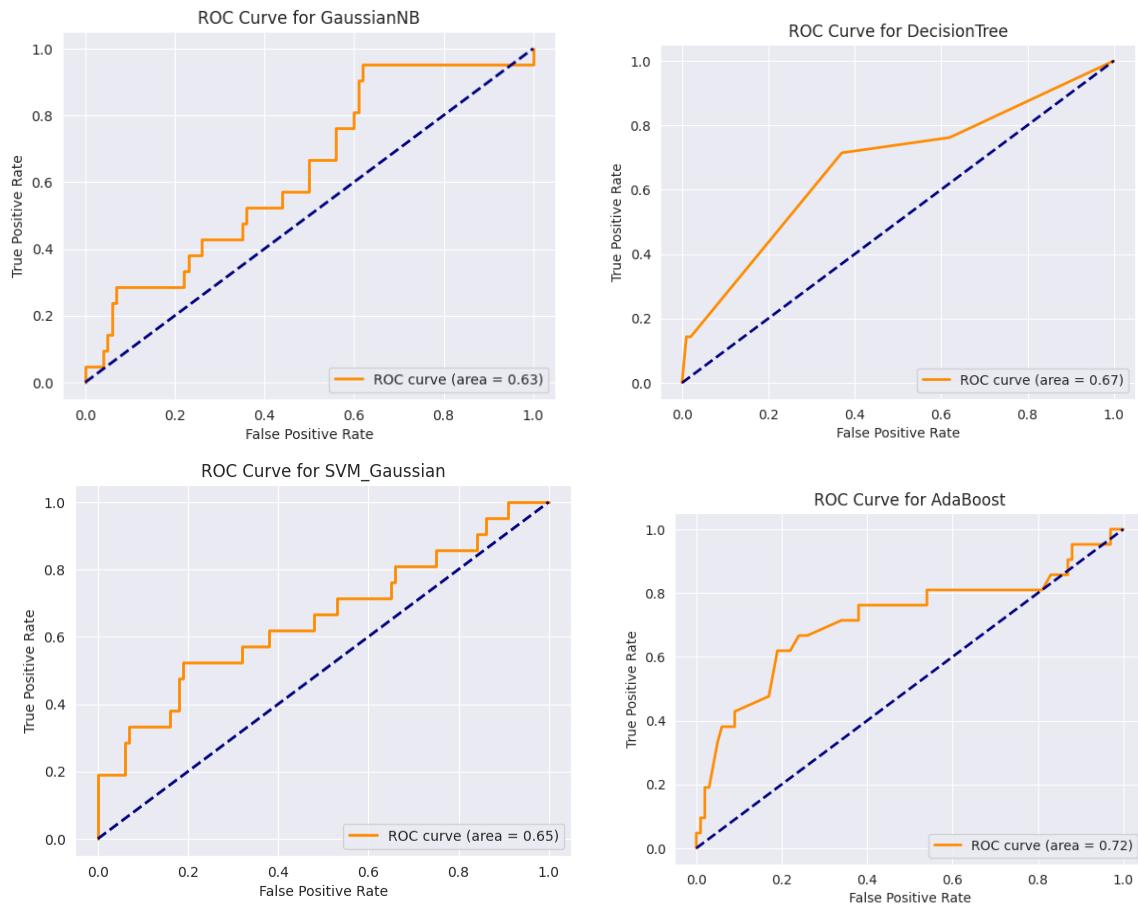


Figure 122 ROC curves of the ten models applied.

## 6.5 Website Implementation

The website development process was divided into two parts. The client-side was developed using HTML, CSS, JavaScript, and JQuery, whereas the server-side was developed using Python’s web framework “Flask”. A templating language for Python, Jinja, was employed to establish data communication between the server and client. The following sections show a demonstration of the website functionalities execution.

### 6.5.1 Login

The ‘Login’ interface allows users to login by entering their username and password. The system then checks the validity of a user’s entry by comparing it to the database data. Suppose the entry matches a record saved in the database. In that case, the user will be able to access the system with specific privileges based on the user type (admin, medical specialist, registered user). Otherwise, a user will be prevented from accessing the system. Sessions were used to deal with the tasks related to logging in, logging out, and remembering the users’ sessions for long periods. To ensure users’ safety, the SHA256 hashing algorithm was utilized to store the hash values of users’ passwords. Figure 123 below displays the ‘Login’ interface.

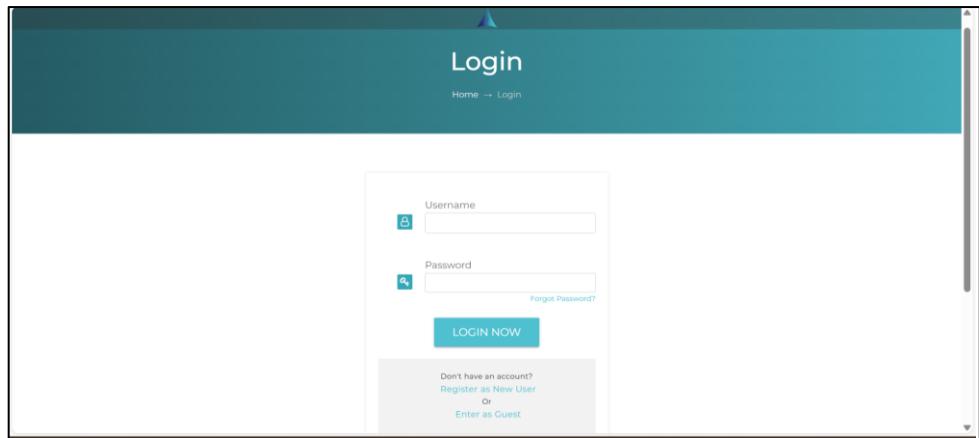


Figure 123 Login' Interface

### 6.5.2 Forgot Password Page

In case of a forgotten password, the user can click on the “Forgot Password” link to be redirected to the ‘Forgot Password’ interface, as shown in Figure 124. The email entered existence is validated using a MongoDB query, in which a new random password is generated if the email exists using the ‘random’ library in Python. The generated password includes 5 random alphabets and 3 random digits stored as hash processed value. Afterward, the random password is sent to the user’s email. Figure 125 illustrates a demo of an email sent to the user.

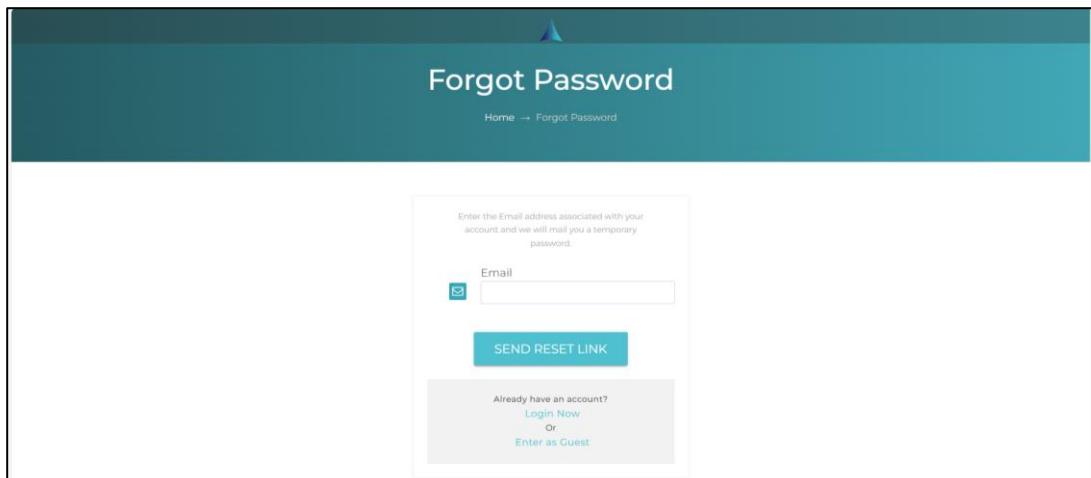


Figure 124 Forgot Password' Interface

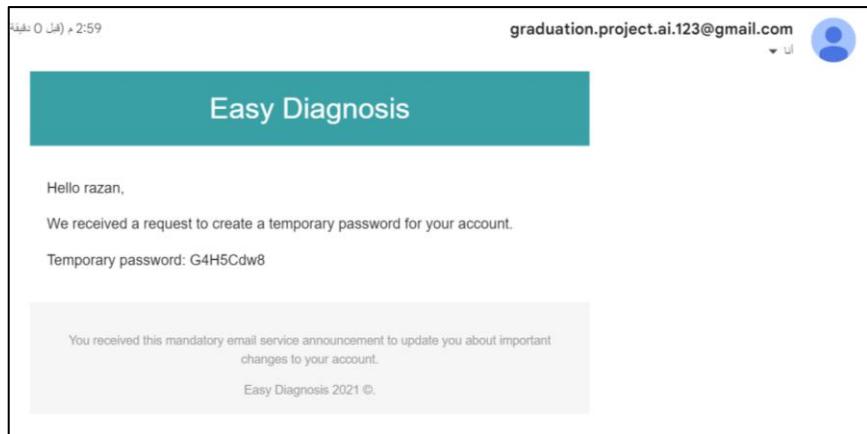


Figure 125 Confirmation Email

### 6.5.3 Registration Page

In the case of an unregistered user, the user can click on the ‘Register as New User’ link to be redirected to the registration page. The users enter their information on a page developed using the WTForms library. The MongoDB query is then used to check the validity of the data entered. If the data entered is valid, an email message is sent to confirm the user’s registration. Figure 126 shows the registration interface.

Figure 126 Registration Form’ Interface

### 6.5.4 Email Confirmation Token

The “URLSafeTimedSerializer” function provided by the “itsdangerous” library in Python was utilized to confirm the user’s email address. This function generates a Uniform Resource Locator (URL) that contains the user’s email, and the creation timestamp hashed with salt to scramble the information in the URL. The URL stays legitimate for 24 hours from when a user receives it. Figures 127 and 128 show instances of confirmation messages.

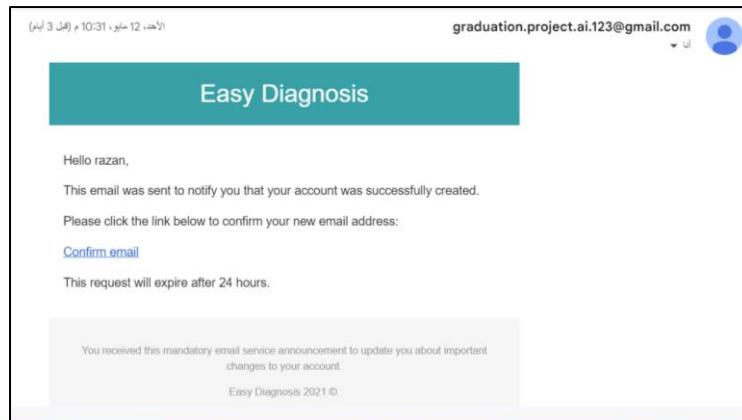


Figure 127 Confirmation email after the user creates an account.

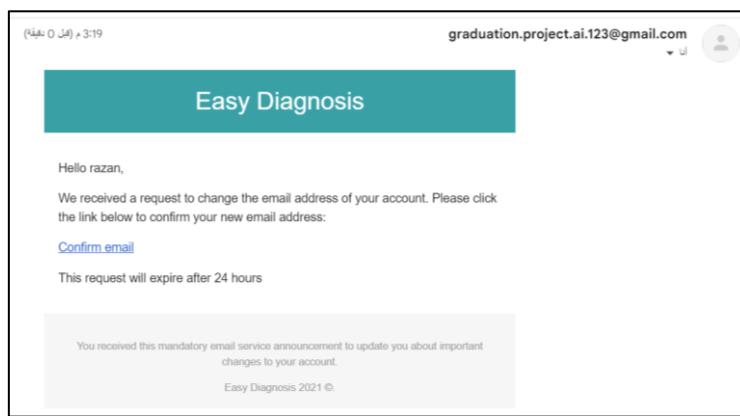


Figure 128 Confirmation email after a user changes their email address.

After clicking on the URL present in the confirmation email, three potential cases might occur: The email is confirmed effectively, the URL is invalid or has been expired, or the email has just been confirmed. Figures 129, 130, and 131 below display the three cases separately.

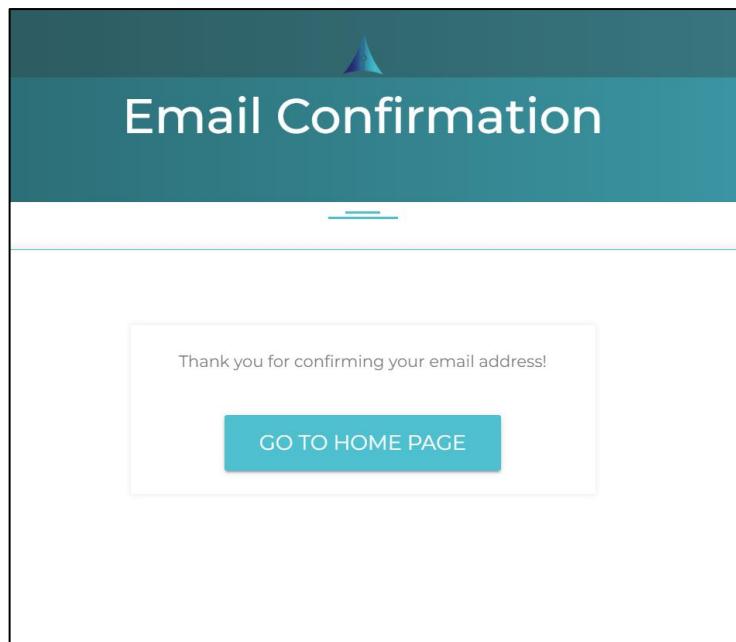


Figure 129 The email is confirmed successfully.

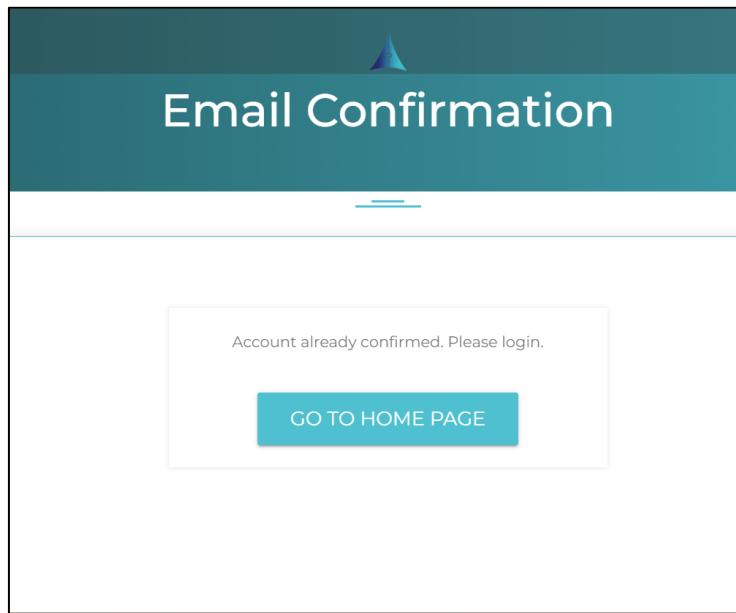


Figure 130 URL for an email has been already confirmed.

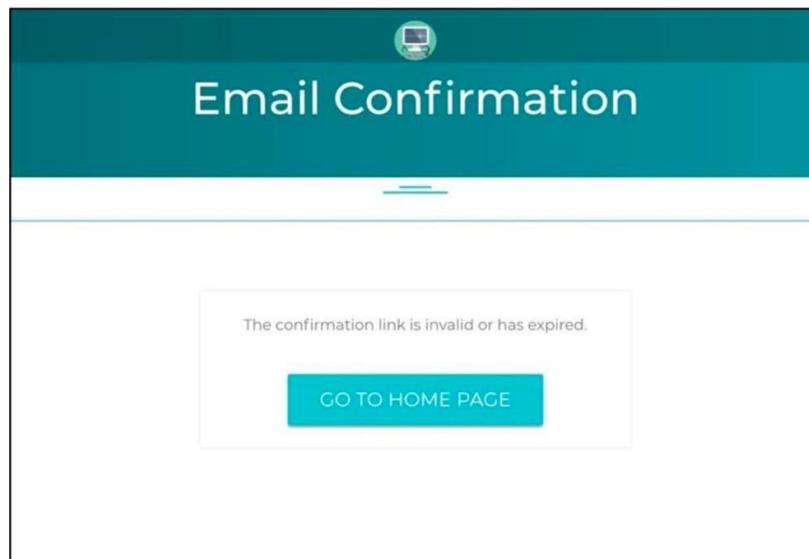


Figure 131 URL has been expired.

### 6.5.5 Profile Page

The “Profile” interface is common between the three system users (Admin, Medical specialist, Laboratory specialists, and Registered user), with minor differences. The information about a user is displayed through this interface with their privileges based on the username stored in the session after the user has logged in. Figures 132, 133, 134, and 135 demonstrate the dashboard that includes links to the interfaces, where each user can perform their functionalities. All users can modify their password and email address. Registered users can also change their name, birthdate, and gender, as shown in Figure 135. In addition, the registered user can delete their account, which will be deleted from the database where the username is the same as the username stored in the login session.

A screenshot of the 'Easy Diagnosis' Admin's Profile page. The top navigation bar includes a logo, the title 'Easy Diagnosis', a 'Logout' link, and a breadcrumb trail 'Home → Profile'. The main content area is divided into two sections: 'Dashboard' on the left and 'Profile' on the right. The 'Dashboard' section contains links for 'Profile', 'Manage User', and 'Rebuild Model'. The 'Profile' section displays the user information for 'razan' (Admin) with the email 'razan3khalid@gmail.com'. It also includes 'Change Email' and 'Change Password' buttons.

Figure 132 Admin's Profile

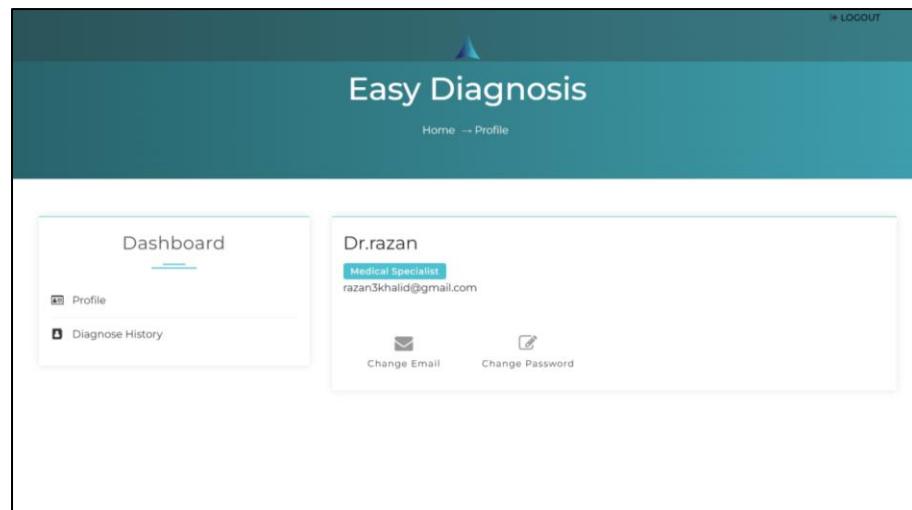


Figure 133 Medical Specialist's Profile

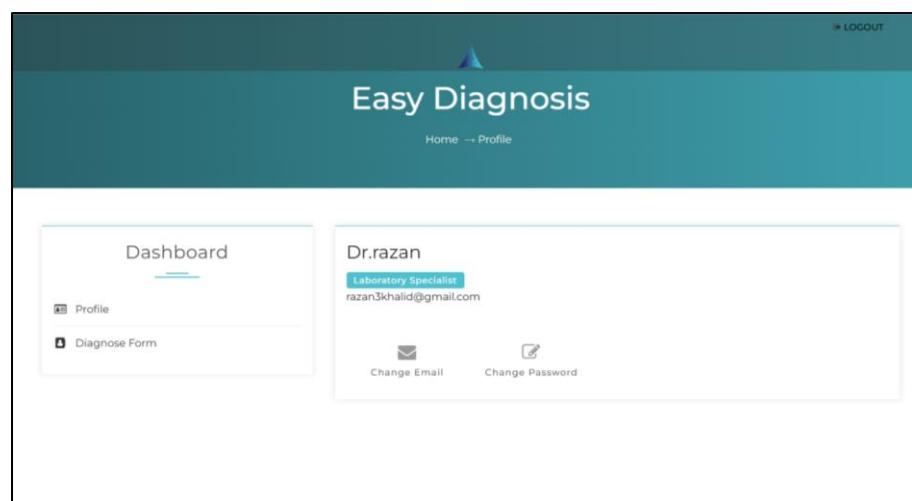


Figure 134 Laboratory Specialist's Profile

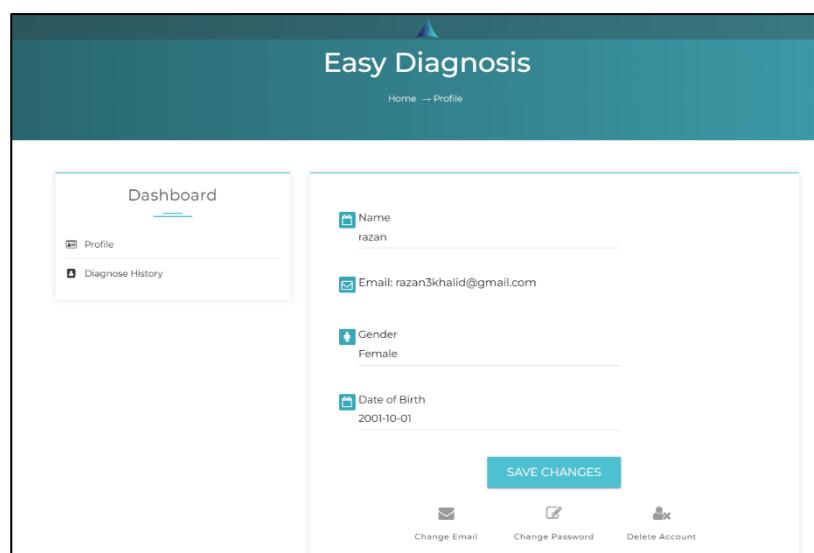
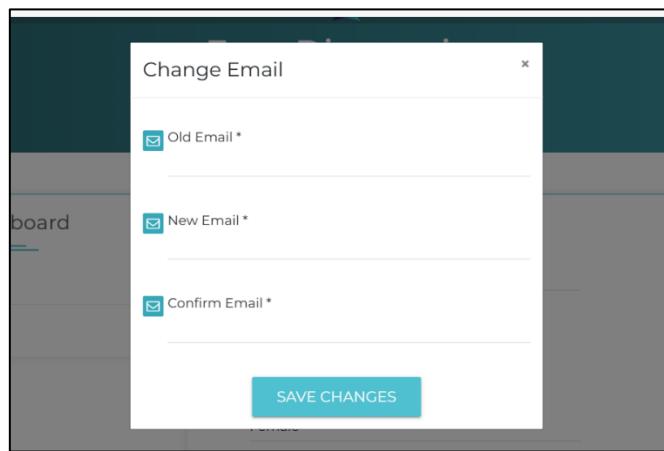


Figure 135 Registered User's Profile

#### 6.5.5.1 Change Email

The users can change their email addresses by clicking on the “Change Email” button, which redirects them to the “Change Email” interface.

In the “Change Email” interface, the users are asked to enter their old emails for verification purposes, as shown in Figure 136. Besides, the users are requested to enter their new email, which will be validated for: its format using a regular expression, whether it matches the old email, and whether it matches an email of another account. If no errors occur, the email address will be updated in the database, and an email will be sent to the user. As demonstrated in Figure 128, a confirmation link is provided to ensure that the email receiver is the change requester.



The screenshot shows a modal window titled "Change Email". Inside the window, there are three input fields: "Old Email \*", "New Email \*", and "Confirm Email \*". Each field has a small icon of an envelope next to it. Below the fields is a blue button labeled "SAVE CHANGES". The background of the modal is white, and the overall design is clean and modern.

Figure 136 Change Email

#### 6.5.5.2 Change Password

The users can change their email addresses by clicking on the “Change Password” button, which redirects them to the “Change Password” interface.

In the “Change Password” interface, the users are asked to enter their old password for verification purposes, as shown in Figure 137. Besides, the users are requested to enter their new password, which will be validated for including at least 8 alphanumeric characters using regular expressions, and whether it matches the old password. If no errors occur, the password will be updated, and an email will be sent to confirm the change, as demonstrated in Figure 138.

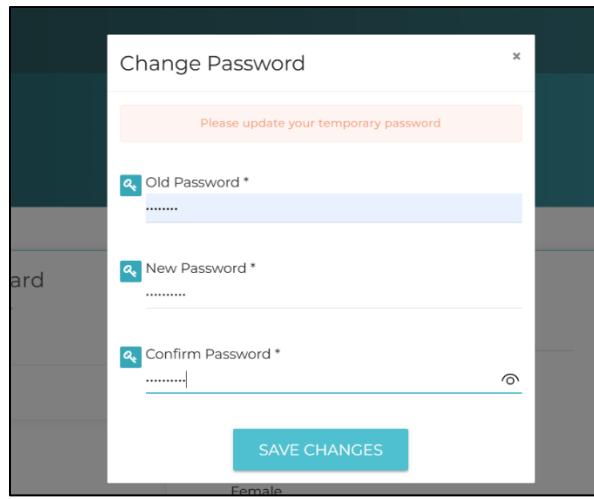


Figure 137 Change Password

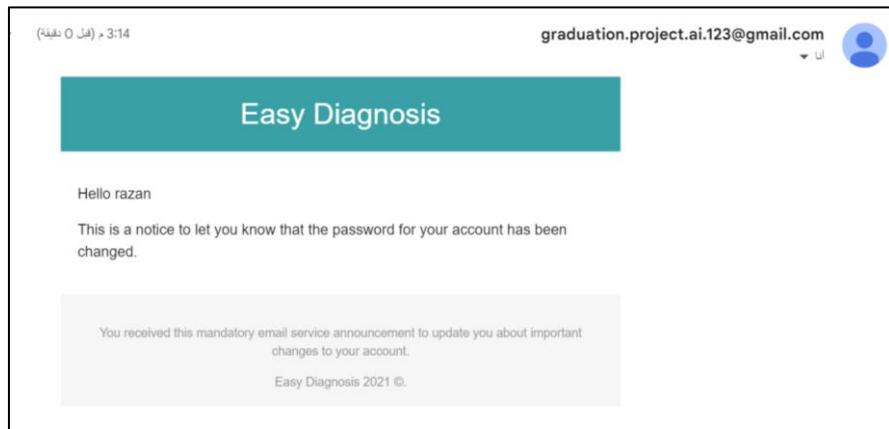


Figure 138 Change Password Confirmation Email

### 6.5.6 Diagnose Interface

The “Diagnose” interface enables medical specialists, registered users, and guests to perform diagnosis by entering the required information. The diagnosis interface uses a MongoDB query to get the name of the active model, which will be used to download the model file from the server, then the user information form will be tested using the loaded model. The result will indicate whether a patient will possibly have the disease or not. Figure 139 below shows the diagnosis form for registered users and guests.

The screenshot shows a web-based medical diagnostic application. At the top, a dark teal header bar contains the word "Diagnose" in white. Below the header is a navigation menu with links to various diseases and conditions. The main content area features a form for diagnosing "Chronic Kidney Disease". The form includes two input fields: "Blood Urea Nitrogen \*" and "Creatinine \*", each preceded by a small blue dropdown icon. A large blue "DIAGNOSE" button is positioned at the bottom right of the form.

Figure 139 Diagnosis form

The “Diagnosis” interface for the medical specialist differs slightly, as shown in Figure 140. The medical staff can display the users’ information by entering a patient’s MRN. Using the MongoDB query, the patient information will be displayed if it exists in the database. Otherwise, the medical specialist can fill in the patients’ information and register them in the system by storing their information in the database.

Figure 140 'Medical Specialist Diagnosis' Interface

The 'Diagnosis' interface for the laboratory specialists differs from the medical specialists, registered users, and the guests, as shown in Figure 141. The laboratory specialists can choose the desired disease/s from a dropdown list and the features will be displayed based on the selection. To determine if a patient's MRN is stored in the system, laboratory specialists can enter the patient's MRN and then click the 'Retrieve' button. If the MRN exists, the laboratory specialists can then fill in the required features, and the results will be saved in the database, which can then be accessed by the medical specialists and registered users.

Figure 141 'Laboratory Specialist Diagnosis' Interface

### 6.5.7 Diagnose Result

Figure 142 below shows the ‘Result’ interface, in which, after a diagnosis, the website transmits the result, disease, and the accuracy to the ‘Result’ interface. Afterward, the website redirects the user to the ‘Result’ interface using the render\_template method.

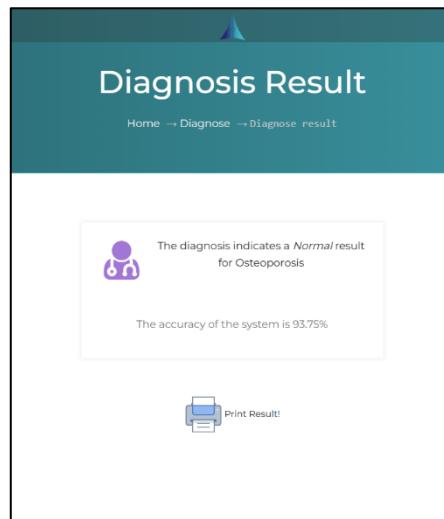


Figure 142 Result Interface

### 6.5.8 View History

The medical specialists and registered users can view past diagnosis results using the “View History” interface, as shown in Figure 143. A medical specialist has the privilege of viewing the history of existing users by entering their MRN. A MongoDB query is used to retrieve the previous diagnosis results. However, a registered user has only the privilege to view their past diagnosis results, as shown in Figure 144.

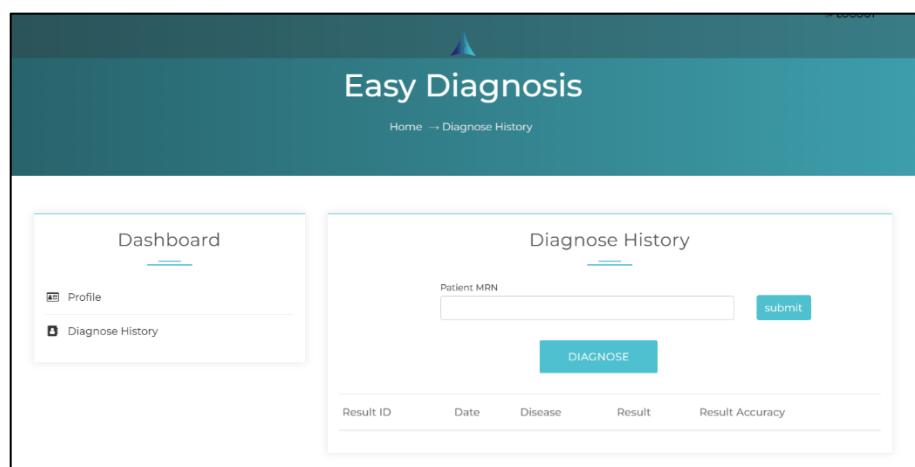


Figure 143 Medical Specialist’s View of Diagnosis History’ interface

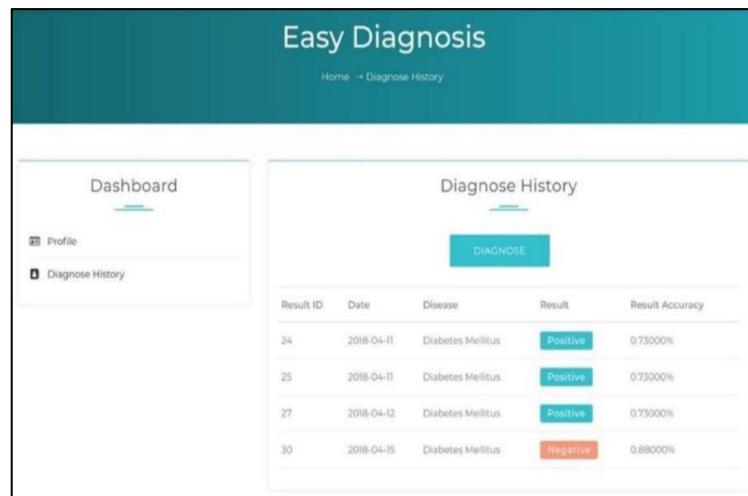


Figure 144 Registered User's View of Diagnosis History' Interface

### 6.5.9 Manage Users

The “Manage User” interface can only be accessed by the admin to manage the system’s users, as shown in Figure 145. Using JavaScript, the admin can delete a specific user by searching for their username. The chosen usernames will be returned to the server-side handling methods, where they are deleted from the database. Before the users’ accounts are deleted, their email addresses are saved. Later, an email is sent to inform them that their accounts have been deleted.

Easy Diagnosis			
Manage User			
<a href="#">Add user</a> <a href="#">Remove user</a>			
<input type="text" value="Search by username.."/>			
Username	Name	Role	Current Email
ranaqht	Rana	admin	ranaalqahtanii99@gmail.com
mashaelalmusairii	mashael	laboratory specialist	mashaelalmusairii@gmail.com
nawalthelearner	Nawal	medical specialist	nawalthelearner@gmail.com
admin	admin	admin	admin@admin.com
rafeqht	Rafeq Ali	laboratory specialist	rafeqht@gmail.com
refalqht	Refal Alqahtani	medical specialist	refalqht102@gmail.com

Figure 145 Manage Users Interface

The admin can add a user, as shown in Figure 146, by entering the user’s name, email, and role. These inputs values are verified at the client-side by utilizing JavaScript and verified at the server-side by comparing them to the database data, such as checking if the entered email is already taken by another user. Afterward, a username and temporary password are provided and sent to the new user, as demonstrated in Figure 147.

Figure 146 Add User

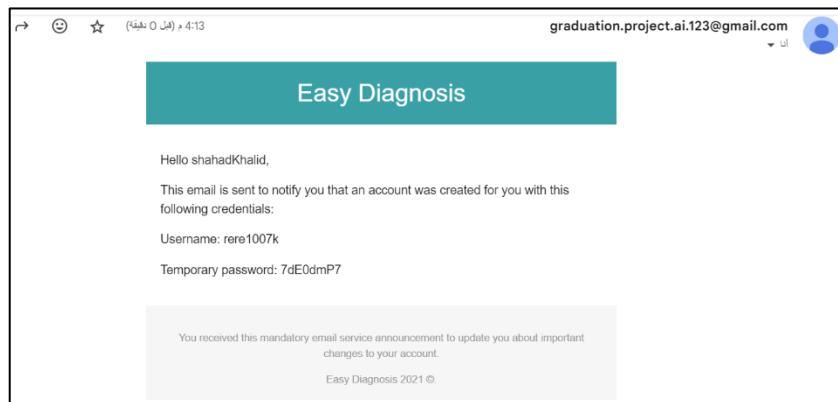


Figure 147 Add User Email

### 6.5.10 Rebuild Diagnostic Model

The “Rebuild Diagnostic Model” interface can only be accessed by the admin. The steps of rebuilding the diagnostic model consist of two processes. The first process is to upload the dataset file that must be in the “.csv” format, while the second process is to select the disease and insert the training percentage. Figure 148 shows the rebuilt model interface.

Figure 148 Rebuild Model

After validating the input data, they are returned to the handling method, which stores the dataset file in the data folder at the server to build the model. The dataset is then divided into training and testing sets by a percentage specified by the admin. Afterward, the training set is fed into the classifier for training. The resulted model is named by the current date and time and is saved in the model's folder. Subsequently, the model is tested using the testing set to calculate the accuracy. Finally, the model's information is inserted into the database as a temp model.

The models are displayed to the admin in a table, as shown in Figure 148. Once the admin is fulfilled with one of those models, he/she can select the model and click the “Update Demonstrative Model” button, which upgrades the current operating model for that disease within the database, in which the model will be utilized over the diagnosis procedure. The other produced temp models for that disease will be deleted from both the database and the model's folder. The admin can also see the current working models for both diseases by clicking the “Show Current Active Model” button.

## Chapter 7: Software Testing Plan

The testing plan, which includes test strategies to evaluate the project's functionality and soundness, is defined in this chapter.

### 7.1 Test Items

The Easy Diagnosis website is effective and easy to use. In order to verify the validity of the features covered in the software requirements specification (SRS) document, the software testing division will test each and every webpage. Software design standards (SDS) also included the implementation and user interfaces for websites. In addition, every webpage on the website has undergone testing to guarantee optimal performance in terms of user interface standards. One of the test elements to guarantee the correctness of the data on the website is the data and database integrity.

### 7.2 Features to be Tested

All of the tested features are listed in the criteria, which are explained in the project proposal's SRS and SDS chapters.

### 7.3 Approach

All of the test types that have been included on the Easy Diagnosis website are included in the list below:

- Testing the integrity of data and databases.
- User interface testing
- integration testing
- component testing
- Testing of interfaces.
- Testing for verification and validation.
- Examining security.
- Execution evaluation.
- Limitations.
- Testing in beta.

#### 7.3.1 Data and Database Integrity Testing

To provide a dependable and precise outcome and facilitate the identification of any corruption, the database has to be checked independently from the website.  
Method:

Use database operations like insert, update, and delete to test them with both valid and erroneous data.

Check that any tables' data that has been added or altered is compatible with the database.

Verify the information that was retrieved is accurate.

Completion Criteria:

There must be no data corruption or inconsistency in the database's performance.

Special considerations:

- To facilitate the visualization and detection of faults, a sample of the database will be used for the database test.

### 7.3.1.1 Test Cases

Test cases are included but are not limited to the following items in Table 109.

Test ID	Test Description	Expected Result	Verified (Yes/No)
1	Update a record	The record is updated	Yes
2	Insert a record	The record is inserted into the specified table	Yes
3	Delete a record	<ul style="list-style-type: none"> <li>• The specified record is deleted.</li> <li>• If the specified record ID to be deleted acts as a foreign key in other tables, then the proper deletion action will occur on the affected records.</li> </ul>	Yes
4	Duplicate value of a primary key	Forbidden Procedure	Yes
5	Retrieve records from a table	Records are retrieved from the correct table and match the criteria specified upon retrieval.	Yes
6	Insert or update a value to “null” in a column that has the constraint “Not Null”	Forbidden Procedure	Yes
7	Insert a record that has field types that are inconsistent with inserted data	Forbidden Procedure	Yes
8	Insert a record with data length that is greater than the specified field length	Forbidden Procedure	Yes
9	Insert a record with a foreign key value that does not match any primary key values in the other table.	Forbidden Procedure	Yes

Table 109 Data and database integrity test cases

### 7.3.2 Component Testing

The type of testing used to evaluate each website component's functionality independently is called component testing.

Technique:

Prior to integrating any component with other website components, evaluate each one's performance independently. The use cases listed in the SRS document serve as the foundation

for the testing cases. Entering both valid and incorrect data allows you to verify the following during this process:

- When using valid data, all components function as intended to provide the desired outcomes.
- When using incorrect data, all components show the relevant error messages and don't function at all until they are fed valid data.

Completion criteria:

- Every part functions flawlessly and without any errors.

Special consideration:

- Divide all the parts into smaller sub-functions and test one sub-function at a time.

### 7.3.2.1 Test Cases

#### 7.3.2.1.1 Login

The following Table 110 shows test cases for the ‘Login’ function.

Test ID	Login_001
Prerequisite	The user has an account
Test Procedure	<p>Each test procedure was applied separately to all roles (Admin, Medical Specialist, Laboratory Specialist, and Registered User). Press the “Login” button after performing each of the following:</p> <ol style="list-style-type: none"> <li>1. Type the correct username and password.</li> <li>2. Type the correct username capitalizing any small letter in it and correct the password.</li> <li>3. Leave username and/or password empty.</li> <li>4. Type the correct username and wrong password.</li> <li>5. Type the correct username and type the correct password capitalizing any small letter in it.</li> </ol>
Expected Result	<p>The result of each of the previously mentioned test procedures:</p> <ol style="list-style-type: none"> <li>1. Login successfully and redirect the user to their interface.</li> <li>2. Login successfully and redirect the user to their interface.</li> <li>3. Display an error "Invalid Entries" and clear the username and password fields.</li> <li>4. An error message is displayed “Invalid Username, Password.” and clear the username and password fields.</li> <li>5. An error message is displayed “Invalid Username, Password.” and clear the username and password fields.</li> </ol>
Actual Result	Matching the expected results.
Verified (Yes/No)	Yes

Table 110 Test Cases of ‘Login’ feature

### 7.3.2.1.2 Reset Password

Table 111 shows test cases for the ‘Reset Password’ function.

Test ID	<b>Reset_Password_001</b>
Prerequisite	The user has an account
Test Procedure	<p>The test procedures are applied separately after clicking the “Send Reset Link” button.</p> <ol style="list-style-type: none"> <li>1. Type an existing email address.</li> <li>2. Type a non-existing email address.</li> <li>3. Type an invalid email syntax.</li> </ol>
Expected Result	<p>The result of the previous separate procedures:</p> <ol style="list-style-type: none"> <li>1. An email that includes the user’s temporary password has been sent to the specified email address.</li> <li>2. An error message is displayed: “this email is not registered”.</li> <li>3. An error message is displayed: “Incorrect email address, email address should be in the form of someone@example.com”.</li> </ol>
Actual Result	The same as the expected results.
Verified (Yes/No)	Yes

Table 111 Test Cases of ‘Reset Password’ feature

### 7.3.2.1.3 Create Account

Table 112 shows test cases for the ‘Create Account’ function.

Test ID	<b>Create_Account_001</b>
Prerequisite	None
Test Procedure	<p>The test procedures are applied separately after clicking the “Create Account” button.</p> <ol style="list-style-type: none"> <li>1. Fill all the fields with valid data.</li> <li>2. Leave some or all the fields empty.</li> <li>3. Auto generated MRN.</li> <li>4. Type all fields with valid data except: <ul style="list-style-type: none"> <li>a. Invalid email syntax.</li> <li>b. Password with only numbers.</li> <li>c. Password with only letters.</li> <li>d. Password is less than the required length.</li> <li>e. Password does not match the confirmed password.</li> <li>f. Name not starting with a letter.</li> <li>g. Name with more than 40 characters.</li> <li>h. Name with less than 3 characters.</li> <li>i. Not selecting a gender.</li> <li>j. Not selecting a birthdate.</li> </ul> </li> </ol>
Expected Result	The result of the previous separate procedures:

	<ol style="list-style-type: none"> <li>1. Successfully creating an account and new records are added to the account record in the database.</li> <li>2. An error message is displayed: “You should fill in the required fields”.</li> <li>3. Each of the following will generate the following result:             <ol style="list-style-type: none"> <li>a. An error message is displayed “Incorrect email address, email address should be in the form of someone@example.com”.</li> <li>b. An error message is displayed: “Password should contain letter and number”.</li> <li>c. An error message is displayed: “Password should contain letter and number”.</li> <li>d. An error message is displayed: “ Password should contain at least 6 characters”.</li> <li>e. An error message is displayed: “ Password doesn’t match the confirmed password”.</li> <li>f. An error message is displayed: “ Name should start with letters”.</li> <li>g. An error message is displayed: “ Name is too long”.</li> <li>h. An error message is displayed: “ Name is too short”.</li> <li>i. An error message is displayed: “Must select gender”.</li> <li>j. An error message is displayed: “Must select a date of birth”.</li> </ol> </li> <li>4. Updating Patient’s information and new records are added to the account record in the database.</li> </ol>
Actual Result	The same as the expected results.
Verified (Yes/No)	Yes

Table 112 Test Cases of ‘Create Account’ feature

#### 7.3.2.1.4 Update Profile

Table 113 shows the ‘Update Profile’ feature test cases. Test cases include, but are not limited to, the following items in the table.

Test ID	Update_Profile_001
Prerequisite	The registered user must be logged into the website.
Test Procedure	<p>The test procedures are done independently after clicking the “Save Changes” button.</p> <ol style="list-style-type: none"> <li>1. Fill all fields with valid data.</li> <li>2. Leave the name field empty.</li> <li>3. Fields are typed with valid data except:             <ol style="list-style-type: none"> <li>a. The name includes numbers.</li> </ol> </li> </ol>

	<ul style="list-style-type: none"> <li>b. The name is longer than 40 characters.</li> <li>c. The name is less than 3 characters.</li> </ul>
Expected Result	<p>The outcome of the prior independent procedures:</p> <ol style="list-style-type: none"> <li>1. Successfully update the registered user profile information with the new entries and display a confirmation message: “profile updated”.</li> <li>2. An error message is displayed: “The name is a required field”.</li> <li>3. The following are the expected outcomes: <ul style="list-style-type: none"> <li>a. An error message is displayed: “Invalid Name, numbers are not allowed”.</li> <li>b. An error message is displayed: “Name is too long”.</li> <li>c. An error message is displayed: “Name is too short”.</li> </ul> </li> </ol>
Actual Result	Matching the expected outcome.
Verified (Yes/No)	Yes

Table 113 Test cases of 'Update Profile' feature

#### 7.3.2.1.5 Change Email

Table 114 shows the ‘Change Email’ feature test cases. Test cases include, but are not limited to, the following items in the table.

Test ID	Change_Email_001
Prerequisite	The user must be logged into the website.
Test Procedure	<p>The test procedures are done independently after clicking the “Change Email” button.</p> <ol style="list-style-type: none"> <li>1. Fill all fields with valid data.</li> <li>2. Some or all the fields are empty.</li> <li>3. Fields are typed with valid data except: <ul style="list-style-type: none"> <li>a. Old email does not match with the logged-in user record.</li> <li>b. Invalid email format.</li> <li>c. The confirmed email does not match the new email.</li> <li>d. The old email matches the new email.</li> <li>e. The new email already exists in another account.</li> </ul> </li> </ol>
Expected Result	<p>The outcome of the prior independent procedures:</p> <ol style="list-style-type: none"> <li>1. A confirmation email is sent to the user’s old email address. Then email address is updated with the new email, and a confirmation message is displayed to the user: “Email is changed”.</li> <li>2. Change email is rejected; an error message is displayed: “All mandatory data must be filled”.</li> <li>3. The following are the expected outcomes: <ul style="list-style-type: none"> <li>a. An error message is displayed: “Old email is incorrect”.</li> <li>b. An error message is displayed: “Incorrect email form, email must be in the form of someone@example.com”.</li> </ul> </li> </ol>

	<ul style="list-style-type: none"> <li>c. An error message is displayed: “Confirmed email does not match the new email”.</li> <li>d. An error message is displayed: “New email matches the old email”.</li> <li>e. An error message is displayed: “New email already exists in another account”.</li> </ul>
Actual Result	Matching the expected outcome.
Verified (Yes/No)	Yes

Table 114 Test cases of 'Change Email' feature

#### 7.3.2.1.6 Change Password

Table 115 shows the ‘Change Password’ feature test cases. Test cases include, but are not limited to, the following items in the table.

Test ID	Change_Password_001
Prerequisite	The user must be logged into the website.
Test Procedure	<p>The test procedures are done independently after clicking the “Change Password” button.</p> <ol style="list-style-type: none"> <li>1. Fill all fields with valid data.</li> <li>2. Some or all the fields are empty.</li> <li>3. Fields are typed with valid data except:           <ol style="list-style-type: none"> <li>a. The old password does not match with the logged in user record.</li> <li>b. The new password’s length is less than 8 characters.</li> <li>c. The new password does not include any digits.</li> <li>d. The new password does not include any characters.</li> <li>e. The confirmed password does not match the new password.</li> <li>f. The old password matches the new password.</li> </ol> </li> </ol>
Expected Result	<p>The outcomes of prior independent procedures:</p> <ol style="list-style-type: none"> <li>1. Users’ password is successfully changed, and an informative email is sent to the logged in user’s email address.</li> <li>2. Change password is rejected; an error message is displayed: “All mandatory data must be filled”.</li> <li>3. The following are the expected outcomes:           <ol style="list-style-type: none"> <li>a. An error message is displayed: “Old password is incorrect”.</li> <li>b. An error message is displayed: “Password must be at least 8 alphanumeric characters”.</li> <li>c. An error message is displayed: “Password must include digits and characters”.</li> <li>d. An error message is displayed: “Password must include digits and characters”.</li> </ol> </li> </ol>

	<p>e. An error message is displayed: “Confirmed password does not match the new password”.</p> <p>f. An error message is displayed: “Old password matches the new password”.</p>
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 115 Test cases of 'Change Password' feature

### 7.3.2.1.7 Add User

Table 116 presents the test cases of ‘Add User’ functionality.

Test ID	Add_User_001
Prerequisite	Admin is logged into the system.
Test Procedure	<p>The test procedures are applied separately after clicking “Add user” button.</p> <ol style="list-style-type: none"> <li>1. Fill all the fields with valid data about the new user and select the correct role.</li> <li>2. Some or all the fields are empty.</li> <li>3. Fields are filled with valid data except:           <ol style="list-style-type: none"> <li>a. The email already exists.</li> <li>b. Invalid email format.</li> <li>c. The Name contains digits.</li> <li>d. The Name is more than 40 characters.</li> <li>e. The Name less than 3 characters.</li> </ol> </li> </ol>
Expected Result	<p>The results of the previous separate procedures:</p> <ol style="list-style-type: none"> <li>1. Add new user successfully; new record is added to the account table.</li> <li>2. Add new user is rejected, an error message is displayed: “Please fill all the required fields”.</li> <li>3. Add new user is rejected and the expected responses are:           <ol style="list-style-type: none"> <li>a. An error message is displayed “Email already exists in another account”.</li> <li>b. An error message is displayed “Incorrect email form, email should be in the form someone@example.com”.</li> <li>c. An error message is displayed “Name should contain characters only”.</li> <li>d. An error message is displayed: “Name is too long”.</li> <li>e. An error message is displayed: “Name is too short”.</li> </ol> </li> </ol>
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 116 Add User Test Case

### 7.3.2.1.8 Delete User

Table 117 presents the test cases of ‘Delete User’ functionality.

Test ID	Delete_User_001
Prerequisite	Admin is logged into the system.
Test Procedure	The test procedures are applied separately after clicking the “Remove” button: <ol style="list-style-type: none"> <li>1. Select a user to be removed from the table.</li> <li>2. Click the button without selecting a user.</li> </ol>
Expected Result	The results of the previous separate procedures: <ol style="list-style-type: none"> <li>1. The user will be removed successfully; the record will be removed from the database.</li> <li>2. An error message is displayed: “A user must be selected”.</li> </ol>
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 117 Delete User Test Case

### 7.3.2.1.9 Rebuild Model

Table 118 presents the test cases of “Rebuild Model” functionality.

Test ID	Rebuild_Model_001
Prerequisite	Admin is logged into the system.
Test Procedure	The test procedures are applied separately after clicking “Generate Model” button. <ol style="list-style-type: none"> <li>1. Upload the dataset file with “.csv” extension and provide training percentage.</li> <li>2. Fields are filled with valid data except:               <ol style="list-style-type: none"> <li>a. File not uploaded.</li> <li>b. Upload a dataset file with an invalid extension.</li> <li>c. Disease not selected.</li> </ol> </li> </ol>
Expected Result	The results of the previous separate procedures: <ol style="list-style-type: none"> <li>1. The system starts training and testing using the uploaded dataset; then a model is generated with information about the model, such as classification accuracy and the total number of instances.</li> <li>2. The following are the expected outcomes:               <ol style="list-style-type: none"> <li>a. An error message is displayed: “Dataset file not uploaded”.</li> <li>b. An error message is displayed: “Dataset file should have extension .csv”.</li> <li>c. An error message is displayed: “Disease not selected”.</li> </ol> </li> </ol>
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 118 Rebuild Model Test Case

### 7.3.2.1.10 Update Model

Table 119 presents the test cases of ‘Update Model’ functionality.

Test ID	Update_Model_001
Prerequisite	Admin is logged into the system.
Test Procedure	The test procedures are applied separately after clicking “Update Diagnostic Model” button. 1. Select a model with the desired accuracy from the model's table. 2. Click the button without selecting a model.
Expected Result	The results of the previous separate procedures: 1. The chosen model is stored with accuracy for future reference and the other trained models are deleted. 2. An error message is displayed “A model should be selected”.
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 119 Update Model Test Case

### 7.3.2.1.11 Medical Specialist’s View of Diagnosis History

Table 120 presents the test cases of ‘Diagnosis History’ functionality.

Test ID	View_Diagnosis_History_001
Prerequisite	The medical specialist is logged into the system.
Test Procedure	Test the following procedures: 1. Fill in a valid MRN. 2. Medical specialist enters an invalid MRN.
Expected Result	1. The patient’s diagnosis history is displayed. 2. An error message is displayed: “Invalid MRN”.
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 120 Medical Specialist’s View of Diagnosis History Test Case

### 7.3.2.1.12 Registered User’s View of Diagnosis History

Table 121 presents the test cases of “Diagnosis History” functionality.

Test ID	View_Diagnosis_History_002
Prerequisite	The registered user is logged into the system.
Test Procedure	The diagnosis history is displayed.
Expected Result	The user’s diagnosis history is displayed.
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 121 Registered User’s View of Diagnosis History Test Case

### 7.3.2.1.13 Laboratory Specialist's Diagnosis

Table 122 presents the test cases of “Diagnosis” functionality.

Test ID	LaboratorySpecialist_Diagnosis _001
Prerequisite	The laboratory specialist is logged into the system.
Test Procedure	<ol style="list-style-type: none"> <li>1. Press the “Check if patient exists” button after filling it with correct MRN.</li> <li>2. Press the “Check if patient exists” button after filling it with wrong MRN.</li> <li>3. Press the “Save” button after performing all the following:             <ol style="list-style-type: none"> <li>a. Enter the patients’ MRN.</li> <li>b. Choosing the selected disease/s.</li> <li>c. Enter all the test results.</li> </ol> </li> <li>4. Press the “Save” button with leaving some or all fields empty.</li> </ol>
Expected Result	<ol style="list-style-type: none"> <li>1. Show a message that the patients’ MRN exists “Valid MRN”.</li> <li>2. Show an error message that the patients’ MRN does not exist “Invalid patient MRN, please enter a valid MRN”.</li> <li>3. The data is saved successfully to the dataset.</li> <li>4. Show an error message if the following:             <ol style="list-style-type: none"> <li>a. An error message is displayed when some or all the fields are empty “Enter all required fields”.</li> <li>b. An error message is displayed when not entering numerical value “Invalid value, some required text fields must contain numeric values only”.</li> </ol> </li> </ol>
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 122 Laboratory Specialist’s Diagnosis Test Case

### 7.3.2.1.14 Diagnose Diabetes Mellitus Disease

Table 123 presents the test cases of ‘Diagnose Diabetes Mellitus Disease’ functionality.

Test ID	Diagnose_Diabetes_Mellitus_Disease
Prerequisite	The user logged into the system.
Test Procedure	<ol style="list-style-type: none"> <li>1. Press the “Diagnose” button after performing all of the following:             <ol style="list-style-type: none"> <li>a. Enter a valid Sugar Level test result.</li> <li>b. Enter a valid Hematocrit Level test result.</li> <li>c. Enter a valid Mean Platelet Volume (MPV) test result.</li> </ol> </li> <li>2. Press the “Diagnose” button while leaving some or all fields empty.</li> </ol>
Expected Result	<ol style="list-style-type: none"> <li>1. Successfully show the diagnose result.</li> <li>2. Show an error message if all the fields are empty “Enter all required fields”. If some are missing, then each of the following will generate an error message:</li> </ol>

	<ul style="list-style-type: none"> <li>a. An error message is displayed “Invalid Sugar Level test result, must enter numbers only”.</li> <li>b. An error message is displayed “Invalid Hematocrit Level test result, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid MPV test result, must enter numbers only”.</li> </ul>
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 123 Diagnose Diabetes Mellitus Disease Test Case

#### 7.3.2.1.15 Diagnose Chronic Kidney Disease

Table 124 presents the test cases of “Diagnose Chronic Kidney Disease” functionality.

Test ID	Diagnose_Chronic_Kidney_Disease
Prerequisite	The user logged into the system.
Test Procedure	<ol style="list-style-type: none"> <li>1. Press the “Diagnose” button after performing all of the following:           <ol style="list-style-type: none"> <li>a. Enter a valid Blood Urea Nitrogen test result.</li> <li>b. Enter a valid Creatinine test result.</li> </ol> </li> <li>2. Press the “Diagnose” button while leaving some or all fields empty.</li> </ol>
Expected Result	<ol style="list-style-type: none"> <li>1. Successfully show the diagnose result.</li> <li>2. Show an error message if all the fields are empty “Enter all required fields”. If some are missing, then each of the following will generate an error message:           <ol style="list-style-type: none"> <li>a. An error message is displayed “Invalid Blood Urea Nitrogen test result, must enter numbers only”.</li> <li>b. An error message is displayed “Invalid Creatinine test result, must enter numbers only”.</li> </ol> </li> </ol>
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 124 Diagnose Chronic Kidney Disease Test Case

#### 7.3.2.1.16 Diagnose Coronary Heart Disease

Table 125 presents the test cases of ‘Diagnose Coronary Heart Disease’ functionality.

Test ID	Diagnose_Coronary_Heart_Disease
Prerequisite	The user logged into the system.
Test Procedure	<ol style="list-style-type: none"> <li>1. Press the “Diagnose” button after performing all of the following:           <ol style="list-style-type: none"> <li>a. Select Gender from the drop-down list.</li> <li>b. Enter a valid Age.</li> <li>c. Enter a valid Mean Corpuscular Hemoglobin (MCH) test result.</li> </ol> </li> </ol>

	<p>d. Enter a valid Mean Corpuscular Hemoglobin Concentration (MCHC) test result.</p> <p>e. Enter a valid Red Cell Distribution Width (RDW) test result.</p> <p>f. Enter a valid Platelet Count test result.</p> <p>g. Enter a valid MPV test result.</p> <p>h. Enter a valid Mononucleosis Absolute test result.</p> <p>i. Enter a valid Basophil Instrument % test result.</p> <p>j. Enter a valid Basophil Instrument Absolute test result.</p> <p>k. Enter a valid Potassium test result.</p> <p>l. Enter a valid Hemoglobin test result.</p> <p>m. Enter a valid Gamma-Glutamyl Transpeptidase (GGTP) test result.</p> <p>n. Enter a valid Serum Glutamic-Oxaloacetic Transaminase (SGOT) test result.</p> <p>o. Enter a valid Serum Glutamic-Pyruvic Transaminase (SGPT) test result.</p> <p>p. Enter a valid Neutrophil Granulocyte Instrument test result.</p> <p>q. Enter a valid Neutrophil Granulocyte Instrument Absolute test result.</p> <p>r. Enter a valid Anion Gap test result.</p> <p>2. Press the “Diagnose” button with leaving some or all fields empty.</p>
Expected Result	<p>1. Successfully show the diagnose result.</p> <p>2. Show an error message if all the fields are empty “Enter all required fields”. If some are missing, then each of the following will generate an error message:</p> <ul style="list-style-type: none"> <li>a. An error message is displayed “The Gender is empty, choose from the drop-down list”.</li> <li>b. An error message is displayed “Invalid Age, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid MCH test result, you must enter numbers only”.</li> <li>d. An error message is displayed “Invalid MCHC test result, you must enter numbers only”.</li> <li>e. An error message is displayed “Invalid RDW test result, you must enter numbers only”.</li> <li>f. An error message is displayed “Invalid Platelet Count test result, you must enter numbers only”.</li> <li>g. An error message is displayed “Invalid MPV test result, you must enter numbers only”.</li> <li>h. An error message is displayed “Invalid Mononucleosis Absolute test result, you must enter numbers only”.</li> </ul>

	<ul style="list-style-type: none"> <li>i. An error message is displayed “Invalid Basophil Instrument % test result, you must enter numbers only”.</li> <li>j. An error message is displayed “Invalid Basophil Instrument Absolute test result, you must enter numbers only”.</li> <li>k. An error message is displayed “Invalid Potassium test result, you must enter numbers only”.</li> <li>l. An error message is displayed “Invalid Hemoglobin test result, you must enter numbers only”.</li> <li>m. An error message is displayed “Invalid GGTP test result, you must enter numbers only”.</li> <li>n. An error message is displayed “Invalid SGOT test result, you must enter numbers only”.</li> <li>o. An error message is displayed “Invalid SGPT test result, you must enter numbers only”.</li> <li>p. An error message is displayed “Invalid Neutrophil Granulocyte Instrument test result, you must enter numbers only”.</li> <li>q. An error message is displayed “Invalid Neutrophil Granulocyte Instrument Absolute test result, you must enter numbers only”.</li> <li>r. An error message is displayed “Invalid Anion Gap test result, you must enter numbers only”.</li> </ul>
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 125 Diagnose Coronary Heart Disease Test Case

#### 7.3.2.1.17 Diagnose Asthma Disease

Table 126 presents the test cases of ‘Diagnosis Asthma Disease’ functionality.

Test ID	Diagnose_Asthma_Disease
Prerequisite	The user logged into the system.
Test Procedure	<ol style="list-style-type: none"> <li>1. Press the “Diagnose” button after performing all the following:           <ol style="list-style-type: none"> <li>a. Select Gender from the drop-down list.</li> <li>b. Enter a valid Age.</li> <li>c. Enter a valid Basophil Instrument % test result.</li> <li>d. Enter a valid Hematocrit test result.</li> <li>e. Enter a valid Hemoglobin test result.</li> <li>f. Enter a valid MCH test result.</li> <li>g. Enter a valid MCHC test result.</li> <li>h. Enter a valid MPV test result.</li> <li>i. Enter a valid White Blood Cell count.</li> </ol> </li> <li>2. Press the “Diagnose” button while leaving some or all fields empty.</li> </ol>

Expected Results	<ol style="list-style-type: none"> <li>1. Successfully show the diagnose result.</li> <li>2. Show an error message if all the fields are empty “Enter all required fields”.             <ol style="list-style-type: none"> <li>a. An error message is displayed “The Gender is empty, choose from the drop-down list”.</li> <li>b. An error message is displayed “Invalid Age, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid Basophil Instrument % test result, must enter numbers only”.</li> <li>d. An error message is displayed “Invalid Hematocrit test result, must enter numbers only”.</li> <li>e. An error message is displayed “Invalid Hemoglobin test result, must enter numbers only”.</li> <li>f. An error message is displayed “Invalid MCH test result, must enter numbers only”.</li> <li>g. An error message is displayed “Invalid MCHC test result, must enter numbers only”.</li> <li>h. An error message is displayed “Invalid MPV test result, must enter numbers only”.</li> <li>i. An error message is displayed “Invalid White Blood Cell count, must enter numbers only”.</li> </ol> </li> </ol>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 126 Diagnose Asthma Disease Test Case

#### 7.3.2.1.18 Diagnose Thyroid Cancer Disease

Table 127 presents the test cases of ‘Diagnosis Thyroid Cancer Disease’ functionality.

Test ID	Diagnose_Thyroid_Cancer_Disease
Prerequisite	The user logged into the system.
Test Procedure	<ol style="list-style-type: none"> <li>1. Press the “Diagnose” button after performing all of the following:             <ol style="list-style-type: none"> <li>a. Select Gender status from the drop-down list.</li> <li>b. Enter a valid Age.</li> <li>c. Enter a valid Hematocrit.</li> <li>d. Enter a valid MCHC.</li> <li>e. Enter a valid MPV.</li> <li>f. Enter a valid Red Blood Cell count.</li> <li>g. Enter a valid White Blood Cell count.</li> </ol> </li> <li>2. Press the “Diagnose” button while leaving some or all fields empty.</li> </ol>
Expected Results	<ol style="list-style-type: none"> <li>1. Successfully show the diagnose result.</li> </ol>

	<p>2. Show an error message if all the fields are empty “Enter all required fields”. If some is missing, then each of the following will generate an error message:</p> <ul style="list-style-type: none"> <li>a. An error message is displayed “The Gender is empty, choose from the drop-down list”.</li> <li>b. An error message is displayed “Invalid Age, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid Hematocrit, must enter numbers only”.</li> <li>d. An error message is displayed “Invalid MCHC, must enter numbers only”.</li> <li>e. An error message is displayed “Invalid MPV, must enter numbers only”.</li> <li>f. An error message is displayed “Invalid Red Blood Cell count, must enter numbers only”.</li> <li>g. An error message is displayed “Invalid White Blood Cell count, must enter numbers only”.</li> </ul>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 127 Diagnose Thyroid Cancer Disease Test Case

#### 7.3.2.1.19 Diagnose Schizophrenia Disease

Table 128 presents the test cases of ‘Diagnosis Schizophrenia Disease’ functionality.

Test ID	Diagnose_Schizophrenia_Disease
Prerequisite	The user logged into the system.
Test Procedure	<p>1. Press the “Diagnose” button after performing all the following:</p> <ul style="list-style-type: none"> <li>a. Enter a valid Age.</li> <li>b. Enter a valid White Blood Cell.</li> <li>c. Enter a valid Hemoglobin.</li> <li>d. Enter a valid Hematocrit.</li> <li>e. Enter a valid Mean Corpuscular Volume (MCV).</li> <li>f. Enter a valid MCH.</li> <li>g. Enter a valid MCHC.</li> <li>h. Enter a valid Platelet.</li> <li>i. Enter a valid MPV.</li> <li>j. Enter a valid Aspartate Aminotransferase (AST).</li> <li>k. Enter a valid Total Protein.</li> <li>l. Enter a valid Gamma Glutamyl transferase (GGT).</li> </ul> <p>2. Press the “Diagnose” button while leaving some or all fields empty.</p>
Expected Results	<p>1. Successfully show the diagnose result.</p>

	<p>2. Show an error message if all the fields are empty “Enter all required fields”. If some is missing, then each of the following will generate an error message:</p> <ul style="list-style-type: none"> <li>a. An error message is displayed “Invalid Age, must enter numbers only”.</li> <li>b. An error message is displayed “Invalid White Blood Cell, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid Hemoglobin, must enter numbers only”.</li> <li>d. An error message is displayed “Invalid Hematocrit, must enter numbers only”.</li> <li>e. An error message is displayed “Invalid MCV, must enter numbers only”.</li> <li>f. An error message is displayed “Invalid MCH, must enter numbers only”.</li> <li>g. An error message is displayed “Invalid MCHC, must enter numbers only”.</li> <li>h. An error message is displayed “Invalid Platelet, must enter numbers only”.</li> <li>i. An error message is displayed “Invalid MPV, must enter numbers only”.</li> <li>j. An error message is displayed “Invalid AST, must enter numbers only”.</li> <li>k. An error message is displayed “Invalid Total Protein, must enter numbers only”.</li> <li>l. An error message is displayed “Invalid GGT, must enter numbers only”.</li> </ul>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 128 Diagnose Schizophrenia Disease Test Case

#### 7.3.2.1.20 Diagnose Glaucoma Disease

Table 129 presents the test cases of ‘Diagnosis Glaucoma Disease’ functionality.

Test ID	Diagnose_Glaucoma_Disease
Prerequisite	The user logged into the system.
Test Procedure	<p>1. Press the “Diagnose” button after performing all of the following:</p> <ul style="list-style-type: none"> <li>a. Enter a valid At test result.</li> <li>b. Enter a valid Ean test result.</li> <li>c. Enter a valid Mhci test result.</li> <li>d. Enter a valid Vaci test result.</li> <li>e. Enter a valid Varg test result.</li> <li>f. Enter a valid Vars test result.</li> </ul>

	<p>g. Enter a valid Tmi test result.</p> <p>2. Press the “Diagnose” button while leaving some or all fields empty.</p>
Expected Results	<p>1. Successfully show the diagnose result.</p> <p>2. Show an error message if all the fields are empty “Enter all required fields”. If some is missing, then each of the following will generate an error message:</p> <ul style="list-style-type: none"> <li>a. An error message is displayed “Invalid At test result, must enter numbers only”.</li> <li>b. An error message is displayed “Invalid Ean test result, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid Mhci test result, must enter numbers only”.</li> <li>d. An error message is displayed “Invalid Vasi test result, must enter numbers only”.</li> <li>e. An error message is displayed “Invalid Varg test result, must enter numbers only”.</li> <li>f. An error message is displayed “Invalid Vars test result, must enter numbers only”.</li> <li>g. An error message is displayed “Invalid Tmi test result, must enter numbers only”.</li> </ul>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 129 Diagnose Glaucoma Disease Test Case

### 7.3.2.1.21Diagnose Alzheimer’s Disease

Table 130 presents the test cases of ‘Diagnosis Alzheimer’s Disease’ functionality.

Test ID	Diagnose_Alzheimer’s_Disease
Prerequisite	The user logged into the system.
Test Procedure	<p>1. Press the “Diagnose” button after performing all of the following:</p> <ul style="list-style-type: none"> <li>a. Select Gender status from the drop-down list.</li> <li>b. Enter a valid Age.</li> <li>c. Enter a valid Pulse Ox test result.</li> <li>d. Enter a valid Respiratory Rate test result.</li> <li>e. Enter a valid BP – Diastolic test result.</li> <li>f. Enter a valid Red Blood Cell count test result.</li> <li>g. Enter a valid White Blood Cell count test result.</li> <li>h. Enter a valid Hemoglobin test result.</li> <li>i. Enter a valid Hematocrit test result.</li> <li>j. Enter a valid MCV test result.</li> <li>k. Enter a valid MCH test result.</li> <li>l. Enter a valid RDW test result.</li> <li>m. Enter a valid MPV test result.</li> </ul>

	2. Press the “Diagnose” button while leaving some or all fields empty.
Expected Results	<ol style="list-style-type: none"> <li>1. Successfully show the diagnose result.</li> <li>2. Show an error message if all the fields are empty “Enter all required fields”. If some is missing, then each of the following will generate an error message:             <ol style="list-style-type: none"> <li>a. An error message is displayed “The Gender is empty, choose from the drop-down list”.</li> <li>b. An error message is displayed “Invalid Age, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid Pulse Ox, must enter numbers only”.</li> <li>d. An error message is displayed “Invalid Respiratory Rate, must enter numbers only”.</li> <li>e. An error message is displayed “Invalid BP - Diastolic, must enter numbers only”.</li> <li>f. An error message is displayed “Invalid Red Blood Cell count, must enter numbers only”.</li> <li>g. An error message is displayed “Invalid White Blood Cell count, must enter numbers only”.</li> <li>h. An error message is displayed “Invalid Hemoglobin, must enter numbers only”.</li> <li>i. An error message is displayed “Invalid Hematocrit, must enter numbers only”.</li> <li>j. An error message is displayed “Invalid Mean Corpuscular Volume (MCV), must enter numbers only”.</li> <li>k. An error message is displayed “Invalid MCH, must enter numbers only”.</li> <li>l. An error message is displayed “Invalid RDW, must enter numbers only”.</li> <li>m. An error message is displayed “Invalid MPV, must enter numbers only”.</li> </ol> </li> </ol>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 130 Diagnose Alzheimer's Disease Test Case

#### 7.3.2.1.22 Diagnose Lung Cancer Disease

Table 131 presents the test cases of ‘Diagnosis Lung Cancer Disease’ functionality.

Test ID	Diagnose_Lung_Cancer_Disease
Prerequisite	The user logged into the system.

Test Procedure	<p>1. Press the “Diagnose” button after performing all the following:</p> <ol style="list-style-type: none"> <li>Select Gender from the drop-down list.</li> <li>Enter a valid Age.</li> <li>Select Smoking status from the drop-down list.</li> <li>Select Yellow Fingers status from the drop-down list.</li> <li>Select Anxiety status from the drop-down list.</li> <li>Select Wheezing status from the drop-down list.</li> <li>Enter a valid Peer Pressure test result.</li> <li>Select Chronic Disease status from the drop-down list.</li> <li>Select Fatigue status from the drop-down list.</li> <li>Select Allergy status from the drop-down list.</li> <li>Select Coughing status from the drop-down list.</li> <li>Select Alcohol status from the drop-down list.</li> <li>Select Shortness of Breath status from the drop-down list.</li> <li>Select Swallowing Difficulty status from the drop-down list.</li> <li>Select Chest Pain status from the drop-down list.</li> </ol> <p>2. Press the “Diagnose” button with leaving some or all fields empty.</p>
Expected Results	<p>1. Successfully show the diagnose result.</p> <p>2. Show an error message if all the fields are empty “Enter all required fields”. If the Age is missing, then following error message will generate:</p> <ol style="list-style-type: none"> <li>An error message is displayed “The Gender is empty, choose from the drop-down list”.</li> <li>An error message is displayed “Invalid Age, must enter numbers only”.</li> <li>An error message is displayed “The Smoking is empty, choose from the drop-down list”.</li> <li>An error message is displayed “The Yellow Fingers is empty, choose from the drop-down list”.</li> <li>An error message is displayed “The Anxiety is empty, choose from the drop-down list”.</li> <li>An error message is displayed “The Wheezing is empty, choose from the drop-down list”.</li> <li>An error message is displayed “Invalid Peer Pressure, must enter numbers only”.</li> <li>An error message is displayed “The Chronic Disease is empty, choose from the drop-down list”.</li> <li>An error message is displayed “The Fatigue is empty, choose from the drop-down list”.</li> <li>An error message is displayed “The Allergy is empty, choose from the drop-down list”.</li> <li>An error message is displayed “The Coughing is empty, choose from the drop-down list”.</li> </ol>

	<ol style="list-style-type: none"> <li>l. An error message is displayed “The Alcohol is empty, choose from the drop-down list”.</li> <li>m. An error message is displayed “The Shortness Of Breath is empty, choose from the drop-down list”.</li> <li>n. An error message is displayed “The Swallowing Difficulty is empty, choose from the drop-down list”.</li> <li>o. An error message is displayed “The Chest Pain is empty, choose from the drop-down list”.</li> </ol>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 131 Diagnose Lung Cancer Disease Test Case

#### 7.3.2.1.23 Diagnose Rheumatoid Arthritis Disease

Table 132 presents the test cases of ‘Rheumatoid Arthritis Disease’ functionality.

Test ID	Diagnose_Rheumatoid_Arthritis_Disease
Prerequisite	The user logged into the system.
Test Procedure	<ol style="list-style-type: none"> <li>1. Press the “Diagnose” button after performing all the following:           <ol style="list-style-type: none"> <li>a. Select Gender from the drop-down list.</li> <li>b. Enter a valid Age.</li> <li>c. Enter a valid Albumin test result.</li> <li>d. Enter a valid Alkaline phosphatase test result.</li> <li>e. Enter a valid Blood Urea Nitrogen test result.</li> <li>f. Enter a valid Chloride test result.</li> <li>g. Enter a valid Carbon dioxide test result.</li> <li>h. Enter a valid Creatinine test result.</li> <li>i. Enter a valid Direct Bilirubin test result.</li> <li>j. Enter a valid GGTP test result.</li> <li>k. Enter a valid Hemoglobin test result.</li> <li>l. Enter a valid Hematocrit Level test result.</li> <li>m. Enter a valid Potassium test result.</li> <li>n. Enter a valid Lactic Acid Dehydrogenase test result.</li> <li>o. Enter a valid MCH test result.</li> <li>p. Enter a valid MPV test result.</li> <li>q. Enter a valid MCHC test result.</li> <li>r. Enter a valid MCV test result.</li> <li>s. Enter a valid Sodium test result.</li> <li>t. Enter a valid Platelet test result.</li> <li>u. Enter a valid Red Blood Cell count.</li> <li>v. Enter a valid RDW test result.</li> <li>w. Enter a valid SGOT test result.</li> <li>x. Enter a valid SGPT test result.</li> <li>y. Enter a valid Total Bilirubin test result.</li> </ol> </li> </ol>

	<p>z. Enter a valid Total Protein test result.</p> <p>2. Press the “Diagnose” button with leaving some or all fields empty.</p>
Expected Results	<ol style="list-style-type: none"> <li>1. Successfully show the diagnose result.</li> <li>2. Show an error message if all the fields are empty “Enter all required fields”. If the Age is missing, then following error message will generate:             <ol style="list-style-type: none"> <li>a. An error message is displayed “The Gender is empty, choose from the drop-down list”.</li> <li>b. An error message is displayed “Invalid Age, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid Albumin, must enter numbers only”.</li> <li>d. An error message is displayed “Invalid Alkaline phosphatase, must enter numbers only”.</li> <li>e. An error message is displayed “Invalid Blood Urea Nitrogen, must enter numbers only”.</li> <li>f. An error message is displayed “Invalid Chloride, must enter numbers only”.</li> <li>g. An error message is displayed “Invalid Carbon dioxide, must enter numbers only”.</li> <li>h. An error message is displayed “Invalid Creatinine, must enter numbers only”.</li> <li>i. An error message is displayed “Invalid Direct Bilirubin, must enter numbers only”.</li> <li>j. An error message is displayed “Invalid GGTP, must enter numbers only”.</li> <li>k. An error message is displayed “Invalid Hemoglobin, must enter numbers only”.</li> <li>l. An error message is displayed “Invalid Hematocrit Level, must enter numbers only”.</li> <li>m. An error message is displayed “Invalid Potassium, must enter numbers only”.</li> <li>n. An error message is displayed “Invalid Lactic Acid Dehydrogenase, must enter numbers only”.</li> <li>o. An error message is displayed “Invalid MCH, must enter numbers only”.</li> <li>p. An error message is displayed “Invalid MPV, must enter numbers only”.</li> <li>q. An error message is displayed “Invalid MCHC, must enter numbers only”.</li> <li>r. An error message is displayed “Invalid MCV, must enter numbers only”.</li> </ol> </li> </ol>

	<ul style="list-style-type: none"> <li>s. An error message is displayed “Invalid Sodium, must enter numbers only”.</li> <li>t. An error message is displayed “Invalid Platelet, must enter numbers only”.</li> <li>u. An error message is displayed “Invalid Red Blood Cell count, must enter numbers only”.</li> <li>v. An error message is displayed “Invalid RDW, must enter numbers only”.</li> <li>w. An error message is displayed “Invalid SGOT, must enter numbers only”.</li> <li>x. An error message is displayed “Invalid SGPT, must enter numbers only”.</li> <li>y. An error message is displayed “Invalid Total Bilirubin, must enter numbers only”.</li> <li>z. An error message is displayed “Invalid Total Protein, must enter numbers only”.</li> </ul>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 132 Diagnose Rheumatoid Arthritis Disease Test Case

#### 7.3.2.1.24 Diagnose Hypothyroidism Disease

Table 133 presents the test cases of ‘Diagnosis Hypothyroidism Disease’ functionality.

Test ID	Diagnose_Hypothyroidism_Disease
Prerequisite	The user logged into the system.
Test Procedure	<ol style="list-style-type: none"> <li>1. Press the “Diagnose” button after performing all the following:           <ol style="list-style-type: none"> <li>a. Enter a valid Age.</li> <li>b. Enter a valid BP - Systolic test result.</li> <li>c. Enter a valid Respiratory Rate test result.</li> <li>d. Enter a valid MCV test result.</li> <li>e. Enter a valid Pulse Ox test result.</li> </ol> </li> <li>2. Press the “Diagnose” button while leaving some or all fields empty.</li> </ol>
Expected Results	<ol style="list-style-type: none"> <li>1. Successfully show the diagnose result.</li> <li>2. Show an error message if all the fields are empty “Enter all required fields”. If some is missing, then each of the following will generate an error message:           <ol style="list-style-type: none"> <li>a. An error message is displayed “Invalid Age, must enter numbers only”.</li> <li>b. An error message is displayed “Invalid BP – Systolic, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid Respiratory Rate, must enter numbers only”.</li> </ol> </li> </ol>

	<p>d. An error message is displayed “Invalid MCV, must enter numbers only”.</p> <p>e. An error message is displayed “Invalid Pulse Ox, must enter numbers only”.</p>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

*Table 133 Diagnose Hypothyroidism Disease Test Case*

#### 7.3.2.1.25 Diagnose Prostate Cancer Disease

Table 134 presents the test cases of ‘Diagnosis Prostate Cancer Disease’ functionality.

Test ID	<b>Diagnose_Prostate_Cancer_Disease</b>
Prerequisite	The user logged into the system.
Test Procedure	<ol style="list-style-type: none"> <li>1. Press the “Diagnose” button after performing all the following:             <ol style="list-style-type: none"> <li>a. Enter a valid Perimeter test result.</li> <li>b. Enter a valid Area test result.</li> <li>c. Enter a valid Smoothness test result.</li> <li>d. Enter a valid Compactness test result.</li> </ol> </li> <li>2. Press the “Diagnose” button with leaving some or all fields empty.</li> </ol>
Expected Results	<ol style="list-style-type: none"> <li>1. Successfully show the diagnose result.</li> <li>2. Show an error message if all the fields are empty “Enter all required fields”. If some is missing, then each of the following will generate an error message:             <ol style="list-style-type: none"> <li>a. An error message is displayed “Invalid Perimeter, must enter numbers only”.</li> <li>b. An error message is displayed “Invalid Area, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid Smoothness, must enter numbers only”.</li> <li>d. An error message is displayed “Invalid Compactness, must enter numbers only”.</li> </ol> </li> </ol>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

*Table 134 Diagnose Prostate Cancer Disease Test Case*

#### 7.3.2.1.26 Diagnose Cervical Cancer Disease

Table 135 presents the test cases of ‘Diagnosis Cervical Cancer Disease’ functionality.

Test ID	<b>Diagnose_Cervical_Cancer_Disease</b>
Prerequisite	The user logged into the system.

Test Procedure	<ol style="list-style-type: none"> <li>1. Press the “Diagnose” button after performing all the following:             <ol style="list-style-type: none"> <li>a. Enter a valid STDs: Number of Diagnosis.</li> <li>b. Select STDs Condylomatosis status from the drop-down list.</li> <li>c. Select STDs Syphilis status from the drop-down list.</li> <li>d. Select STDs HIV status from the drop-down list.</li> <li>e. Select STDs HPV status from the drop-down list.</li> <li>f. Select Dx status from the drop-down list.</li> <li>g. Select Dx: CIN status from the drop-down list.</li> <li>h. Select Dx: HPV status from the drop-down list.</li> </ol> </li> <li>2. Press the “Diagnose” button while leaving some or all fields empty.</li> </ol>
Expected Results	<ol style="list-style-type: none"> <li>1. Successfully show the diagnose result.</li> <li>2. Show an error message if all the fields are empty “Enter all required fields”. If some is missing, then each of the following will generate an error message:             <ol style="list-style-type: none"> <li>a. An error message is displayed “Invalid STDs: Number of Diagnosis, must enter numbers only”.</li> <li>b. An error message is displayed “The STDs Condylomatosis is empty, choose from the drop-down list”.</li> <li>c. An error message is displayed “The STDs Syphilis is empty, choose from the drop-down list”.</li> <li>d. An error message is displayed “The STDs HIV is empty, choose from the drop-down list”.</li> <li>e. An error message is displayed “The STDs HPV is empty, choose from the drop-down list”.</li> <li>f. An error message is displayed “The Dx is empty, choose from the drop-down list”.</li> <li>g. An error message is displayed “The Dx: CIN is empty, choose from the drop-down list”.</li> <li>h. An error message is displayed “The Dx: HPV is empty, choose from the drop-down list”.</li> </ol> </li> </ol>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 135 Diagnose Cervical Cancer Disease Test Case

### 7.3.2.1.27 Diagnose Multiple Sclerosis Disease

Table 136 presents the test cases of ‘Diagnosis Multiple Sclerosis Disease’ functionality.

Test ID	Diagnose_Multiple_Sclerosis_Disease
Prerequisite	The user logged into the system.

Test Procedure	<ol style="list-style-type: none"> <li>1. Press the “Diagnose” button after performing all the following:             <ol style="list-style-type: none"> <li>a. Enter a valid Age.</li> <li>b. Enter a valid Alanine Transaminase (ALT) test result.</li> <li>c. Enter a valid Lactate Dehydrogenase (LDH) test result.</li> <li>d. Enter a valid Creatinine test result.</li> <li>e. Enter a valid Blood Urea Nitrogen test result.</li> <li>f. Enter a valid Total Bilirubin test result.</li> <li>g. Enter a valid GGT test result.</li> <li>h. Enter a valid Alkaline Phosphatase test result.</li> <li>i. Enter a valid AST test result.</li> <li>j. Enter a valid Platelet test result.</li> <li>k. Enter a valid BP – Systolic test result.</li> </ol> </li> <li>2. Press the “Diagnose” button while leaving some or all fields empty.</li> </ol>
Expected Results	<ol style="list-style-type: none"> <li>1. Successfully show the diagnose result.</li> <li>2. Show an error message if all the fields are empty “Enter all required fields”. If some is missing, then each of the following will generate an error message:             <ol style="list-style-type: none"> <li>a. An error message is displayed “Invalid Age, must enter numbers only”.</li> <li>b. An error message is displayed “Invalid ALT, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid LDH, must enter numbers only”.</li> <li>d. An error message is displayed “Invalid Creatinine, must enter numbers only”.</li> <li>e. An error message is displayed “Invalid Blood Urea Nitrogen, must enter numbers only”.</li> <li>f. An error message is displayed “Invalid Total Bilirubin, must enter numbers only”.</li> <li>g. An error message is displayed “Invalid GGT, must enter numbers only”.</li> <li>h. An error message is displayed “Invalid Alkaline Phosphatase, must enter numbers only”.</li> <li>i. An error message is displayed “Invalid AST, must enter numbers only”.</li> <li>j. An error message is displayed “Invalid Platelet, must enter numbers only”.</li> <li>k. An error message is displayed “Invalid BP - Systolic, must enter numbers only”.</li> </ol> </li> </ol>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

*Table 136 Diagnose Multiple Sclerosis Disease Test Case*

### 7.3.2.1.28 Diagnose Liver Cirrhosis Disease

Table 137 presents the test cases of ‘Diagnosis Liver Cirrhosis Disease’ functionality.

Test ID	Diagnose_Liver_Cirrhosis_Disease
Prerequisite	The user logged into the system.
Test Procedure	<ol style="list-style-type: none"> <li>1. Press the “Diagnose” button after performing all the following:             <ol style="list-style-type: none"> <li>a. Select Gender from the drop-down list.</li> <li>b. Enter a valid Age.</li> <li>c. Enter a valid N_Days.</li> <li>d. Select Hepatomegaly status from the drop-down list.</li> <li>e. Select Spiders status from the drop-down list.</li> <li>f. Enter a valid Edema test result.</li> <li>g. Enter a valid Cholesterol test result.</li> <li>h. Enter a valid Copper test result.</li> <li>i. Enter a valid SGOT test result.</li> <li>j. Enter a valid Platelet test result.</li> <li>k. Enter a valid Prothrombin test result.</li> <li>l. Select Ascites status from the drop-down list.</li> <li>m. Enter a valid Serum Bilirubin test result.</li> <li>n. Enter a valid Albumin test result.</li> <li>o. Enter a valid Alkaline Phosphatase test result.</li> <li>p. Enter a valid Triglycerides test result.</li> <li>q. Enter a valid Drug test result.</li> <li>r. Select status from the drop-down list.</li> </ol> </li> <li>2. Press the “Diagnose” button while leaving some or all fields empty.</li> </ol>
Expected Results	<ol style="list-style-type: none"> <li>1. Successfully show the diagnose result.</li> <li>2. Show an error message if all the fields are empty “Enter all required fields”. If some is missing, then each of the following will generate an error message:             <ol style="list-style-type: none"> <li>a. An error message is displayed “The Gender is empty, choose from the drop-down list”.</li> <li>b. An error message is displayed “Invalid Age, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid N_Days, must enter numbers only”.</li> <li>d. An error message is displayed “The Hepatomegaly is empty, choose from the drop-down list”.</li> <li>e. An error message is displayed “The Spiders is empty, choose from the drop-down list”.</li> <li>f. An error message is displayed “Invalid Edema, must enter numbers only”.</li> </ol> </li> </ol>

	<ul style="list-style-type: none"> <li>g. An error message is displayed “Invalid Cholesterol, must enter numbers only”.</li> <li>h. An error message is displayed “Invalid Copper, must enter numbers only”.</li> <li>i. An error message is displayed “Invalid SGOT, must enter numbers only”.</li> <li>j. An error message is displayed “Invalid Platelet, must enter numbers only”.</li> <li>k. An error message is displayed “Invalid Prothrombin, must enter numbers only”.</li> <li>l. An error message is displayed “The Ascites is empty, choose from the drop-down list”.</li> <li>m. An error message is displayed “Invalid Serum Bilirubin, must enter numbers only”.</li> <li>n. An error message is displayed “Invalid Albumin, must enter numbers only”.</li> <li>o. An error message is displayed “Invalid Alkaline Phosphatase, must enter numbers only”.</li> <li>p. An error message is displayed “Invalid Triglycerides, must enter numbers only”.</li> <li>q. An error message is displayed “Invalid Drug, must enter numbers only”.</li> <li>r. An error message is displayed “The Status is empty, choose from the drop-down list”.</li> </ul>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 137 Diagnose Liver Cirrhosis Disease Test Case

#### 7.3.2.1.29 Diagnose Chronic Obstructive Pulmonary Disease

Table 138 presents the test cases of ‘Diagnosis Chronic Obstructive Pulmonary Disease’ functionality.

Test ID	Diagnose_Chronic_Obstructive_Pulmonary_Disease
Prerequisite	The user logged into the system.
Test Procedure	<p>1. Press the “Diagnose” button after performing all the following:</p> <ul style="list-style-type: none"> <li>a. Select Gender status from the drop-down list.</li> <li>b. Enter a valid Age.</li> <li>c. Enter a valid Smoking test result.</li> <li>d. Enter a valid Imagery part minimum test result.</li> <li>e. Enter a valid Imagery part average test result.</li> <li>f. Enter a valid Real part minimum test result.</li> <li>g. Enter a valid Real part average test result.</li> </ul> <p>2. Press the “Diagnose” button while leaving some or all fields empty.</p>

Expected Results	<ol style="list-style-type: none"> <li>1. Successfully show the diagnose result.</li> <li>2. Show an error message if all the fields are empty “Enter all required fields”. If some is missing, then each of the following will generate an error message:           <ol style="list-style-type: none"> <li>a. An error message is displayed “The Gender is empty, choose from the drop-down list”.</li> <li>b. An error message is displayed “Invalid Age, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid Smoking, must enter numbers only”.</li> <li>d. An error message is displayed “Invalid Imagery part minimum, must enter numbers only”.</li> <li>e. An error message is displayed “Invalid Imagery part average, must enter numbers only”.</li> <li>f. An error message is displayed “Invalid Real part minimum, must enter numbers only”.</li> <li>g. An error message is displayed “Invalid Real part average, must enter numbers only”.</li> </ol> </li> </ol>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 138 Diagnose Chronic Obstructive Pulmonary Disease Test Case

#### 7.3.2.1.30 Diagnose Parkinson’s Disease

Table 139 presents the test cases of ‘Diagnosis Parkinson’s Disease’ functionality.

Test ID	Diagnose_Parkinson’s_Disease
Prerequisite	The user logged into the system.
Test Procedure	<p>1. Press the “Diagnose” button after performing all the following:</p> <ol style="list-style-type: none"> <li>a. Select Gender status from the drop-down list.</li> <li>b. Enter a valid Age.</li> <li>c. Enter a valid Anion Gap test result.</li> <li>d. Enter a valid ALT test result.</li> <li>e. Enter a valid LDH test result.</li> <li>f. Enter a valid White Blood Cells test result.</li> <li>g. Enter a valid Red Blood Cells test result.</li> <li>h. Enter a valid Hemoglobin test result.</li> <li>i. Enter a valid Hematocrit test result.</li> <li>j. Enter a valid Sodium test result.</li> <li>k. Enter a valid Potassium test result.</li> <li>l. Enter a valid Chloride test result.</li> <li>m. Enter a valid Carbon Dioxide test result.</li> <li>n. Enter a valid Creatinine test result.</li> <li>o. Enter a valid Total Protein test result.</li> </ol>

	<p>p. Enter a valid Albumin test result.</p> <p>q. Enter a valid Blood Urea Nitrogen test result.</p> <p>r. Enter a valid Total Bilirubin test result.</p> <p>s. Enter a valid Direct Bilirubin test result.</p> <p>t. Enter a valid GGT test result.</p> <p>u. Enter a valid MCV test result.</p> <p>v. Enter a valid MCH test result.</p> <p>w. Enter a valid MCHC test result.</p> <p>x. Enter a valid Alkaline Phosphatase test result.</p> <p>y. Enter a valid RDW test result.</p> <p>z. Enter a valid AST test result.</p> <p>2. Press the “Diagnose” button while leaving some or all fields empty.</p>
Expected Results	<p>1. Successfully show the diagnose result.</p> <p>2. Show an error message if all the fields are empty “Enter all required fields”. If some is missing, then each of the following will generate an error message:</p> <ul style="list-style-type: none"> <li>a. An error message is displayed “The Gender is empty, choose from the drop-down list”.</li> <li>b. An error message is displayed “Invalid Age, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid Anion Gap, must enter numbers only”.</li> <li>d. An error message is displayed “Invalid ALT, must enter numbers only”.</li> <li>e. An error message is displayed “Invalid LDH, must enter numbers only”.</li> <li>f. An error message is displayed “Invalid White Blood Cells, must enter numbers only”.</li> <li>g. An error message is displayed “Invalid Red Blood Cells, must enter numbers only”.</li> <li>h. An error message is displayed “Invalid Hemoglobin, must enter numbers only”.</li> <li>i. An error message is displayed “Invalid Hematocrit, must enter numbers only”.</li> <li>j. An error message is displayed “Invalid Sodium, must enter numbers only”.</li> <li>k. An error message is displayed “Invalid Potassium, must enter numbers only”.</li> <li>l. An error message is displayed “Invalid Chloride, must enter numbers only”.</li> <li>m. An error message is displayed “Invalid Carbon Dioxide, must enter numbers only”.</li> </ul>

	<ul style="list-style-type: none"> <li>n. An error message is displayed “Invalid Creatinine, must enter numbers only”.</li> <li>o. An error message is displayed “Invalid Total Protein, must enter numbers only”.</li> <li>p. An error message is displayed “Invalid Albumin, must enter numbers only”.</li> <li>q. An error message is displayed “Invalid Blood Urea Nitrogen, must enter numbers only”.</li> <li>r. An error message is displayed “Invalid Total Bilirubin, must enter numbers only”.</li> <li>s. An error message is displayed “Invalid Direct Bilirubin, must enter numbers only”.</li> <li>t. An error message is displayed “Invalid GGT, must enter numbers only”.</li> <li>u. An error message is displayed “Invalid MCV, must enter numbers only”.</li> <li>v. An error message is displayed “Invalid MCH, must enter numbers only”.</li> <li>w. An error message is displayed “Invalid MCHC, must enter numbers only”.</li> <li>x. An error message is displayed “Invalid Alkaline Phosphatase, must enter numbers only”.</li> <li>y. An error message is displayed “Invalid RDW, must enter numbers only”.</li> <li>z. An error message is displayed “Invalid AST, must enter numbers only”.</li> </ul>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 139 Diagnose Parkinson's Disease Test Case

#### 7.3.2.1.31 Diagnose Hepatitis C Disease

Table 140 presents the test cases of ‘Diagnosis Hepatitis C Disease’ functionality.

Test ID	Diagnose_Hepatitis_C_Disease
Prerequisite	The user logged into the system.
Test Procedure	<ol style="list-style-type: none"> <li>1. Press the “Diagnose” button after performing all the following:           <ol style="list-style-type: none"> <li>a. Enter a valid Age.</li> <li>b. Enter a valid Total Protein test result.</li> <li>c. Enter a valid Total Bilirubin test result.</li> <li>d. Enter a valid Direct Bilirubin test result.</li> <li>e. Enter a valid GGT test result.</li> <li>f. Enter a valid Alkaline Phosphatase test result.</li> <li>g. Enter a valid Lymphocyte - Instrument % test result.</li> </ol> </li> </ol>

	<p>h. Enter a valid Neutrophil Granulocyte - Instrument Absolute test result.</p> <p>i. Enter a valid Platelet test result.</p> <p>j. Enter a valid Basophil - Instrument % test result.</p> <p>k. Enter a valid BP – Systolic test result.</p> <p>l. Enter a valid Fall Risk - Morse test result.</p> <p>m. Enter a valid Body Mass Index test result.</p> <p>n. Enter a valid International Normalized Ratio test result.</p> <p>2. Press the “Diagnose” button with leaving some or all fields empty.</p>
Expected Results	<p>1. Successfully show the diagnose result.</p> <p>2. Show an error message if all the fields are empty “Enter all required fields”. If some is missing, then each of the following will generate an error message:</p> <ul style="list-style-type: none"> <li>a. An error message is displayed “Invalid Age, must enter numbers only”.</li> <li>b. An error message is displayed “Invalid Total Protein, must enter numbers only”.</li> <li>c. An error message is displayed “Invalid Total Bilirubin, must enter numbers only”.</li> <li>d. An error message is displayed “Invalid Direct Bilirubin, must enter numbers only”.</li> <li>e. An error message is displayed “Invalid GGT, must enter numbers only”.</li> <li>f. An error message is displayed “Invalid Alkaline Phosphatase, must enter numbers only”.</li> <li>g. An error message is displayed “Invalid Lymphocyte - Instrument %, must enter numbers only”.</li> <li>h. An error message is displayed “Invalid Neutrophil Granulocyte - Instrument Absolute, must enter numbers only”.</li> <li>i. An error message is displayed “Invalid Platelet, must enter numbers only”.</li> <li>j. An error message is displayed “Invalid Basophil - Instrument %, must enter numbers only”.</li> <li>k. An error message is displayed “Invalid BP – Systolic, must enter numbers only”.</li> <li>l. An error message is displayed “Invalid Fall Risk - Morse, must enter numbers only”.</li> <li>m. An error message is displayed “Invalid Body Mass Index, must enter numbers only”.</li> <li>n. An error message is displayed “Invalid International Normalized Ratio, must enter numbers only”.</li> </ul>
Actual Result	Matching the expected result.

Verified (Yes\No)	Yes
----------------------	-----

Table 140 Diagnose Hepatitis C Disease Test Case

### 7.3.2.1.32 Diagnose Depression Disease

Table 141 presents the test cases of ‘Diagnosis Depression Disease’ functionality.

Test ID	Diagnose_Depression_Disease
Prerequisite	The user logged into the system.
Test Procedure	<p>1. Press the “Diagnose” button after performing all the following:</p> <ul style="list-style-type: none"> <li>a. Enter a valid Age.</li> <li>b. Enter a valid Household Size.</li> <li>c. Enter a valid Education Level.</li> <li>d. Enter a valid Value of livestock.</li> <li>e. Enter a valid Value of durable goods.</li> <li>f. Enter a valid Value of savings.</li> <li>g. Enter a valid number of Land owned.</li> <li>h. Enter a valid level of Consumed Alcohol.</li> <li>i. Enter a valid level of Consumed Tobacco.</li> <li>j. Enter a valid Education expenditure.</li> <li>k. Enter a valid non-ag business flow expense, monthly.</li> <li>l. Enter a valid Livestock sales and meat revenue, monthly.</li> <li>m. Enter a valid Total expense, monthly.</li> <li>n. Enter a valid Whole day without food.</li> <li>o. Enter a valid Non-durable Investments.</li> <li>p. Enter a valid Amount received using M-Pesa.</li> <li>q. Select Marital status from the drop-down list.</li> <li>r. Enter a valid number of Children.</li> <li>s. Enter a valid hh_children.</li> <li>t. Select Non-agricultural business owner status from the drop-down list.</li> <li>u. Select Early Survey status from the drop-down list.</li> <li>v. Select Saved money using M-Pesa status from the drop-down list.</li> </ul> <p>2. Press the “Diagnose” button while leaving some or all fields empty.</p>
Expected Results	<p>1. Successfully show the diagnose result.</p> <p>2. Show an error message if all the fields are empty “Enter all required fields”. If some is missing, then each of the following will generate an error message:</p> <ul style="list-style-type: none"> <li>a. An error message is displayed “Invalid Age, must enter numbers only”.</li> <li>b. An error message is displayed “Invalid Household Size, must enter numbers only”.</li> </ul>

	<ul style="list-style-type: none"> <li>c. An error message is displayed “Invalid Education Level, must enter numbers only”.</li> <li>d. An error message is displayed “Invalid Value of livestock, must enter numbers only”.</li> <li>e. An error message is displayed “Invalid Value of durable goods, must enter numbers only”.</li> <li>f. An error message is displayed “Invalid Value of savings, must enter numbers only”.</li> <li>g. An error message is displayed “Invalid Land owned, must enter numbers only”.</li> <li>h. An error message is displayed “Invalid Consumed Alcohol, must enter numbers only”.</li> <li>i. An error message is displayed “Invalid Consumed Tobacco, must enter numbers only”.</li> <li>j. An error message is displayed “Invalid Education expenditure, must enter numbers only”.</li> <li>k. An error message is displayed “Invalid Non-ag business flow expenses, monthly, must enter numbers only”.</li> <li>l. An error message is displayed “Invalid Livestock sales and meat revenue, monthly, must enter numbers only”.</li> <li>m. An error message is displayed “Invalid Total expenses, monthly, must enter numbers only”.</li> <li>n. An error message is displayed “Invalid Whole days without food, must enter numbers only”.</li> <li>o. An error message is displayed “Invalid Non-durable Investments, must enter numbers only”.</li> <li>p. An error message is displayed “Invalid Amount received using M-Pesa, must enter numbers only”.</li> <li>q. An error message is displayed “The Marital status is empty, choose from the drop-down list”.</li> <li>r. An error message is displayed “Invalid number of Children, must enter numbers only”.</li> <li>s. An error message is displayed “Invalid number of hh_children, must enter numbers only”.</li> <li>t. An error message is displayed “The Non-agricultural business owner is empty, choose from the drop-down list”.</li> <li>u. An error message is displayed “The Saved money using M-Pesa is empty, choose from the drop-down list”.</li> <li>v. An error message is displayed “The Early Survey is empty, choose from the drop-down list”.</li> </ul>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

*Table 141 Diagnose Depression Disease Test Case*

### 7.3.2.1.33 Diagnose Epileptic Seizure Disease

Table 142 presents the test cases of ‘Diagnosis Epileptic Seizure Disease’ functionality.

Test ID	Diagnose_Epileptic_Seizure_Disease
Prerequisite	The user logged into the system.
Test Procedure	<p>1. Press the “Diagnose” button after performing all the following:</p> <ul style="list-style-type: none"> <li>a. Select Gender from the drop-down list.</li> <li>b. Enter a valid Age.</li> <li>c. Enter a valid Basophil Instrument % test result.</li> <li>d. Enter a valid Hematocrit test result.</li> <li>e. Enter a valid Hemoglobin test result.</li> <li>f. Enter a valid MCH test result.</li> <li>g. Enter a valid MCHC test result.</li> <li>h. Enter a valid MPV test result.</li> <li>i. Enter a valid White Blood Cell count.</li> <li>j. Select the number of non-psychiatric comorbidities from the drop-down list.</li> <li>k. Enter the number of prior AEDs (Anti-Epileptic Drugs) as an integer.</li> <li>l. Select the presence of Asthma from the drop-down list.</li> <li>m. Select the presence of Migraine from the drop-down list.</li> <li>n. Select the presence of Chronic Pain from the drop-down list.</li> <li>o. Select the presence of Diabetes from the drop-down list.</li> <li>p. Select the presence of non-metastatic cancer from the drop-down list.</li> <li>q. Enter the number of non-seizure, non-psychiatric medications as an integer.</li> <li>r. Enter the number of current AEDs as an integer.</li> <li>s. Enter the baseline as a decimal value.</li> <li>t. Enter the median duration of seizures as a decimal value.</li> <li>u. Enter the number of seizure types as an integer.</li> <li>v. Select if there's an injury with seizure from the drop-down list.</li> <li>w. Select if it's Catamenial from the drop-down list.</li> <li>x. Select the trigger of sleep deprivation from the drop-down list.</li> <li>y. Select if there's an Aura from the drop-down list.</li> <li>z. Select if there are Ictal Eye Closures from the drop-down list.</li> <li>aa. Select if there are Ictal Hallucinations from the drop-down list.</li> <li>bb. Select if there are Oral automatisms from the drop-down list.</li> <li>cc. Select if there's Incontinence from the drop-down list.</li> <li>dd. Select if there are Limb Automatisms from the drop-down list.</li> <li>ee. Select if there are Ictal Tonic-clonic from the drop-down list.</li> <li>ff. Select if there are Muscle Twitching from the drop-down list.</li> <li>gg. Select if there's Hip Thrusting from the drop-down list.</li> </ul>

	<p>hh. Select if there's Post-ictal Fatigue from the drop-down list.</p> <p>ii. Select if there's a Head Injury from the drop-down list.</p> <p>jj. Select if there are Psychological Traumatic Events from the drop-down list.</p> <p>kk. Select if there's Concussion without Loss of Consciousness from the drop-down list.</p> <p>ll. Select if there's Concussion with Loss of Consciousness from the drop-down list.</p> <p>mm. Select if there's Severe Traumatic Brain Injury from the drop-down list.</p> <p>nn. Select if there's Opioid usage from the drop-down list.</p> <p>oo. Select if there's a history of Sexual Abuse from the drop-down list.</p> <p>pp. Select if there's a history of Physical Abuse from the drop-down list.</p> <p>qq. Select if there's a history of Rape from the drop-down list.</p> <p>2. Press the “Diagnose” button while leaving some or all fields empty.</p>
Expected Results	<p>1. Successfully show the diagnose result.</p> <p>2. Show an error message if all the fields are empty “Enter all required fields”.</p> <p>a. An error message is displayed “The Gender is empty, choose from the drop-down list”.</p> <p>b. An error message is displayed “Invalid Age, must enter numbers only”.</p> <p>c. An error message is displayed “Invalid Basophil Instrument % test result, must enter numbers only”.</p> <p>d. An error message is displayed “Invalid Hematocrit test result, must enter numbers only”.</p> <p>e. An error message is displayed “Invalid Hemoglobin test result, must enter numbers only”.</p> <p>f. An error message is displayed “Invalid MCH test result, must enter numbers only”.</p> <p>g. An error message is displayed “Invalid MCHC test result, must enter numbers only”.</p> <p>h. An error message is displayed “Invalid MPV test result, must enter numbers only”.</p> <p>i. An error message is displayed “Invalid White Blood Cell count, must enter numbers only”.</p> <p>j. An error message is displayed "Please select the number of non-psychiatric comorbidities".</p> <p>k. An error message is displayed "Please enter the number of prior AEDs as a whole number".</p> <p>l. An error message is displayed "Please select the presence of Asthma from the drop-down list".</p>

- m. An error message is displayed "Please select the presence of Migraine from the drop-down list".
- n. An error message is displayed "Please select the presence of Chronic Pain from the drop-down list".
- o. An error message is displayed "Please select the presence of Diabetes from the drop-down list".
- p. An error message is displayed "Please select the presence of non-metastatic cancer from the drop-down list".
- q. An error message is displayed "Please enter the number of non-seizure, non-psychiatric medications as a whole number".
- r. An error message is displayed "Please enter the number of current AEDs as a whole number".
- s. An error message is displayed "Please enter the baseline as a decimal value".
- t. An error message is displayed "Please enter the median duration of seizures as a decimal value".
- u. An error message is displayed "Please enter the number of seizure types as a whole number".
- v. An error message is displayed "Please select if there's an injury with seizure from the drop-down list".
- w. An error message is displayed "Please select if it's Catamenial from the drop-down list".
- x. An error message is displayed "Please select the trigger of sleep deprivation from the drop-down list".
- y. An error message is displayed "Please select if there's an Aura from the drop-down list".
- z. An error message is displayed "Please select if there are Ictal Eye Closures from the drop-down list".
- aa. An error message is displayed "Please select if there are Ictal Hallucinations from the drop-down list".
- bb. An error message is displayed "Please select if there are Oral automatisms from the drop-down list".
- cc. An error message is displayed "Please select if there's Incontinence from the drop-down list".
- dd. An error message is displayed "Please select if there are Limb Automatisms from the drop-down list".
- ee. An error message is displayed "Please select if there are Ictal Tonic-clonic from the drop-down list".
- ff. An error message is displayed "Please select if there are Muscle Twitching from the drop-down list".
- gg. An error message is displayed "Please select if there's Hip Thrusting from the drop-down list".
- hh. An error message is displayed "Please select if there's Post-ictal

Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 142 Diagnose Epileptic Seizure Disease Test Case

### 7.3.2.1.34 Diagnose Osteoporosis Disease

Table 143 presents the test cases of ‘Diagnosis Osteoporosis Disease’ functionality.

Test ID	Diagnose_Osteoporosis_Disease
Prerequisite	The user logged into the system.
Test Procedure	<p>1. Press the “Diagnose” button after performing all the following:</p> <ul style="list-style-type: none"> <li>a. Select Gender from the drop-down list.</li> <li>b. Enter a valid Age.</li> <li>c. Enter a valid Basophil Instrument % test result.</li> <li>d. Enter a valid Hematocrit test result.</li> <li>e. Enter a valid Hemoglobin test result.</li> <li>f. Enter a valid MCH test result.</li> <li>g. Enter a valid MCHC test result.</li> <li>h. Enter a valid MPV test result.</li> <li>i. Enter a valid White Blood Cell count.</li> <li>j. Select the presence of Diabetes from the drop-down list.</li> <li>k. Select the presence of Hypothyroidism from the drop-down list.</li> <li>l. Select the presence of Seizure Disorder from the drop-down list.</li> <li>m. Select the consumption of Alcohol from the drop-down list.</li> <li>n. Select the smoking status from the drop-down list.</li> <li>o. Select if there's estrogen use from the drop-down list.</li> <li>p. Select the presence of Joint Pain from the drop-down list.</li> <li>q. Select if there's a history of Fracture from the drop-down list.</li> <li>r. Select if the patient is on Dialysis from the drop-down list.</li> <li>s. Select if there's a family history of Osteoporosis from the drop-down list.</li> <li>t. Enter the maximum walking distance as an integer.</li> <li>u. Select the daily eating habits from the drop-down list.</li> <li>v. Enter the BMI (Body Mass Index) as a decimal value.</li> <li>w. Select the site of testing from the drop-down list.</li> <li>x. Select if there's Obesity from the drop-down list.</li> <li>y. Enter the patient's weight as an integer.</li> <li>z. Enter the patient's height as an integer.</li> </ul> <p>2. Press the “Diagnose” button while leaving some or all fields empty.</p>
Expected Results	<ol style="list-style-type: none"> <li>1. Successfully show the diagnose result.</li> <li>2. Show an error message if all the fields are empty “Enter all required fields”.</li> </ol>

- a. An error message is displayed "The Gender is empty, choose from the drop-down list".
- b. An error message is displayed "Invalid Age, must enter numbers only".
- c. An error message is displayed "Invalid Basophil Instrument % test result, must enter numbers only".
- d. An error message is displayed "Invalid Hematocrit test result, must enter numbers only".
- e. An error message is displayed "Invalid Hemoglobin test result, must enter numbers only".
- f. An error message is displayed "Invalid MCH test result, must enter numbers only".
- g. An error message is displayed "Invalid MCHC test result, must enter numbers only".
- h. An error message is displayed "Invalid MPV test result, must enter numbers only".
- i. An error message is displayed "Invalid White Blood Cell count, must enter numbers only".
- j. An error message is displayed "Please select the presence of Diabetes from the drop-down list".
- k. An error message is displayed "Please select the presence of Hypothyroidism from the drop-down list".
- l. An error message is displayed "Please select the presence of Seizure Disorder from the drop-down list".
- m. An error message is displayed "Please select the consumption of Alcohol from the drop-down list".
- n. An error message is displayed "Please select the smoking status from the drop-down list".
- o. An error message is displayed "Please select if there's estrogen use from the drop-down list".
- p. An error message is displayed "Please select the presence of Joint Pain from the drop-down list".
- q. An error message is displayed "Please select if there's a history of Fracture from the drop-down list".
- r. An error message is displayed "Please select if the patient is on Dialysis from the drop-down list".
- s. An error message is displayed "Please select if there's a family history of Osteoporosis from the drop-down list".
- t. An error message is displayed "Please enter the maximum walking distance as a whole number".
- u. An error message is displayed "Please select the daily eating habits from the drop-down list".

	<p>v. An error message is displayed "Please enter the BMI as a decimal value".</p> <p>w. An error message is displayed "Please select the site from the drop-down list".</p> <p>x. An error message is displayed "Please select if there's Obesity from the drop-down list".</p> <p>y. An error message is displayed "Please enter the patient's weight as a whole number".</p> <p>z. An error message is displayed "Please enter the patient's height as a whole number".</p>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 143 Diagnose Osteoporosis Disease Test Case

#### 7.3.2.1.35 Diagnosis Sickle Cell Anemia Disease

Table 144 presents the test cases of ‘Diagnosis Sickle Cell Anemia Disease’ functionality.

Test ID	Diagnose_Sickle Cell Anemia _Disease
Prerequisite	The user logged into the system.
Test Procedure	<p>1. Press the “Diagnose” button after performing all the following:</p> <ul style="list-style-type: none"> <li>a. Select Gender from the drop-down list.</li> <li>b. Enter a valid Age.</li> <li>c. Enter a valid Tribe.</li> <li>d. Enter a valid Hemoglobin (HB) test result.</li> <li>e. Enter a valid Packed cell volume (PCV) test result.</li> <li>f. Enter a valid Red Blood Cells (RBCs) test result.</li> <li>g. Enter a valid Mean Cell Volume (MCV) test result.</li> <li>h. Enter a valid Mean Cell Hemoglobin (MCH) test result.</li> <li>i. Enter a valid Mean Cell Hemoglobin Concentration (MCHC) test result.</li> <li>j. Enter a valid Total White Blood Cells (TWBCs) count.</li> <li>k. Enter a valid Platelet Counts (PLTs) count.</li> </ul> <p>2. Press the “Diagnose” button while leaving some or all fields empty.</p>
Expected Results	<p>1. Successfully show the diagnostic result.</p> <p>2. Show an error message if all the fields are empty “Enter all required fields”. If some is missing, then each of the following will generate an error message:</p> <ul style="list-style-type: none"> <li>a. An error message is displayed “The Gender is empty, choose from the drop-down list”.</li> <li>b. An error message is displayed “Invalid Age, must enter numbers only”.</li> </ul>

	<p>c. An error message is displayed “Invalid Tribe, must enter numbers only”.</p> <p>d. An error message is displayed “Invalid Hemoglobin (HB) test result, must enter numbers only”.</p> <p>e. An error message is displayed “Invalid Packed cell volume (PCV) test result, must enter numbers only”.</p> <p>f. An error message is displayed “Invalid Red Blood Cells (RBCs) test result, must enter numbers only”.</p> <p>g. An error message is displayed “Invalid Mean Cell Volume (MCV) test result, must enter numbers only”.</p> <p>h. An error message is displayed “Invalid Mean Cell Hemoglobin (MCH) test result, must enter numbers only”.</p> <p>i. An error message is displayed “Invalid Mean Cell Hemoglobin Concentration (MCHC) test result, must enter numbers only”.</p> <p>j. An error message is displayed “Invalid Total White Blood Cells (TWBCs) count, must enter numbers only”.</p> <p>k. An error message is displayed “Invalid Platelet Counts (PLTs) count, must enter numbers only”.</p>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 144 Table 143 Diagnose Sickle Cell Anemia Disease Test Case

### 7.3.2.1.3 View Diagnostic Result

Table 145 presents the test cases of ‘Diagnosis Result’ functionality.

Test ID	Diagnostic_Result_001
Prerequisite	The user logged into the system.
Test Procedure	<ol style="list-style-type: none"> <li>Access the ‘Result’ interface directed from the diagnosis interface.</li> <li>Clicking the print button.</li> </ol>
Expected Results	<ol style="list-style-type: none"> <li>Display correct diagnosis as advised by the model and display the model’s accuracy.</li> <li>Displays the print options and sends a print job to the operating system.</li> </ol>
Actual Result	Matching the expected result.
Verified (Yes\No)	Yes

Table 145 View Diagnostic Result

### 7.3.3 Integration Testing

Integration testing is carried out to verify that the components are properly interacting with one another as the system develops, after component isolation and individual testing.

Technique:

- The test begins with the primary interface and adds components one at a time so that the system's error points may be quickly and easily detected.

Completion criteria:

- Every component of the system functions as it should.

Special Considerations:

- In the event that an error arises during the integration process, the component will undergo testing to address the error.

Test cases confirm the functionality of the various parts beginning with each user's main interface:

- Admin Home Interface.
- Medical Specialist Home Interface.
- Laboratory Specialists Home Interface.
- Registered User Home Interface.
- Guest Home Interface.

### 7.3.3.1 Test Cases

#### 7.3.3.1.1 Admin Home Interface

Table 146 presents the integration between components starting from the admin's home interface.

Test ID	Admin_Home_Interface_001
Prerequisite	Admin is logged into the system.
Test Procedure	<p>Try the following procedures separately:</p> <ol style="list-style-type: none"> <li>Access 'profile' interface and the component testing cases are done.</li> <li>Access 'manage user' interface and the component testing cases are done.</li> <li>Access 'rebuild model' interface and the component testing cases are done.</li> <li>Logout from the system and returned to 'login' interface.</li> </ol>
Expected Result	<ol style="list-style-type: none"> <li>Each of the previous interfaces should be displayed when requested with each function performing properly.</li> <li>The logged in admin's username should be maintained during the session.</li> <li>All confirmation and error messages are displayed correctly.</li> </ol>
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 146 Integration Test Case of 'Admin Home' Interface

#### 7.3.3.1.2 Medical Specialist Home Interface

Table 147 presents the integration between components starting from the medical specialists' home interface.

Test ID	<b>MedicalSpecialist_Home_Interface_001</b>
Prerequisite	Medical Specialist is logged into the system.
Test Procedure	<p>Try the following procedures separately:</p> <ol style="list-style-type: none"> <li>1. Access ‘profile’ interface and the component testing cases are done.</li> <li>2. Access ‘diagnosis history’ interface and the component testing cases are done.</li> <li>3. Access ‘diagnose’ interface, and the component testing cases are done.</li> <li>4. Logout from the system and returned to ‘login’ interface.</li> </ol>
Expected Result	<ol style="list-style-type: none"> <li>1. Each of the previous interfaces should be displayed when requested with each function performing properly.</li> <li>2. The patient’s medical data should flow properly to the ‘diagnose’ interface and the diagnosis results should flow properly to the diagnosis result’ interface.</li> <li>3. The logged in medical specialist’s username should be maintained during the session.</li> <li>4. All confirmation and error messages are displayed correctly.</li> </ol>
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 147 Integration Test Case of ‘Medical Specialist Home’ Interface

### 7.3.3.1.3 Laboratory Specialist Home Interface

Table 148 presents the integration between components starting from the laboratory specialists’ home interface.

Test ID	<b>LaboratorySpecialist_Home_Interface_001</b>
Prerequisite	Laboratory Specialist is logged into the system.
Test Procedure	<p>Try the following procedures separately:</p> <ol style="list-style-type: none"> <li>1. Access ‘profile’ interface and the component testing cases are done.</li> <li>2. Access ‘Laboratory Specialists Easy Diagnosis Interface’ interface and the component testing cases are done.</li> <li>3. Logout from the system and returned to ‘login’ interface.</li> </ol>
Expected Results	<ol style="list-style-type: none"> <li>1. Each of the previous interfaces should be displayed when requested with each function performing properly.</li> <li>2. The logged-in laboratory specialist’s username should be maintained during the session.</li> <li>3. All confirmation and error messages are displayed correctly.</li> </ol>
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 148 Integration Test Case of ‘Laboratory Specialist Home’ Interface

#### 7.3.3.1.4 Registered User Home Interface

Table 149 presents the integration between components starting from the registered user's home interface.

Test ID	RegisteredUser_Home_Interface_001
Prerequisite	Registered user is logged into the system.
Test Procedure	<p>Try the following procedures separately:</p> <ol style="list-style-type: none"> <li>1. Access 'profile' interface and the component testing cases are done.</li> <li>2. Access 'diagnosis history' interface and the component testing cases are done.</li> <li>3. Access 'diagnose' interface and the component testing cases are done.</li> <li>4. Logout from the system and return to 'login' interface.</li> </ol>
Expected Results	<ol style="list-style-type: none"> <li>1. Each of the previous interfaces should be displayed when requested with each function performing properly.</li> <li>2. The registered user's medical data should flow properly to the 'diagnose' interface and the diagnosis results data should flow properly to the 'diagnosis result' interface.</li> <li>3. The logged in registered user's username should be maintained during the session.</li> <li>4. All confirmation and error messages are displayed correctly.</li> </ol>
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 149 Integration Test Case of 'Registered User Home' Interface

#### 7.3.3.1.5 Guest User Home Interface

Table 150 presents the integration between components starting from the guest's home interface.

Test ID	Guest_Home_interface_001
Prerequisite	None.
Test Procedure	<p>Try the following procedures separately:</p> <ol style="list-style-type: none"> <li>1. Access 'diagnose' interface and the component testing cases are done.</li> <li>2. The guest's medical data should flow properly to the 'diagnosis result' interface.</li> </ol>
Expected Result	<ol style="list-style-type: none"> <li>1. Each of the previous interfaces should be displayed when requested with each function performs properly.</li> <li>2. All confirmation and error messages are displayed correctly.</li> </ol>
Actual Result	Matching the expected outcomes.
Verified (Yes/No)	Yes

Table 150 Integration Test Case of 'Guest Home' Interface

#### **7.3.4 User Interface Testing**

User interface (UI) testing is the process of determining whether user interfaces highlight needs and execute appropriately. It also seeks to gauge how satisfied users are with the way the system user interfaces (UIs). It is thus the most important kind of exam.

Technique:

Several of the system's user interface tests are explained below, depending on the particular requirements:

- Consistency and clarity across browsers, devices, and interfaces.
- The components' dimensions and locations.
- The resolution of the zooming in and out.
- Clarity, color, and font size.
- Grammar and spelling errors in labels and communications.
- Clarity, size, and alignment of images.
- When scaling the interfaces, components shouldn't overlap.
- Every hyperlink needs to be accessible.

Completion Criteria:

- UI successfully satisfies the needs of users.

#### **7.3.5 Interface Testing**

Interface testing examines the communication and data flow between external and system components. This is done to ensure that all interactions between these components are as intended, including the right parameters being sent to the called component in the right order, the called component performing as intended, and the calling operation and called components accessing the most recent information at the same speed.

Our system's mail server and database server are its external components. Information about users and patients is stored in the database. To evaluate how the components interacted with the database, a variety of test scenarios were used.

For new users, the system sends their temporary passwords and usernames over the mail server. An email confirmation is sent to complete the information modification. To test how the various components interact with the mail server, a variety of test cases were used.

#### **7.3.6 Validation and Verification Testing**

Validation and Verification (V&V) guarantees that the system is constructed correctly in accordance with the given specifications and that it satisfies the demands of the users. Validation is carried out toward the conclusion of the development process, and verification is frequently done throughout this time.

Technique:

- Verification testing involves looking over the test cases, code, requirements, and design specifications. Validation testing is carried out to ensure that the system satisfies the needs of the users.

Completion Criteria:

- The system was successfully validated and confirmed.

### **7.3.7 Security Testing**

Making sure that only authorized users have access to the webpages under their privileges is part of security testing.

Technique:

- In order to access the website, a working account and password are necessary. An error notice will appear if the password or username are incorrect.

Completion Criteria:

- Only those with permission can access the website through their accounts.

### **7.3.8 Performance Testing**

Performance testing is done to evaluate the fault tolerance, recovery, availability, and response time of the website. It entails load testing, which is used to track how well a system performs under various loads. Since the most time-consuming process is generating the diagnostic model, which takes 1.35 seconds, the response time of the website should not exceed 2 seconds. The website must also be available around-the-clock in order to offer diagnostic services whenever needed. If a user makes a mistake, the suggested website has to confirm their input and provide a helpful message explaining how to fix it.

Technique:

- Increasing the number of concurrent users, functions, and database accesses, which steadily increases the system's demand. The system's behavior is examined under different load conditions.

Completion Criteria:

- Successfully finish the test cases for several users in the allotted time frame and without any errors.

### **7.3.9 Constraints**

Every deadline specified in the SPMP section needs to be adhered to.

### **7.3.10 Beta Testing**

A beta version is made available, and some users are asked to test the system and report any issues.

### **7.3.11 Acceptance Testing**

No publication of the acceptance version is planned.

### **7.3.12 Test report (ABET\* Mandatory)**

The purpose of testing is to ensure that the system works as intended. Various requirements were tested using different approaches. Including interfaces functionality and design. Any issues faced during the testing that refer to problems on the system were solved. Not being able to register on the system or facing some difficulty in diagnosing results was some of the problems faced during the testing stage.

## **7.4 Pass/Fail Testing**

The following conditions should be satisfied by the Easy Diagnosis system:

- Every feature should function in accordance with the SRS document's specifications.
- Every test should pass and include no errors.
- Every test case including essential functionality ought to be successful.

## **7.5 Testing Process**

The testing deliverables, tasks, responsibilities, resources, and timetable are covered in this section.

### **7.5.1 Test Deliverables**

Deliveries of these papers would occur either during or following the testing phase:

- Software Test Plan (STP).
- User manual document.

### **7.5.2 Testing Tasks**

Table 151 summarizes the testing tasks and how they depend on other tasks.

Task #	Task Name	Skills Required	Dependencies
1	Prepare SRS and SDS documents	Writing, designing	-
2	Develop test cases	Testing skills	1
3	Prepare STP document	Writing, Analytical	2
4	Prepare the hardware and software test environment	Technical	3
5	Executing all the test activities	Basic computer skills	4
6	Debugging errors occurred during testing	Programming skills	5
7	Maintain the application	Programming skills	6

Table 151 Testing Tasks

### **7.5.3 Responsibilities**

Each component and integration test case outlined in the document, as well as creating the software test strategy, are the responsibilities of the team members. The team has the responsibility of debugging and fixing any faults that are discovered.

#### 7.5.4 Resources

Table 152 shows all the resources used for website testing.

Resource	Description
Hardware	Windows and Mac Laptops.
Software	<ul style="list-style-type: none"><li>MongoDB.</li><li>Web browsers: Chrome, Firefox, and Safari.</li></ul>
Human	The project team members who are skilled in programming, analysis and managing databases.

Table 152 Software Testing Resources

#### 7.5.5 Schedule

Table 153 shows the schedule for the testing tasks.

Task	Date
Complete the SRS document	4-Nov-2023
Complete the SDS document	16-Nov-2023
Develop all test cases	25-April-2024
Complete the STP document	3-May-2024
Prepare the hardware and software test environment	4-May-2024
Executing all the test activities	6-May-2024
Debug and resolve all errors encountered during testing	16-May-2024

Table 153 Testing Tasks Schedule

#### 7.5.6 Environmental Requirements

This section contains a summary of potential hazards and assumptions as well as the environmental requirements for the server, software, and hardware.

##### 7.5.6.1 Hardware

The following hardware is needed to finish the testing activities:

- Laptops running Mac OS X.
- Windows 10.

##### 7.5.6.2 Software

The following software prerequisites are required to carry out the testing activities:

- Internet browser.
- MongoDB.

##### 7.5.6.3 Server

To participate in the testing activities, the server has to meet certain requirements:

- Fast database query response times.
- Large capacity for storing.

#### **6.5.6.4 Risks and Assumptions**

A coding method change might create a delay in the software development life cycle and impact the test plan. The following is a list of the presumptions and possible hazards.

##### **7.5.6.4.1 Test Item Availability**

There might be a delay if any unit is unavailable. Integration testing may suffer if the unit is unavailable. The test team will test alternative units up until the unavailable unit becomes available, as an assumption.

##### **7.5.6.4.2 Test Resources Availability**

Technical problems with resources, including loading on Database Management Systems (DBMS), may exist. The test team will assume that they can get in touch with the support team and get a speedy remedy.

##### **7.5.6.4.3 Time Constraints**

Any alteration to the schedule may have an impact on the test plan and result in a delay. The test team has assumed that future adjustments would involve lengthening working hours and modifying the timetable.

##### **7.5.6.4.4 Change Management Procedure**

Every programmer may believe that the test plan has to be changed. They have to go through a series of steps, nevertheless, to get the suggested permission for the modification. The following are the steps involved in modifying the test plan:

- First, before making any changes, the supervisor and team members need to be consulted. This might take place in person or through online meetings.
- Secondly, they ought to determine whether to accept it or reject it.
- Lastly, the change process moves forward once the modification is documented if the team members and the supervisor approve the change.

#### **7.6 Comparison of design choice: (ABET requirement)**

In this section, we provide evidence for the design decisions we took in our project by showing how the outcomes outperform earlier research using similar datasets. We demonstrate the advantage of our concept in producing better results through analysis and graph representations.

##### **7.6.1 Osteoporosis disease**

The Knee X-ray Osteoporosis Database is the dataset utilized for osteoporosis prediction and it contains both images and clinical data [69]. Our study uses clinical data to the diagnoses of osteoporosis and osteopenia. We applied preprocessing techniques, such as dealing with class

imbalance and handling outliers. We trained five classifiers using Sequential Forward Feature Selection (SFFS) to identify important features and applied SMOTETomek technique. The Random Forest achieved the best accuracy of 91.11% and 91% recall, 92% precision, and 91% F1 score with SMOTETomek and 20 features. Additionally, we improved the interpretability of the model by utilizing XAI methods such as SHAP and LIME.

The earlier research used transfer learning to develop a CNN-based method for osteoporosis identification from the x-ray's images part of the dataset. While achieving a similar accuracy of 91.1%, our study offers enhanced interpretability through explicit feature selection and XAI methods. Furthermore, because our approach uses clinical data rather than images for diagnosis, it may be more affordable and easily accessible for clinics and hospitals.

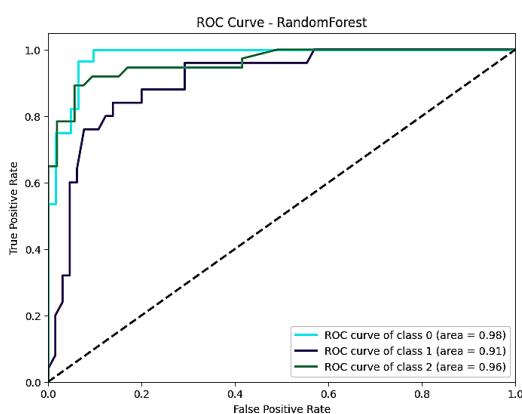


Figure 149 ROC-AUC for the RF classifier

### 7.6.2 Epileptic seizure disease

Our research on the preemptive diagnosis of epileptic seizures using machine learning algorithms, leveraging clinical data. Our result outperformed previous study conducted on the same dataset [68]. While both studies delve into neurological conditions involving seizures, they approach the problem from different angles. Our study focuses on preemptive diagnosis using machine learning algorithms to distinguish between epileptic and non-epileptic seizures, employing diagnostic, clinical, and demographic data. Through the optimization of models and utilization of feature selection techniques, Logistic Regression emerged as the most effective method with an accuracy of 87.67%. Conversely, the other study validates the dissociative seizure likelihood score (DSLS) to differentiate between epileptic and non-epileptic seizures based on clinical and medication history. Despite local variability, the DSLS accurately predicted the diagnosis in 81% of patients.

While our study emphasizes the application of machine learning in predictive modeling, the other study highlights the utility of clinical scoring systems in aiding diagnostic decision-making for challenging patient populations.

Our findings underscore the significance of our design choices in achieving superior results and demonstrate the potential of machine learning in enhancing predictive modeling for epileptic seizure diagnosis. In the ROC curve, we see that the Logistic Regression model exhibits a robust discrimination ability among the machine learning models evaluated. The Logistic Regression demonstrates high Area Under the Curve (AUC) values for all classes, 0.95 for Class 0 which represents epileptic seizure class, 0.96 for Class 1 that represents both epileptic seizure and non-epileptic seizure class, and 0.95 for Class 2 that represents non-epileptic seizure class, implying a strong predictive performance with high true positive rates and low false positive rates.

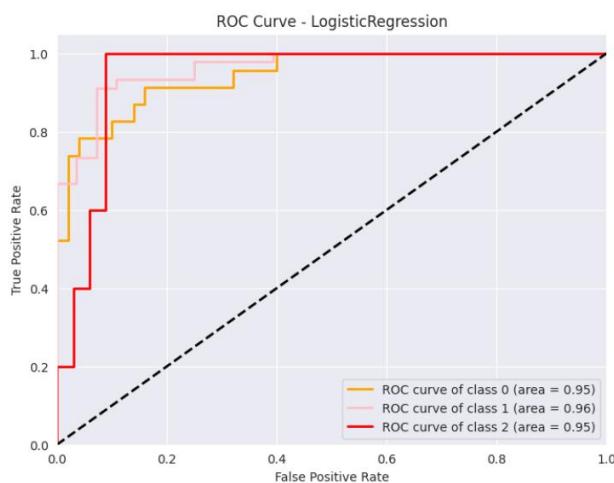
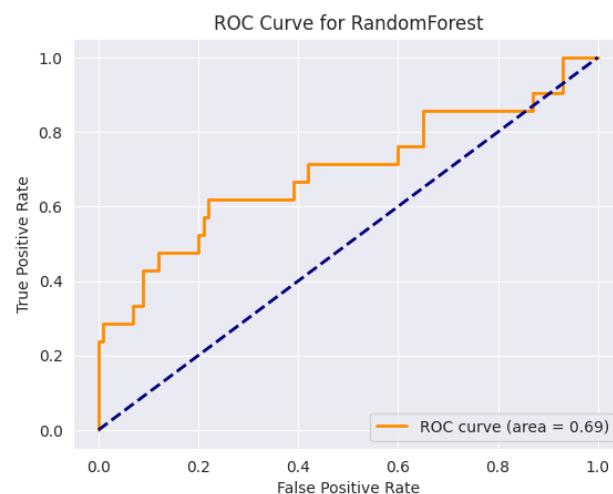


Figure 150 a graph shows AUC for LR classifier.

### 7.6.3 Sickle cell anemia disease

Our research on the Pre-emptive Diagnosis of Sickle Cell Anemia using Clinical Data revolves around using machine learning techniques to identify features associated with sickle cell anemia and predict and detect the disease early. Specifically, our approach utilizes Random Forest and achieves an accuracy of 86.77% using only 5 features out of 10. We employed feature selection methods to identify the most informative features. This demonstrates the effectiveness of machine learning in analyzing the dataset and extracting meaningful patterns related to sickle cell anemia.

In comparison to our approach, a similar study used the same dataset [70] was not focused on the prediction and early diagnosis of sickle cell anemia. Rather, their objective centered around understanding the characteristics that affect sickle cell anemia within a specific population. While our project focused on using machine learning for predictive modeling through feature selection, the other study contributed crucial epidemiological data essential for understanding the disease's prevalence and characteristics within a specific population. The ROC curves displayed across this figure measure the performance of the model utilized for predicting clinical outcomes. Each curve illustrates the true positive rate against the false positive rate for different thresholds, where a higher Area Under the Curve (AUC) indicates more effective model performance. In this analysis, the Random Forest model report an AUC of 0.69, indicating a strong capability to distinguish between patients with sickle cell anemia and healthy patients accurately.



*Figure151 151 a graph shows AUC of RF classifier.*

## Chapter 8: Conclusion

### 8.1 introduction

In this project, clinical data will be utilized to apply computational intelligence techniques for the preemptive diagnosis of diseases such as sickle cell anemia, osteoporosis, and epileptic seizures. Early diagnosis of the above chronic diseases may be very valuable to the health sector and to raising the standard of living in society. Additionally, it may significantly slow their spread and reduce the fatality rate. This project is expected to positively impact the health and economic sectors by addressing the mentioned issues.

Even if a number of studies were conducted to identify the proposed chronic diseases at an early stage, most of them used imaging technologies to obtain data that might not be available in every hospital. No studies explored the preemptive diagnosis of the specified chronic diseases using datasets from Saudi Arabia. Yet, these studies offer valuable insights to reach the aim of developing a preventive diagnostic system suitable for deployment in Saudi hospitals. By employing Saudi clinical datasets that may be used in hospitals with minimal resources, this project is intended to close the gap, decreasing the potential risks associated with the late identification of some chronic diseases.

The system was built using the Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), Gradient Boosting (GBoost), Extreme Gradient Boosting (XGBoost), AdaBoost, Logistic Regression, Decision Tree, SVM\_Gaussian, and GaussianNB algorithms, all of which performed well. The system's users include the admin, registered user, laboratory specialist, medical specialist, and guest. Each user has certain requirements, constraints, functionalities, and software attributes specified.

Throughout the development of the system interfaces, mock-ups that demonstrate how the system functions and operates were utilized. Furthermore, a flowchart was also created to define and model the data design and system architecture. Additionally, each function was written in Pseudocode, with the components' operations and constraints specified.

### 8.2 Finding and Contributions

The Kingdom of Saudi Arabia's public health care will eventually be improved with the help of the preemptive diagnosis system. As a result, the population's health will improve faster and there will be fewer fatalities. The system's simplicity will make the diagnosis procedure more efficient. Moreover, it will benefit the Kingdom's data mining research field.

### 8.3 Entrepreneurship Impact

According to a recent IDC study, worldwide spending on AI systems will reach \$600 billion in 2026, up from \$118 billion in 2022 [124]. AI and machine learning are transforming the way businesses connect with consumers and provide more in less time. ML has the potential to be employed in a variety of industries, including healthcare. As a result, incorporating AI into a business might provide a new breeding ground for entrepreneurial opportunities.

Our preemptive diagnostic system, Easy Diagnosis, utilizes clinical data to identify various chronic diseases in their early stages, even before symptoms manifest, aiming to enhance healthcare systems. We envision Easy Diagnosis evolving into a substantial healthcare-focused startup, offering subscription services to interested medical centers and government entities. The system seamlessly integrates with any hospital laboratory, analyzing patient clinical data and preemptively diagnosing chronic diseases through computational intelligence techniques. Subsequently, proactive medical centers, staffed with psychologists, receive and communicate results to patients with professional insights. Additionally, individuals, including patients and guests, can access the website anytime, inputting data for self-diagnosis. Ultimately, doctors gain the ability to preemptively deduce their patients' medical status and review results.

#### **8.4 Issue Faced**

Dealing with issues and facing challenges cannot be avoided during the process of senior project implementation and other stages. Solving these issues and finding a solution to them is essential to continue and enhance the project. Finding a reliable dataset was the first issue we faced. Clinical datasets are rarely found and available. Since we started the implementation stage, most of these issues were technical and coding related. One of these problems is finding suitable machine-learning algorithms that result in the best performance and accuracy. Another issue is cleaning our dataset since the datasets contain problems like missing values, irrelevant features, and unbalanced data, dealing with these problems was time consuming. Many other issues related to the modeling part and evaluating the performance were faced and resolved. Developing the website has its share of issues as well, installing required libraries didn't work smoothly. In addition to other issues such as storage problems.

#### **8.5 Lessons Learned**

By means of this project, we have enhanced and improved what we have learned in machine learning. From analyzing and pre-processing our dataset. To deal with several models and select the best methods to enhance the performance of the model. Furthermore, we also became familiar with the process of searching for a suitable dataset, what are the sources that could help us, and what criteria the dataset should have. We were also provided with medical knowledge about the diseases we work with. While editing the website, we learned about several language including HTML, Java Script, CSS, and JQuery. In addition to build a new data using a program called MongoDB. The project teaches us the real value of time and that every second matter, we learned how to manage our time properly to finish the required tasks on time. In addition, we were given the opportunity to write a research paper about our machine-learning model, this helped improve our writing and searching skills. Working as a team let us learn from each other, we exchanged a lot of experiences and knowledge. We faced challenges together and fixed any issues facing the success of our project. The feedback we received from our supervisors was a great help with these problems. These lessons we learned will help us in real life.

## **8.6 Client requirements**

Meeting clients' requirements is essential for product providers. These requirements need to be understood to ensure clients' satisfaction. This project preemptive diagnosis enhances patient care by identifying potential health issues before they become critical. Here is a list of potential customers need that can be met:

- Early detection of chronic disease
- Reducing unnecessary hospital visits
- Reducing pressure on hospitals
- Availability
- Easy access

## **8.6 Recommendations for future works**

To support the Kingdom of Saudi Arabia's healthcare system, we suggest extending the project to cover other chronic conditions on the website for proactive diagnosis. Additionally, we strongly recommend machine learning to the medical professions in the Kingdom.

## References

- [1] “Non communicable diseases,” World Health Organization, <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases> (accessed Oct. 8, 2023).
- [2] M. Heine and S. Hanekom, “Chronic disease in low-resource settings: Prevention and management throughout the continuum of care-a call for papers,” International journal of environmental research and public health, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9966920/#:~:text=It%20is%20well%20described%20that,individuals%20with%20multimorbidity%20%5B7%5D> (accessed Oct. 8, 2023).
- [3] “National Plan for Osteoporosis Prevention and Management in the Kingdom of Saudi Arabia,” Ministry of Health, <https://www.moh.gov.sa/en/Ministry/MediaCenter/Publications/Documents/NPOPM-2018.pdf> (accessed Oct. 8, 2023).
- [4] A. Al Rumayyan et al., “The prevalence of active epilepsy in the Kingdom of Saudi Arabia: A cross-sectional study,” Neuroepidemiology, <https://pubmed.ncbi.nlm.nih.gov/36209733/> (accessed Oct. 8, 2023).
- [5] Bin Zuair, A. et al. (2023) ‘The burden of sickle cell disease in Saudi Arabia: A single-institution large retrospective study’, International Journal of General Medicine, Volume 16, pp. 161–171. doi:10.2147/ijgm.s393233.
- [6] “Epilepsy,” American Association of Neurological Surgeons, <https://www.aans.org/en/Patients/Neurosurgical-Conditions-and-Treatments/Epilepsy> (accessed Oct. 8, 2023).
- [7] “Osteoporosis,” National Institute on Aging, <https://www.nia.nih.gov/health/osteoporosis> (accessed Oct. 8, 2023).
- [8] “Osteoporosis,” National Institute of Arthritis and Musculoskeletal and Skin Diseases, <https://www.niams.nih.gov/health-topics/osteoporosis#:~:text=Osteoporosis%20in%20Men%20and%20Women%20is%20a%20bone%20disease%20that%20develops%20when%20bone%20mineral,Pregnancy%2C%20Breastfeeding%2C%20and%20Bone%20Health> (accessed Oct. 8, 2023).
- [9] J. Peng, E. C. Jury, P. Dönnes, and C. Ciurtin, “Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: Applications and challenges,” Frontiers, <https://www.frontiersin.org/articles/10.3389/fphar.2021.720694/full> (accessed Oct. 8, 2023).
- [10] S. Ganiger and R. K M M, “Chronic Diseases Diagnosis using Machine Learning,” ResearchGate, [https://www.researchgate.net/publication/335576533\\_Chronic\\_Diseases\\_Diagnosis\\_using\\_Machine\\_Learning](https://www.researchgate.net/publication/335576533_Chronic_Diseases_Diagnosis_using_Machine_Learning) (accessed Oct. 8, 2023).

- [11] A. M. Alkhotani, “Teachers and epilepsy in Saudi Arabia: Gaps in knowledge and potential roles,” *International Journal of General Medicine*, vol. Volume 15, pp. 795–801, 2022. doi:10.2147/ijgm.s349302
- [12] S. Muhammad Usman, S. Khalid, and M. H. Aslam, “Epileptic seizures prediction using Deep Learning Techniques,” IEEE Access, vol. 8, pp. 39998–40007, 2020. doi:10.1109/access.2020.2976866
- [13] Ministry Portal Team, “Osteoporosis,” Saudi Ministry of Health, <https://www.moh.gov.sa/HealthAwareness/EducationalContent/Diseases/OrthopedicDiseases/Pages/001.aspx> (accessed Sep. 30, 2023).
- [14] Jang, R. et al. (2021) ‘Prediction of osteoporosis from simple hip radiography using deep learning algorithm’, *Scientific Reports*, 11(1). doi :10.1038/s41598-021-99549-6.
- [15] Ministry Portal Team, “Sickle Cell Anemia (Sickle Cell Anemia),” Saudi Ministry of Health, <https://www.moh.gov.sa/HealthAwareness/EducationalContent/Diseases/Hematology/Pages/SickleCell-Anemia.aspx> (accessed Sep. 30, 2023).
- [16] R. G. Zaini, “Sickle-cell Anemia and Consanguinity among the Saudi Arabian,” iMedPub Journals, vol. 8, Jun. 2016.
- [17] فريق بوابة وزارة الصحة [17], “International epilepsy day,” Ministry Of Health Saudi Arabia, <https://www.moh.gov.sa/en/HealthAwareness/healthDay/2020/Pages/HealthDay-2020-02-10.aspx> (accessed Dec. 2, 2023).
- [18] Chen, W., Wang, Y., Ren, Y., Jiang, H., Du, G., Zhang, J., & Li, J. (2023). An automated detection of epileptic seizures EEG using CNN classifier based on feature fusion with high accuracy. *BMC Medical Informatics and Decision Making*, 23(1). <https://doi.org/10.1186/s12911-023-02180-w>
- [19] Ode, R., Fujiwara, K., Miyajima, M. et al. "Development of an epileptic seizure prediction algorithm using R–R intervals with self-attentive autoencoder". *Artif Life Robotics* 28, 403–409 (2022). Doi: <https://doi.org/10.1007/s10015-022-00832-0>
- [20] J.Nazari , A. Nasrabadi , M. Menhaj, and S Raiesdana . "Epilepsy seizure prediction with few-shot learning method". *Brain Inform.* 2022;9(1):21. Published 2022 Sep 16. doi:10.1186/s40708-022-00170-8
- [21] M. Tawfik, E. Mahyoub, Z. A. Ahmed, N. M. Al-Zidi, and S. Nimbhore, “Classification of epileptic seizure using machine learning and deep learning based on electroencephalography (EEG),” *Communication and Intelligent Systems*, pp. 179–199, 2022. doi:10.1007/978-981-19-2130-8\_15
- [22] A. M. Hilal et al., “Intelligent Epileptic Seizure Detection and Classification Model Using Optimal Deep Canonical Sparse Autoencoder,” *Biology*, vol. 11, no. 8, p. 1220, Aug. 2022, doi: 10.3390/biology11081220.

- [23] I. Jemal, N. Mezghani, L. Abou-Abbas, and A. Mitiche, "An interpretable deep learning classifier for epileptic seizure prediction using EEG Data," *IEEE Access*, vol. 10, pp. 60141–60150, 2022. doi:10.1109/access.2022.3176367
- [24] O. Ouichka, A. Echtioui, and H. Hamam, "Deep Learning Models for Predicting Epileptic Seizures Using iEEG Signals," *Electronics*, vol. 11, no. 4, p. 605, Feb. 2022, doi: 10.3390/electronics11040605.
- [25] S. Usman, S. Khalid, and Z. Bashir," Epileptic seizure prediction using scalp electroencephalogram signals" *Sciedirect*, Vol. 41, no.1, p. 211-220. Jan.2021, doi: <https://doi.org/10.1016/j.bbe.2021.01.001>
- [26] Almustafa, K. M. (2020). Classification of epileptic seizure dataset using different machine learning algorithms. *Informatics in Medicine Unlocked*, 21, 100444. <https://doi.org/10.1016/J.IMU.2020.100444>
- [27] M. A. Jumaah, A.I.Shihab and A.A.Farhan, "Epileptic seizures detection using DCT-II and KNN classifier in long -Term EEG Signals" *Researchgate*, Feb. 2020, doi: [https://www.researchgate.net/publication/339302243\\_Epileptic\\_Seizures\\_Detection\\_Using\\_DCT-II\\_and\\_KNN\\_Classifier\\_in\\_Long-Term\\_EEG\\_Signals](https://www.researchgate.net/publication/339302243_Epileptic_Seizures_Detection_Using_DCT-II_and_KNN_Classifier_in_Long-Term_EEG_Signals)
- [28] M. Savadkoohi, T. Oladunni, and L. Thompson, "A machine learning approach to epileptic seizure prediction using electroencephalogram (EEG) signal," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 3, pp. 1328–1341, 2020. doi:10.1016/j.bbe.2020.07.004
- [29] C. L. Liu, B. Xiao, W. H. Hsiao, and V. S. Tseng, "Epileptic Seizure Prediction With Multi-View Convolutional Neural Networks," *IEEEExplore*, <https://ieeexplore.ieee.org/document/8910555?denied=> (accessed Oct. 8, 2023).
- [30] S. Usman, M. Usman, and S. Fong, "Epileptic Seizures Prediction Using Machine Learning Methods", *Computational and Mathematical Methods in Medicine*, vol. 2017, Article ID 9074759, 10 pages, 2017. <https://doi.org/10.1155/2017/9074759>
- [31] I. K. Kornek et al., "Epileptic Seizure Prediction Using Big Data and Deep Learning: Toward a Mobile System." IBM, Australia, Dec. 12, 2017.
- [32] Alamri, F.A. *et al.* (2015) 'Knowledge, attitude, and practice of osteoporosis among Saudis', *Journal of the Egyptian Public Health Association*, 90(4), pp. 171–177. doi:10.1097/01.epx.0000475735.83732.fc.
- [33] G. A. Albuquerque et al., "Osteoporosis screening using machine learning and Electromagnetic Waves," *Nature News*, <https://www.nature.com/articles/s41598-023-40104-w> (accessed Oct. 8, 2023).
- [34] X. Wu et al., "Development of machine learning models for predicting osteoporosis in patients with type 2 diabetes mellitus—a preliminary study," *Diabetes, Metabolic Syndrome and Obesity*, vol. Volume 16, pp. 1987–2003, 2023. doi:10.2147/dmso.s406695

- [35] X. Wu and S. Park, “A prediction model for osteoporosis risk using a machine-learning approach and its validation in a large cohort,” Journal of Korean Medical Science, vol. 38, no. 21, 2023. doi:10.3346/jkms.2023.38.e162
- [36] Inui, A. et al. (2023) ‘Screening for osteoporosis from blood test data in elderly women using a machine learning approach’, Bioengineering, 10(3), p. 277. doi:10.3390/bioengineering10030277.
- [37] Y. Ma, Q. Lu, F. Yuan, and H. Chen, “Comparison of the effectiveness of different machine learning algorithms in predicting new fractures after PKP for osteoporotic vertebral compression fractures,” Journal of Orthopaedic Surgery and Research, vol. 18, no. 1, 2023. doi:10.1186/s13018-023-03551-9
- [38] L. Fasihi, B. Tartibian, R. Eslami, and H. Fasihi, “Artificial intelligence used to diagnose osteoporosis from risk factors in clinical data and proposing sports protocols,” Nature News, <https://www.nature.com/articles/s41598-022-23184-y> (accessed Oct. 8, 2023).
- [39] R. Dzierżak and Z. Omietek, “Application of deep convolutional neural networks in the diagnosis of osteoporosis,” Sensors, vol. 22, no. 21, p. 8189, 2022. doi:10.3390/s22218189
- [40] Kwon, Y. et al. (2022) ‘Osteoporosis pre-screening using ensemble machine learning in postmenopausal Korean women’, Healthcare, 10(6), p. 1107. doi:10.3390/healthcare10061107.
- [41] M. Jang et al., “Opportunistic osteoporosis screening using chest radiographs with Deep Learning: Development and external validation with a cohort dataset,” Journal of Bone and Mineral Research, vol. 37, no. 2, pp. 369–377, 2021. doi:10.1002/jbmr.4477
- [42] Ou Yang, W.-Y. et al. (2021) ‘Development of machine learning models for prediction of osteoporosis from Clinical Health Examination Data’, International Journal of Environmental Research and Public Health, 18(14), p. 7635. doi:10.3390/ijerph18147635.
- [43] M. Anam et al., “Osteoporosis prediction for trabecular bone using Machine Learning: A Review,” SSRN, [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3786263](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3786263) (accessed Oct. 8, 2023).
- [44] Yamamoto, N. et al. (2020) ‘Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates’, Biomolecules, 10(11), p. 1534. doi:10.3390/biom10111534.
- [45] X. Yu, C. Ye, and L. Xiang, “Application of artificial neural network in the diagnostic system of osteoporosis,” Neurocomputing, <https://www.sciencedirect.com/science/article/abs/pii/S0925231216306610> (accessed Oct. 8, 2023).
- [46] Yoo, T.K. et al. (2013) ‘Osteoporosis risk prediction for bone mineral density assessment of postmenopausal women using machine learning’, Yonsei Medical Journal, 54(6), p. 1321. doi:10.3349/ymj.2013.54.6.1321.

- [47] M. Darrin et al., “Classification of red cell dynamics with convolutional and recurrent neural networks: A sickle cell disease case study,” Nature News, <https://www.nature.com/articles/s41598-023-27718-w> (accessed Oct. 8, 2023).
- [48] O. B. Ayoade, A. L. Imoize, J. A. Adeloye, and T. O. Oladele, “An Ensemble Models for the Prediction of Sickle Cell Disease from Erythrocytes Smears,” ResearchGate, [https://www.researchgate.net/publication/374046799\\_An\\_Ensemble\\_Models\\_for\\_the\\_Prediction\\_of\\_Sickle\\_Cell\\_Disease\\_from\\_Erythrocytes\\_Smears](https://www.researchgate.net/publication/374046799_An_Ensemble_Models_for_the_Prediction_of_Sickle_Cell_Disease_from_Erythrocytes_Smears) (accessed Oct. 9, 2023).
- [49] D. Nguyen, L. Abraham, and A. Amatya, “Application of deep learning models into the prediction of interleukin ...,” Application of Deep Learning Models into the Prediction of Interleukin-6 and -8 Cytokines in Sickle Cell Anemia Patients, [https://www.researchgate.net/publication/372051232\\_Application\\_of\\_Deep\\_Learning\\_Models\\_into\\_the\\_Prediction\\_of\\_Interleukin-6\\_and\\_-8\\_Cytokines\\_in\\_Sickle\\_Cell\\_Anemia\\_Patients](https://www.researchgate.net/publication/372051232_Application_of_Deep_Learning_Models_into_the_Prediction_of_Interleukin-6_and_-8_Cytokines_in_Sickle_Cell_Anemia_Patients) (accessed Oct. 8, 2023).
- [50] D. C. Saputra, K. Sunat, and T. Ratnaningsih, “A new artificial intelligence approach using extreme learning machine as the potentially effective model to predict and analyze the diagnosis of anemia,” Healthcare, vol. 11, no. 5, p. 697, 2023. doi:10.3390/healthcare11050697
- [51] R. Vohra, A. Hussain, A. K. Dudyala, J. Pahareeya, and W. Khan, “Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting,” PLOS ONE, vol. 17, no. 7, 2022. doi:10.1371/journal.pone.0269685.
- [52] C. Patgiri and A. Ganguly, “Machine Learning Techniques for Automatic Detection of Sickle Cell Anemia using Adaptive Thresholding and Contour-based Segmentation Method,” ResearchGate, [https://www.researchgate.net/publication/369739403\\_Machine\\_Learning\\_Techniques\\_for\\_Automatic\\_Detection\\_of\\_Sickle\\_Cell\\_Anemia\\_using\\_Adaptive\\_Thresholding\\_and\\_Contour-based\\_Segmentation\\_Method](https://www.researchgate.net/publication/369739403_Machine_Learning_Techniques_for_Automatic_Detection_of_Sickle_Cell_Anemia_using_Adaptive_Thresholding_and_Contour-based_Segmentation_Method) (accessed Oct. 9, 2023).
- [53] M. Shahzad et al., “Identification of anemia and its severity level in a peripheral blood smear using 3-tier deep neural network,” Applied Sciences, vol. 12, no. 10, p. 5030, 2022. doi:10.3390/app12105030
- [54] S. Srivastava, R. N. K, R. Srinivasan, N. K. Nambison, and S. S. Gorthi, “Diagnosis of sickle cell anemia using AutoML on UV-vis absorbance ...,” Diagnosis of sickle cell anemia using AutoML on UV-Vis absorbance spectroscopy data, [https://www.researchgate.net/publication/356602448\\_Diagnosis\\_of\\_sickle\\_cell\\_anemia\\_using\\_AutoML\\_on\\_UV-Vis\\_absorbance\\_spectroscopy\\_data](https://www.researchgate.net/publication/356602448_Diagnosis_of_sickle_cell_anemia_using_AutoML_on_UV-Vis_absorbance_spectroscopy_data) (accessed Oct. 8, 2023).
- [55] S. Yeruva, M. S. Varalakshmi, B. P. Gowtham, Y. H. Chandana, and PESN. K. Prasad, “Identification of sickle cell anemia using deep neural networks,” Emerging Science Journal, vol. 5, no. 2, pp. 200–210, 2021. doi:10.28991/esj-2021-01270

- [56] Abdulkarim, H.A. et al. (2020) ‘A deep learning alexnet model for classification of red blood cells in sickle cell anemia’, IAES International Journal of Artificial Intelligence (IJ-AI), 9(2), p. 221. doi:10.11591/ijai.v9.i2.pp221-228.
- [57] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, J. Zhang, and Y. Duan, “Deep learning models for classification of red blood cells in microscopy images to aid in sickle cell anemia diagnosis,” Electronics, vol. 9, no. 3, p. 427, 2020. doi:10.3390/electronics9030427
- [58] Mohammed, A. et al. (2019) Machine learning predicts early-onset acute organ failure in critically ill patients with sickle cell disease [Preprint]. doi:10.1101/614941.
- [59] Alzubaidi, L. et al. (2019) ‘Classification of red blood cells in sickle cell anemia using deep convolutional neural network’, Advances in Intelligent Systems and Computing, pp. 550–559. doi:10.1007/978-3-030-16657-1\_51.
- [60] M. Xu, S. Abidi, M. Dao, and D. P.Papageorgiou, “A deep convolutional neural network for classification of red blood cells in sickle cell anemia,” ResearcgGate, [https://www.researchgate.net/publication/320510877\\_A\\_deep\\_convolutional\\_neural\\_network\\_for\\_classification\\_of\\_red\\_blood\\_cells\\_in\\_sickle\\_cell\\_anemia](https://www.researchgate.net/publication/320510877_A_deep_convolutional_neural_network_for_classification_of_red_blood_cells_in_sickle_cell_anemia) (accessed Oct. 9, 2023).
- [61] Akrimi, J.A. et al. (2014) ‘Classification red blood cells using support Vector Machine’, Proceedings of the 6th International Conference on Information Technology and Multimedia [Preprint]. doi:10.1109/icimu.2014.7066642.
- [62] “About chronic diseases,” Centers for Disease Control and Prevention, <https://www.cdc.gov/chronicdisease/about/index.htm#:~:text=Print-,About%20Chronic%20Diseases,of%20daily%20living%20or%20both> (accessed Oct. 9, 2023).
- [63] Admin, “Software project management plan: Steps and tips,” Designveloper, <https://www.designveloper.com/blog/software-project-management-plan/> (accessed Oct. 9, 2023).
- [64] G. K. Lane, “HOW TO WRITE A software requirements specification (SRS document),” Perforce Software, [https://www.perforce.com/blog/alm/how-write-software-requirements-specification-srs-document#:~:text=A%20software%20requirements%20specification%20\(SRS\)%20is%20a%20document%20that%20describes,stakeholders%20\(business%2C%20users\)](https://www.perforce.com/blog/alm/how-write-software-requirements-specification-srs-document#:~:text=A%20software%20requirements%20specification%20(SRS)%20is%20a%20document%20that%20describes,stakeholders%20(business%2C%20users)) (accessed Oct. 9, 2023).
- [65] “Software design specification.,” PresentationEZE.com, <https://www.presentationeze.com/presentations/software-validation/software-validation-full-details/software-design-specification/#:~:text=Specifically%20the%20software%20design%20specification%20necessary%20code%20to%20be%20produced> (accessed Oct. 9, 2023).
- [66] “Test planning: A detailed guide,” BrowserStack, <https://www.browserstack.com/guide/test->

planning#:~:text=A%20Test%20Plan%20is%20a,correctly%20%E2%80%93%20contr  
olled%20by%20test%20managers (accessed Oct. 9, 2023).

- [67] “SDLC - Waterfall Model,” Online Courses and eBooks Library, [https://www.tutorialspoint.com/sdlc/sdlc\\_waterfall\\_model.htm](https://www.tutorialspoint.com/sdlc/sdlc_waterfall_model.htm) (accessed Oct. 9, 2023).
- [68] S. Lenio et al., “Data & Code from Validation of a predictive calculator to distinguish between patients presenting with dissociative versus epileptic seizures,” [data.mendeley.com](https://data.mendeley.com), vol. 1, Feb. 2021, doi: <https://doi.org/10.17632/cshccr8w3h.1>.
- [69] Majeed Wani, Insha ; Arora, Sakshi (2021), “Knee X-ray Osteoporosis Database”, Mendeley Data, V2, doi: 10.17632/fxjm8fb6mw.2 (Accessed: 29 October 2023).
- [70] Adam, M.A., Adam, N.K. & Mohamed, B.A. "Prevalence of sickle cell disease and sickle cell trait among children admitted to Al Fashir Teaching Hospital North Darfur State, Sudan". BMC Res Notes 12, 659 (2019). <https://doi.org/10.1186/s13104-019-4682-5>.
- [71] L. Breiman, “Random forests,” Mach Learn, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
- [72] F. Livingston, “Implementation of Breiman’s Random Forest Machine Learning Algorithm,” ECE591Q Machine Learning Journal Paper. Fall, 2005.
- [73] N. Aslam et al., “Anomaly Detection Using Explainable Random Forest for the Prediction of Undesirable Events in Oil Wells,” Applied Computational Intelligence and Soft Computing, vol. 2022, 2022, doi: 10.1155/2022/1558381.
- [74] I. Nitze, U. Schulthess, and H. Asche, “COMPARISON OF MACHINE LEARNING ALGORITHMS RANDOM FOREST … ,” 2012.
- [75] A. Ben-Hur and J. Weston, “A user’s guide to support vector machines.,” Methods Mol Biol, vol. 609, pp. 223–239, 2010, doi: 10.1007/978-1-60327-241-4\_13/COVER.
- [76] R. G. Brereton and G. R. Lloyd, “Support Vector Machines for classification and regression,” Analyst, vol. 135, no. 2, pp. 230–267, Jan. 2010, doi: 10.1039/B918972F.
- [77] J. M. Moguerza and A. Muñoz, “Support vector machines with applications,” Statistical Science, vol. 21, no. 3, pp. 322–336, Aug. 2006, doi: 10.1214/088342306000000493.
- [78] P. Janardhanan and F. Sabika, “Effectiveness of Support Vector Machines in Medical Data mining,” 2015.
- [79] “(PDF) Efficient Support Vector Machines for Spam Detection: A Survey.” [https://www.researchgate.net/publication/316075352\\_Efficient\\_Support\\_Vector\\_Machines\\_for\\_Spam\\_Detection\\_A\\_Survey](https://www.researchgate.net/publication/316075352_Efficient_Support_Vector_Machines_for_Spam_Detection_A_Survey) (accessed Nov. 04, 2022).
- [80] C. Authors, “Mathematical and Quantitative Methods,” Acta Universitatis Danubius. Œconomica, vol. 8, no. 5, 2012, Accessed: Nov. 04, 2022. [Online]. Available: <https://dj.univ-danubius.ro/index.php/AUDOE/article/view/1065>

- [81] J. Cardoso-Fernandes, A. C. Teodoro, A. Lima, and E. Roda-Robles, “Semi-automatization of support vector machines to map lithium (Li) bearing pegmatites,” *Remote Sens (Basel)*, vol. 12, no. 14, Jul. 2020, doi: 10.3390/RS12142319.
- [82] H. Yu and S. Kim, “SVM tutorial-classification, regression and ranking,” *Handbook of Natural Computing*, vol. 1–4, pp. 479–506, Jan. 2012, doi: 10.1007/978-3-540-92910-9\_15.
- [83] H. Al-Behadili, A. Grumpe, C. Dopp, and C. Wohler, “Non-linear distance based large scale data classifications,” *Proceedings of 2015 IEEE International Conference on Progress in Informatics and Computing, PIC 2015*, pp. 613–617, Jun. 2016, doi: 10.1109/PIC.2015.7489921.
- [84] M. Ramachandro and R. Bhramaramba, “Classification of gene expression data set using support vectors machine with RBF kernel,” *International Journal of Recent Technology and Engineering*, vol. 8, no. 2, pp. 2907–2913, Jul. 2019, doi: 10.35940/IJRTEB2463.078219.
- [85] M. Ramachandro and R. Bhramaramba, “Classification of gene expression data set using support vectors machine with RBF kernel,” *International Journal of Recent Technology and Engineering*, vol. 8, no. 2, pp. 2907–2913, Jul. 2019, doi: 10.35940/IJRTEB2463.078219.
- [86] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Stat Comput*, vol. 14, no. 3, pp. 199–222, Aug. 2004, doi: 10.1023/B:STCO.0000035301.49549.88.
- [87] I. Adrianto, T. B. Trafalis, and V. Lakshmanan, “Support vector machines for spatiotemporal tornado prediction,” *Int J Gen Syst*, vol. 38, no. 7, pp. 759–776, Oct. 2009, doi: 10.1080/03081070601068629.
- [88] K. Taunk, S. De, S. Verma, and A. Swetapadma, “A brief review of nearest neighbor algorithm for learning and classification,” *2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019*, pp. 1255–1260, May 2019, doi: 10.1109/ICCS45141.2019.9065747.
- [89] S. Yang, H. Jian, Z. Ding, Z. Hongyuan, and C. L. Giles, “IKNN: Informative K-nearest neighbor pattern classification,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4702 LNBI, pp. 248–264, 2007, doi: 10.1007/978-3-540-74976-9\_25/COVER.
- [90] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “KNN model-based approach in classification,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2888, pp. 986–996, 2003, doi: 10.1007/978-3-540-39964-3\_62/COVER.
- [91] R. C. Neath and M. S. Johnson, “Discrimination and Classification,” *International Encyclopedia of Education*, pp. 135–141, Jan. 2010, doi: 10.1016/B978-0-08-044894-7.01312-9.

- [92] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, doi: 10.1145/2939672.
- [93] D. Zhang et al., "IBLP: An XGBoost-Based Predictor for Identifying Bioluminescent Proteins," Comput Math Methods Med, vol. 2021, 2021, doi: 10.1155/2021/6664362.
- [94] N. Dhibe, H. Ghazzai, H. Besbes, and Y. Massoud, "Extreme gradient boosting machine learning algorithm for safe auto insurance operations," 2019 IEEE International Conference on Vehicular Electronics and Safety, ICVES 2019, Sep. 2019, doi: 10.1109/ICVES.2019.8906396.
- [95] Brownlee, J. (2021) Gradient boosting with Scikit-learn, XGBoost, LIGHTGBM, and CatBoost, MachineLearningMastery.com. Available at: <https://machinelearningmastery.com/gradient-boosting-with-scikit-learn-xgboost-lightgbm-and-catboost/> (Accessed: 29 October 2023).
- [96] Breiman, L. (June 1997). "Arcing The Edge" (PDF). Technical Report 486. Statistics Department, University of California, Berkeley.
- [97] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
- [98] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees, Belmont, CA, USA: Wadsworth, 1984.
- [99] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," J. Comput. Syst. Sci., vol. 55, no. 1, pp. 119–139, 1997.
- [100] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.
- [101] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, 1995, pp. 338-345.
- [102] S. Gupta and M. K. Gupta, "Computational Prediction of Cervical Cancer Diagnosis Using Ensemble-Based Classification Algorithm," Comput J, vol. 65, no. 6, pp. 1527–1539, Jun. 2022, doi: 10.1093/COMJNL/BXAA198.
- [103] "Shneiderman's Eight Golden Rules Will Help You Design Better Interfaces | IxDF." <https://www.interaction-design.org/literature/article/shneiderman-s-eight-golden-rules-will-help-you-design-better-interfaces> (Accessed: 29 October 2023).
- [104] G. Battineni, N. Chintalapudi, and F. Amenta, "Machine learning in medicine: Performance calculation of dementia ...," Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM), [https://www.researchgate.net/publication/334054467\\_Machine\\_learning\\_in\\_medicine\\_](https://www.researchgate.net/publication/334054467_Machine_learning_in_medicine_)

Performance\_calculation\_of\_dementia\_prediction\_by\_support\_vector\_machines\_SVM  
(accessed Jan. 6, 2024).

- [105] Author links open overlay panel Pall Oskar Gislason et al., “Random forests for land cover classification,” Pattern Recognition Letters, <https://www.sciencedirect.com/science/article/abs/pii/S0167865505002242?via%3Dhub> (accessed Jan. 21, 2024).
- [106] R. Krishnamoorthi et al., “A novel diabetes healthcare disease prediction framework using machine learning techniques,” Journal of Healthcare Engineering, <https://www.hindawi.com/journals/jhe/2022/1684017/> (accessed Jan. 21, 2024).
- [107] “What is the K-nearest neighbors algorithm?,” IBM, <https://www.ibm.com/topics/knn> (accessed Jan. 21, 2024).
- [108] M. A. Vahedifar, A. Akhtarshenas, M. Sabbaghian, and M. Rafatpanah, “Information Modified K-Nearest Neighbor.” Tehran, Dec. 4, 2023.
- [109] W. Li, Y. Yin, X. Quan, and H. Zhang, “Gene expression value prediction based on XGBoost algorithm,” Frontiers, <https://www.frontiersin.org/articles/10.3389/fgene.2019.01077/full> (accessed Jan. 21, 2024).
- [110] N. K. Sinha, M. Khulal, M. Gurung, and A. Lal, “Developing a web based system for breast cancer prediction using XGboost classifier,” International Journal of Engineering Research & Technology, <https://www.ijert.org/developing-a-web-based-system-for-breast-cancer-prediction-using-xgboost-classifier> (accessed Jan. 21, 2024).
- [111] B. B. & B. Greenwell, “Hands-on machine learning with R,” Chapter 12 Gradient Boosting, <https://bradleyboehmke.github.io/HOML/gbm.html> (accessed Jan. 21, 2024).
- [112] Marcano-Cedeño, A. et al. (2010) Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network, researchgate. Available at: [https://www.researchgate.net/publication/224207758\\_Feature\\_selection\\_using\\_Sequential\\_Forward\\_Selection\\_and\\_classification\\_applying\\_Artificial\\_Metaplasticity\\_Neural\\_Network](https://www.researchgate.net/publication/224207758_Feature_selection_using_Sequential_Forward_Selection_and_classification_applying_Artificial_Metaplasticity_Neural_Network) (Accessed: 23 January 2024).
- [113] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
- [114] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189-1232.
- [115] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.
- [116] Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81-106

- [117] R. A. A. Viadinugroho, "Imbalanced Classification in Python: SMOTE-Tomek Links Method," Medium, Apr. 18, 2021. <https://towardsdatascience.com/imbalanced-classification-in-python-smote-tomek-links-method-6e48dfe69bbc>
- [118] GfG (2023) Chi-square test in machine learning, GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/ml-chi-square-test-for-feature-selection/> (Accessed: 03 April 2024).
- [119] J. Zhu et al., "Adaboost algorithm," 2009. [Online]. Available: <https://www.intlpress.com/site/pub/pages/journals/items/sii/content/vols/0002/0003/a008/>
- [120] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.
- [121] I. Rish, "An empirical study of the naive Bayes classifier," in IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 2001.
- [122] B. Schölkopf et al., "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond," MIT Press, 2002.
- [123] GeeksforGeeks (2024) Feature selection techniques in machine learning, GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/> (Accessed: 04 May 2024).
- [124] M. Shirer, "Worldwide spending on AI-centric systems will pass \$300 billion by 2026, according to IDC," IDC, <https://www.idc.com/getdoc.jsp?containerId=prUS49670322> (accessed Dec. 7, 2023).

## APPENDIX

### Appendix A: Response to comments

Tables below contain all suggestions and comments from the reviewers, and response to each comment.

<b>1- Comment:</b>	<b>Comment was by</b>
Add hyphen.	Mrs. Mehwash Farooqui
<b>Response</b>	
We thank the reviewer for pointing this out. We added a hyphen. Please see page 19, line 8.	

<b>2- Comment:</b>	<b>Comment was by</b>
Change the verb. It should be has.	Mrs. Mehwash Farooqui
<b>Response</b>	
We thank the reviewer for pointing this out. We revised the grammar and fixed it accordingly.	

<b>3- Comment:</b>	<b>Comment was by</b>
It should be who were.	Mrs. Mehwash Farooqui
<b>Response</b>	
We thank the reviewer for pointing this out. We revised the grammar and fixed it accordingly.	

<b>4- Comment:</b>	<b>Comment was by</b>
Correct it. It should be ages of.	Mrs. Mehwash Farooqui
<b>Response</b>	
We thank the reviewer for pointing this out. We revised the grammar and fixed accordingly.	

<b>5- Comment:</b>	<b>Comment was by</b>
Correct it " lightGBM "	Mrs. Mehwash Farooqui
<b>Response</b>	
We thank the reviewer for pointing this out, we correct it.	

<b>6- Comment:</b>	<b>Comment was by</b>
Correct it "Afterwards"	Mrs. Mehwash Farooqui
<b>Response</b>	

We thank the reviewer for pointing this out. We corrected it.

<b>7- Comment:</b>	<b>Comment was by</b>
All references should be in the same format.	Mrs. Mehwash Farooqui
<b>Response</b>	
We thank the reviewer for pointing this out. We've modified them to be the same format.	

<b>8- Comment:</b>	<b>Comment was by</b>
Missing section 1.7, 3.6 and 8.6	Mrs. Mehwash Farooqui
<b>Response</b>	
We thank the reviewer for pointing this out. We add these sections	

## Appendix B: plagiarism check

### Final Report of Graduation Project Proposal ARTI511-G#3.docx

#### ORIGINALITY REPORT

**61** % SIMILARITY INDEX    9% INTERNET SOURCES    11% PUBLICATIONS    55% STUDENT PAPERS

#### PRIMARY SOURCES

1	Submitted to University of Dammam Student Paper	52%
2	www.mdpi.com Internet Source	1%
3	mdpi-res.com Internet Source	<1%
4	fastercapital.com Internet Source	<1%
5	Submitted to Liverpool John Moores University Student Paper	<1%
6	Richard M. Hicks. "Implementing Always On VPN", Springer Science and Business Media LLC, 2022 Publication	<1%
7	Steven Lenio, Wesley T. Kerr, Meagan Watson, Sarah Baker, Chad Bush, Alex Rajic, Laura Strom. "Validation of a predictive calculator to distinguish between patients presenting with	<1%

## **Appendix C.1: Bi-weekly reports 1**

### **ARTI 511 – Project proposal**

**Term 1 – 2023 / 2024**

#### **Status Report #1**

Period from **20/8/ 2023** to **11/10/ 2023**

- Group #:3
- Members of the group:

#	Name	ID	Role
1	Rahaf Mofareh Yaan Allah	2200003935	Leader
2	Fai Saleh Alanazi	2200000931	Member
3	Razan Khalid Alshammari	2200004035	Member
4	Fatimah Abbas Alkhatim	2200001977	Member
5	Shahad Khalid Alghamdi	2200003434	Member

- The outcome of the meetings
  - 1- understanding the project requirements and specifications
  - 2- deciding which chronic diseases to work with
  - 3- deciding what are the datasets that will be used
  - 4- getting feedback from supervisors about works on progress
- Tasks planned to work on during the specified period:
  1. Searching for chronic diseases to be selected
  2. Searching for suitable dataset
  3. Start Writing introduction chapter
  4. Start writing literature reviews chapter
  5. Start writing SPMP
  6. Submitting midterm report for first semester
- Number of group meetings you had during the specified time: 6

## Appendix C.2: Bi-weekly reports 2

### ARTI 511 – Project proposal

Term 1 – 2023 / 2024

#### Status Report #2

Period from 15/10/ 2023 to 6/12/ 2023

- Group #:3
- Members of the group:

#	Name	ID	Role
1	Rahaf Mofareh Yaan Allah	2200003935	Leader
2	Fai Saleh Alanazi	2200000931	Member
3	Razan Khalid Alshammari	2200004035	Member
4	Fatimah Abbas Alkhatim	2200001977	Member
5	Shahad Khalid Alghamdi	2200003434	Member

- The outcome of the meetings
  1. Discussing the methodology to be used in implementing the models
  2. Getting feedback from supervisors during the implementation and research paper writing of osteoporosis disease
  3. Dividing tasks need to be completed for each team member
- Tasks planned to work on during the specified period:
  1. Start writing methodology and SRS chapter
  2. Start writing SDP chapter
  3. Submitting the final report
  4. Start the implementation for osteoporosis disease
  5. Start writing research paper for journal of preemptive diagnose of for osteoporosis disease
- Number of group meetings you had during the specified time: 5

#### Pending tasks

1. Finalizing the models implemented

## **Appendix C.3: Bi-weekly reports 3**

### **ARTI 511 – Project proposal**

**Term 2 – 2023 / 2024**

#### **Status Report #3**

**Period from 14/1/ 2024 to 29/2/ 2024**

- Group #:3
- Members of the group:

#	Name	ID	Role
1	Rahaf Mofareh Yaan Allah	2200003935	Leader
2	Fai Saleh Alanazi	2200000931	Member
3	Razan Khalid Alshammari	2200004035	Member
4	Fatimah Abbas Alkhatim	2200001977	Member
5	Shahad Khalid Alghamdi	2200003434	Member

- The outcome of the meetings
  1. Getting feedback from supervisors during the model implementation and research paper writing for epileptic seizure disease
  2. Getting feedback from supervisors during the model implementation and research paper writing for Sickle cell anemia disease
  3. Dividing tasks need to be complete between team members
- Tasks planned to work on during the specified period:
  1. Start implementing the model for epileptic seizure
  2. Start writing research paper for journal of preemptive diagnose of epileptic seizure
  3. Start implementing the model for Sickle cell anemia
  4. Start writing research paper for journal of preemptive diagnose Sickle cell anemia
  5. Writing conference paper for preemptive diagnose of osteoporosis
  6. Finalizing the osteoporosis model
- Number of group meetings you had during the specified time: 3

## Appendix C.4: Bi-weekly reports 4

### ARTI 511 – Project proposal

Term 2 – 2023 / 2024

#### Status Report #4

Period from 1/3/ 2024 to 18/5/ 2024

- Group #:3
- Members of the group:

#	Name	ID	Role
1	Rahaf Mofareh Yaan Allah	2200003935	Leader
2	Fai Saleh Alanazi	2200000931	Member
3	Razan Khalid Alshammari	2200004035	Member
4	Fatimah Abbas Alkhatim	2200001977	Member
5	Shahad Khalid Alghamdi	2200003434	Member

- The outcome of the meetings
  1. Understanding the process of implementing the website and what requirements need to be met
  2. Dividing tasks need to be complete between team members
- Tasks planned to work on during the specified period:
  1. Writing conference paper for preemptive diagnose of epileptic seizure
  2. Writing conference paper for preemptive diagnose of sickle cell anemia
  3. Start implementing the website
  4. Testing the system
  5. Start writing STP chapter
- Number of group meetings you had during the specified time: 5

# Appendix D.1: Osteoporosis Journal Paper

## Comprehensible Machine Learning-Based Models for the Pre-emptive Diagnosis of Osteoporosis and Osteopenia using Clinical Data

**Sunday O. Olatunji<sup>1\*</sup>, Mohammad Aftab Alam Khan<sup>1\*</sup>, Fai Alanazi<sup>1\*</sup>, Rahaf yaan allah<sup>1\*</sup>, Shahad alghamdi<sup>1\*</sup>, Razan Alshammari<sup>1\*</sup>, Fatimah Alkhatim<sup>1\*</sup>, Mehwash Farooqui<sup>1\*</sup> and Mohammed Imran Basheer Ahmed<sup>1\*</sup>**

<sup>1</sup> College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia.

\* Correspondence: <sup>1</sup> [osunday@iau.edu.sa](mailto:osunday@iau.edu.sa) <sup>2</sup> [mkhan@iau.edu.sa](mailto:mkhan@iau.edu.sa) <sup>3</sup> [2200000931@iau.edu.sa](mailto:2200000931@iau.edu.sa) <sup>4</sup> [2200003935@iau.edu.sa](mailto:2200003935@iau.edu.sa)  
<sup>5</sup> [2200003434@iau.edu.sa](mailto:2200003434@iau.edu.sa) <sup>6</sup> [2200004035@iau.edu.sa](mailto:2200004035@iau.edu.sa) <sup>7</sup> [2200001977@iau.edu.sa](mailto:2200001977@iau.edu.sa) <sup>8</sup> [mfarooqui@iau.edu.sa](mailto:mfarooqui@iau.edu.sa) <sup>9</sup> [mbahmed@iau.edu.sa](mailto:mbahmed@iau.edu.sa)

### **Abstract**

Osteoporosis and osteopenia are bone diseases characterized by low bone density and structural deterioration to an extent that makes the bones brittle and prone to fractures. The World Health Organization (WHO) differentiates osteopenia as a milder form from osteoporosis as a severe form associated with bone fractures based on statistical values. The fact that this disease is often silent, meaning it is not accompanied by noticeable symptoms, typically reveals itself when a bone fracture occurs. By then, the bones have undergone years of severe deterioration in strength and structure, which is why Bone Mineral Density (BMD) testing is essential for identifying osteoporosis and osteopenia and estimating the likelihood of future fractures. BMD is determined via a bone density test that uses dual-energy X-ray absorptiometry (DXA) to quantify the amount of minerals, such as calcium and phosphorus, inside the bones. Nevertheless, this method may be costly, which limits its accessibility. Therefore, this paper investigates the pre-emptive diagnosis of osteoporosis and osteopenia using machine learning techniques, exclusively utilizing clinical data without relying on Bone Mass Density (BMD) measurements. The dataset, which consists of 240 patients, originates from the BMD camp organized by the Unani and Panchkarma Hospital in Srinagar, J&K, India. Employing five machine learning algorithms—Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting (GBoost), and XGBoost — this study leverages GridSearchCV with 10-folds for model optimization. Central to the methodology is the utilization of Sequential Forward Feature Selection (SFFS) to reduce the number of features and enhance predictive accuracy. Significantly, employing the SMOTETomek approach revealed that Random Forest outpaced other methodologies, using 20 features, where it achieved an accuracy of 91.11% with precision, recall, and F1 values of 92%, 91%, and 91% respectively. Additionally, the study used Explainable Artificial Intelligence (XAI) techniques, such as Shapley Additive Explanation (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), to better understand the decision-making processes and the best-performing AI model.

### **1. Introduction**

Chronic diseases result from a combination of genetic, environmental, and lifestyle factors, often lasting a long time. Chronic diseases pose a significant challenge globally, with over three-quarters of the 31.4

million chronic diseases related deaths occurring in regions facing economic challenges. These diseases affect people of all ages and regions, with 17 million chronic diseases related deaths happening before age 70 [1]. Alarmingly, a vast majority of these early deaths occur in areas with limited economic resources [2]. Risk factors like unhealthy diets, lack of exercise, exposure to tobacco smoke, and pollution make children, adults, and the elderly vulnerable. These diseases are on the rise due to factors like rapid urbanization, globalization, and an aging population [1]. They're closely tied to poverty, hindering efforts to reduce it, as chronic diseases bring hefty healthcare costs. Costly treatments and income loss push millions into poverty annually [2].

In Saudi Arabia, chronic diseases are prevalent, with a variety of enduring health conditions such as osteoporosis and osteopenia. According to the Kingdom of Saudi Arabia's National Plan for Osteoporosis Prevention and Management, the prevalence of osteoporosis and its precursor, osteopenia, is 37.8% in men and 28.2% in women over the age of 50. With the projected demographic shift in Saudi Arabia, where the 50+ age group is expected to rise steeply over the next few decades, and life expectancy on the rise, the Kingdom anticipates a surge in osteoporosis cases [3].

Osteoporosis is a condition that arises when there's a decline in bone mineral density and mass, or when the composition and strength of bones change [4]. This weakening of bones can significantly elevate the risk of fractures. On the other hand, osteopenia is less severe than osteoporosis, it can be considered as the precursor of osteoporosis. It simply indicates that there is less bone mineral density than usual, but it is not yet seriously affecting the person. Additionally, often referred to as a "silent" ailment, osteoporosis and osteopenia typically progress without noticeable symptoms, and individuals may only become aware of it when they experience a bone break. This disease is a primary cause of fractures in older men and postmenopausal women, frequently occurring in bones like the hip, spine vertebrae, and wrist. It's important to note that osteoporosis doesn't discriminate based on race or ethnicity, affecting individuals from all backgrounds [5]. While the risk of developing osteoporosis increases with age, it can manifest at any stage of life.

According to the World Health Organization (WHO), Dual-energy x-ray absorptiometry (DXA) is a globally recognized standard for diagnosing osteoporosis and osteopenia through the measurement of bone mineral density (BMD) testing [6] using the T-score as shown in Table 1. The T-score is determined by taking the patient's bone mineral density (BMD) and subtracting the mean BMD (in g/cm<sup>2</sup>) of a reference population of young adults. This difference is then divided by the standard deviation (SD) of the young-adult reference group [7].

Diagnosis	Bone Mass Density Measurement	T-Score Range
Normal	$\mu: \sigma < 1$	$\geq -1$
Osteopenia	$\mu: 1 \leq \sigma < 2.5$	$-1 < -2.5$
Osteoporosis	$\mu: \sigma \geq 2.5$	$\leq -2.5$

Table 154 WHO diagnosis of osteoporosis and Osteopenia based on BMD measurements from DXA

Table 1 shows the classification of patients with the corresponding T score values. Samples with T-scores greater than -1 are diagnosed as normal. Likewise, if the calculated T-score falls between -1 and -2.5, it is designated as osteopenia. T-score values equal to or less than -2.5 are considered osteoporosis. Bone mineral density test results are reported in two forms: the T-score and the Z-score. The T-score represents bone density compared to what would normally be expected in healthy adults of the same sex. It also represents the number of units, called standard deviations, by which bone density is above or below the average. A Z-score is the number of standard deviations above or below what would normally be expected for a person of the same age and sex. If the Z-score is significantly higher or lower than average, additional testing may be needed to determine the cause of the problem [8].

Many studies have contributed to the identification and prediction of this medical disease, but most of them depended on specialized imaging technologies like DXA that may not be available in all healthcare facilities, which limits the general application of these predictive models. Additionally, earlier studies focused mostly on diagnosing the illness after the patient has experienced symptoms. As a response to these limitations, this paper aims to address the issue by employing available clinical data to train and evaluate several types of machine learning algorithms for the preemptive detection of osteoporosis and osteopenia without relying on bone mineral density test results that usually depend on using DXA scan, using only features such as gender, age, Body mass index, and if the patient has diabetes or not.

Over the past decade, the transformative influence of ML in healthcare has been undeniable, particularly in personalized medicine. ML models have demonstrated remarkable capabilities when dealing with complex and voluminous datasets, paving the way for tailored healthcare approaches. By harnessing ML techniques, researchers have unlocked a realm of potential within clinical studies of these conditions, offering exciting prospects for precision medicine in both research and practical healthcare settings [9]. Inspired by the potential of ML to revolutionize early disease identification [10], we are compelled to expand our work into previously uncharted territory.

Knee X-ray Osteoporosis Database from Shri Mata Vaishno Devi University is the dataset used in this study which is obtained from Mendeley Data website. The dataset includes the T-score values from each participant's knee X-ray and Quantitative Ultrasound System [11], together with clinical factors responsible for osteoporosis and osteopenia, such as age, gender, menopausal age, and other disorders including diabetes, fracture history, lifestyle factors, etc. The utilized ML algorithms are Random Forest (RF), Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Gradient Boosting (GBoost), and Extreme Gradient Boosting (XGBoost) which have been chosen after an extensive literature survey. Using 20 features only and SMOTETomek sampling, Random Forest outperformed with an accuracy of 91.11% with a precision, recall and F1 values of 92%, 91%, and 91% respectively. To enhance the transparency and trust in ML algorithms by interpreting the model's decision-making procedure, Explainable Artificial Intelligence (XAI) is applied [12]. Two techniques are used for model explanation and interpretation which are Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanation (SHAP). The remaining segment of this paper is structured as follows. Section 2 is focused on the exploration of literature relevant to the topic. Section 3 contains materials and methods demonstrating dataset, statistical analysis, data preprocessing, utilized ML algorithms, performance measures, and optimization strategies. Section 4 involves the empirical results and further discussions of the results. Section 5 presents an Interpretation of the Final Recommended Model using XAI while section 6 contains discussions. Finally, the conclusion, recommendation, and future work are in section 7.

## 2. Literature Review

Albuquerque et al. [13] proposed a study that aims to develop an automated system for the early detection of osteoporosis. The researchers utilized a dataset consisting of bone mineral density measurements from many patients, reaching 505. Approximately 21.8% of individuals are healthy, while the majority, approximately 78.2%, have low bone mineral density and are affected by osteoporosis. The dataset was collected through DXA scans, which are considered the gold standard for detecting osteoporosis. The methodology involved training a machine learning model on these DXA scan measurements to classify individuals as either osteoporotic or non-osteoporotic. A 5-fold cross-validation is performed, and 20% of the dataset is used for testing. By incorporating electromagnetic wave analysis, Random Forest is the best model, with a sensitivity of 0.853, a specificity of 0.879, and an F1 score of 0.859. The findings emphasize age, BMI, and Osseus' signal attenuation as the most critical criteria in the categorization of osteoporosis.

The effectiveness of the RF model, combined with Osseus measurements, supports early osteoporosis screening, leading to a reduction in costs and improving patients' quality of life.

Moreover, a study by Wu et al. [14], their study aims to create predictive models using ML algorithms to serve as screening mechanisms for identifying osteoporosis type 2 diabetes (T2DM) patients. The data was gathered from 433 participants and employed nine categorical ML algorithms to choose features that are based on clinical and demographic variables. They evaluated the models using a range of metrics, optimized them through 5-fold cross-validation, and assessed feature importance using Shapley additive explanations (SHAP). Furthermore, to identify distinct subgroups, they employed latent class analysis (LCA) within the dataset. ML algorithms had average precision scores ranging from 0.444 to 1.000. The final model, XGBoost, achieved AUROC values of 0.940 in training, 0.772 in 5-fold cross-validation, and 0.872 in testing. The study holds some limitations, such as the cross-sectional nature of this study, which introduces inherent limitations in conducting predictive analysis, and the study's sample size was inadequate to assess the model.

Another study written by Wu and Park [15], This study aimed to build a model to predict the osteoporosis risk using ML in adults over 40 in the Ansan/Anseong cohort and to analyze its association with fractures in the HEXA cohort. In the Ansan/Anseong cohort, 8,842 participants had 109 factors, including demographics, measurements, genetics, nutrition, and lifestyle, manually chosen and integrated into the ML algorithm. The data was randomly split into 80% training with 7,074 samples and 20% testing with 1,768 samples. Using 109 standardized variables, they select LoR, SVM, XGBoost, DT, RF, KNN, and DNN for predicting metabolic status models to improve ROC area, accuracy, and K-fold performance in testing. XGBoost, DNN, and RF produced a highly accurate prediction model with an AUC of 0.86 in the ROC using 10, 15, and 20 features. The top two models for accuracy were XGBoost and RF with 15 features, achieving 0.902 and 0.903, respectively. The study has some limitations; for example, the cross-sectional design measured biomarkers once, but the large sample size helps BMD measured with a peripheral densitometer, though less precise than DEXA, remain suitable for clinical osteoporosis risk assessment.

Inui et al. [16] aimed their study to develop an effective osteoporosis screening technique that did not require DXA scans. The authors employed a machine learning (ML) approach, utilizing patient body mass index (BMI), age, and blood test data as input features. The study included medical data from 2541 women who visited an osteoporosis clinic. Moreover, to identify and predict low BMD in patients, multiple ML models such as LR, DT, RF, Gradient Boosting Trees (GBT), and lightGBM were trained. Furthermore, among the trained models, lightGBM outperformed the other models with the highest accuracy of 83.4% and AUC of 96.1%, indicating its superior predictive performance for screening osteoporosis. The study has some limitations, including its focus on an outpatient clinic in Japan, its use of only female records, and the absence of imaging data like X-rays or CT scans, which could lead to different conclusions.

In a study by Ma et al. [17], the goal of their study is to determine whether various alternative ML models could give a prediction that is better than LoR models and to choose the best model. A retrospective analysis was performed on 529 patients who had Percutaneous kyphoplasty (PKP) at their institution between 6/2017 and 6/2020. They split the dataset into 75% for training and 25% for testing sets and then built the ML models after 10 cross-validations. Afterward, all models were assessed using the testing set, and their performance was evaluated by measuring the AUC for each model. Except for DT, ML algorithms

showed better performance than LoR, and among them, RF demonstrated the highest predictive capability. This study showed some limitations, such as the fact that retrospective studies can introduce selection and subjective bias, and the single-center study sample size remains relatively small.

Furthermore, the authors in [18] presented a deep learning approach using transfer learning with convolutional neural networks (CNNs) to classify knee X-ray images into normal, osteopenia, and osteoporosis disease groups. Using a dataset of 381 knee X-rays from a BMD camp held by the Unani and Panchkarma Hospital in Srinagar, J&K, India. Four CNN architectures were compared. The pretrained AlexNet architecture achieved the best accuracy of 91%, an error rate of 0.09, and a validation loss of 0.54 reducing fracture risk and aiding clinicians. For our study, we selected the same dataset that includes both imaging and clinical data. Nevertheless, our analysis will mainly use clinical data to increase prediction accuracy, in contrast to the image-based approach used in the prior study.

Fasihi et al. [19] conducted a study that aims to explore the use of artificial intelligence in diagnosing osteoporosis based on risk factors derived from clinical data for both women and men. Additionally, the authors propose sports protocols that can potentially mitigate the negative impacts of osteoporosis. The dataset was obtained from three hospitals in Tehran, Iran, specifically from the scanning center for DXA. It included clinical information from a total of 1224 individuals, both men and women. They used various machine learning models, such as RF, KNN, SVM, DT, AB, GB, ET, and MLP analysis, to predict osteoporosis and suggest suitable exercise programs for treatment. The dataset was separated into training (80%) and test (20%) sets, with the results evaluated on the test set. Based on the AUROC curve, the FR algorithm yielded the most accurate predictions for men, achieving an AUROC of 0.91. For women, the GB algorithm performed best with an AUROC of 0.95. In terms of exercise recommendation, the RF algorithm performed best in detecting exercise for healthy individuals as well as those with osteopenia and osteoporosis, with AUROCs of 0.96 and 0.99 in women and men, respectively.

Another study by Dzierzak and Omiotek [20], objective of this study was to determine whether DCNNs could be used to develop a reliable approach for detecting osteoporosis based on CT scans of the spine. The study's research dataset contains CT scans of the L1 spongy tissue from 100 participants, of whom 50 have osteoporosis and 50 are healthy. This study used six DCNN architectures that are pre-trained (MobileNetV2, Xception, InceptionResNetV2, VGG16, VGG19, and ResNet50) with various topological depths. The VGG16 model, which has the lowest topological depths, showed the highest results with 95% accuracy, 96% true positive rate, and 94% true negative rate. The study limitation was having a small dataset with 400 images, and to overcome this, they used pre-trained DCNN models.

In a study by Kwon et al. [21], the researchers aimed to develop a pre-screening method for early diagnosis of osteoporosis in postmenopausal Korean women using machine learning algorithms. The Korea National Health and Nutrition Examination Surveys were utilized to collect the data, which included 1431 postmenopausal women between the ages of 40 and 69. Feature selection techniques were used to identify 20 relevant features affecting osteoporosis. Moreover, three machine learning algorithms, AdaBoost, Gradient Boosting (GBM), and RF, were trained on three different models: A, checkup features, B, survey features, and C, both checkup and survey features. The evaluation process was conducted based on accuracy and the AUROC. Finally, the results of the evaluation showed that Model C had the highest accuracy rates for AdaBoost, GBM, and RF with scores of 84.9%, 82.9%, and 83.2%, respectively, and AUROC scores of 92.1%, 90.8%, and 91.9%, respectively. This study shows some limitations, such as limited data from cross-sectional observational studies and focus on women aged 40–69, which makes generalizing its findings to all Korean populations difficult.

Another study authored by Jang et al. [22], proposed the OsPor-screen model that diagnoses osteoporosis from Chest Radiographs. The model was tested in internal and external datasets. The internal dataset is 13,026 chest radiographs and DXA, while the external dataset is 1089 chest radiographs, both of which

were provided by Asan Medical Center. The model was trained using supervised learning techniques. The internal dataset test result was 82.40% accurate. On the other hand, the external dataset has an accuracy of 77.69%. The model was tested with a dataset that was collected from one center, and the performance of the model could be improved using electronic medical records (EMRs) clinical data. These limiting factors prevent the model from achieving better performance.

A research study was carried out by Jang et al. [23], the research paper introduces a deep neural network model (DNN) that predicts osteoporosis via simple hip radiography. The model was established to be a screening tool instead of using dual-energy X-ray absorptiometry (DXA). DXA is a tool to diagnose osteoporosis, but because of its high cost, it cannot be used widely. Consequently, the model proposed as a solution for this problem. The model used a dataset of a DXA for 1001 females that are over fifty-five years old. The accuracy that the model achieved was 81.2%. Yet, there was a limitation in this study, which was only including females in the study.

In another study by Ou Yang et al. [24], the authors proposed a machine-learning model that can predict the presence of osteoporosis in adults over fifty years old. The study aims to use the proposed model as a screening tool for osteoporosis in clinical practice. This will enable patients to be more alert regarding osteoporosis and prevent any serious complications due to it. The authors also desired to compare the results of this model with traditional methods that were used for predicting the disease. The dataset that was used in this study was from people who registered at the medical center in Taiwan. The dataset contains 5982 data points in total; 3053 of them are for men and 2929 are for women. Men were separated from women, and each had different results in accuracy. The algorithms used in the model are SVM, KNN, ANN, RF, and logistic regression (LoR). To identify the performance-best model Comparing the model's performance was done using AUROC. The model achieves 0.840, 0.821, 0.837, 0.843, and 0.827 for men and 0.807, 0.767, 0.781, 0.811, and 0.772 for women for SVM, KNN, ANN, RF, and LoF, respectively. Some of these algorithms show better performance than traditional methods. Thus, patients can gain an advantage by using the model to predict osteoporosis in the future.

Another study was conducted by Anam et al. [25], which aims to review the application of ML techniques in predicting osteoporosis for trabecular bone. In this review, the authors sought to analyze various datasets that have been used in studies related to osteoporosis prediction for trabecular bone using ML. The methodology provided a comprehensive summary of the ML techniques employed in osteoporosis prediction for trabecular bone. They reviewed several algorithms, such as SVM, RF, neural networks, and deep learning models. The authors emphasized the importance of feature selection and extraction techniques to enhance the accuracy of the predictions. Furthermore, they discussed the importance of cross-validation techniques to validate the performance of the models. The result reported notable achievements in predicting osteoporosis for trabecular bone using ML techniques. They highlighted that the developed models exhibited high accuracy rates when tested on the selected datasets. The review identified that certain algorithms, such as ensemble models and deep learning architectures, yielded superior results in comparison to traditional ML approaches.

The research paper by Yamamoto et al. [26] introduces a method to diagnose osteoporosis and classify it from hip radiographs. The authors used a deep learning algorithm which is CNN. Moreover, the authors wanted to test if adding clinical data to the image-based dataset they used would enhance the performance of the diagnosis and increase its accuracy. ResNet18, ResNet34, GoogleNet, EfficientNet b3, and EfficientNet b4 are the CNN models used to classify hip radiographs into osteoporosis and non-osteoporosis. Each model was assessed first on an image-based dataset containing 1131 hip radiograph

images and then with the same dataset in addition to clinical data. GoogleNet and EfficientNet b3 had the highest accuracy of 0.8407 for the dataset without clinical data. With the clinical data added, there was an improvement in the accuracy for each model; the highest was EfficientNet b3 with 0.8850.

The aim of this research paper, which was achieved by Yu et al. [27], is to establish and apply an ANN model to aid in the diagnosis of osteoporosis using a preprocessed dataset of patients diagnosed with osteoporosis, which consists of X-ray images, clinical symptoms, and demographic information, as well as bone mineral density (BMD) measurements. The researchers applied the MLP-based ANN model to the diagnostic system of osteoporosis. The MLP architecture consists of an input layer, hidden layers, and an output layer where the final diagnostic prediction is provided. The ANN model was trained using a backpropagation algorithm, and the hyperparameters were optimized to achieve the highest possible accuracy. The ANN model achieved a high level of accuracy in diagnosing osteoporosis (90%). The researchers validated the model using a separate test set, further confirming its effectiveness in accurately classifying osteoporosis cases.

A study by Yoo et al. [28], this paper aimed to develop and validate machine learning models to accurately identify the risk of osteoporosis in postmenopausal women. The models were developed using medical records from postmenopausal Korean women acquired as part of the Korea National Health and Nutrition Examination Surveys. Furthermore, prediction models were built using machine learning methods such as SVM, RF, LoR, and ANN. The efficacy of these machine learning models was compared to four traditional clinical decision tools: osteoporosis index of risk (OSIRIS), osteoporosis self-assessment tool (OST), simple computed osteoporosis risk estimation (SCORE), and osteoporosis risk assessment instrument (ORAI). Finally, compared to other approaches, SVM exhibited a much superior AUC of the receiver operating characteristic. SVM had an AUC of 82.7% and an accuracy of 76.7% for predicting osteoporosis risk. This study has some limitations due to its cross-sectional survey and difficulties in addressing drug effects as well as development for Korean women. It also has an imbalanced class distribution. Further research with large, diverse samples is required to validate the findings.

A review of the literature on early osteoporosis and osteopenia prediction indicated that most previous studies on the subject have concentrated on imaging datasets while ignoring the potential of clinical data. Moreover, earlier studies frequently encountered constraints, such as limited sample sizes and a lack of adequate clinical data. Therefore, the purpose of this work is to close this gap by using clinical data to construct a machine-learning model for reliable and more accurate osteoporosis prediction. Furthermore, the study will employ explainable AI (XAI) techniques to provide transparency and build trust in the predictive model, ensuring its effective adoption by medical practitioners across various healthcare settings.

### 3. Materials and Methods

This study developed a pre-emptive model for diagnosing osteoporosis and osteopenia using Python programming language. All operations were conducted with a fixed seed value of 0 to ensure repeatability of results and consistency across runs. As seen in Figure 1, the dataset went through a number of crucial pre-processing steps before modeling to ensure data integrity and feature quality. These steps included handling outliers, checking for duplication, and checking for missing values across all features. After that, LabelEncoder was used to convert the values of categorical columns into numerical representations for better model interpretability. Moreover, to address the class imbalance, the SMOTETomek technique was

used —a hybrid sampling method that combines the advantages of SMOTE (Synthetic Minority Over-sampling Technique) and Tomek links. With the use of Tomek links and SMOTE, it attempts to rectify class imbalance by undersampling the majority class and oversampling the minority class. The dataset was then divided into training and testing sets, using 80% for training and 20% for testing, and standardized using StandardScaler for optimal model performance. This method assumes that data is normally distributed and will scale them such that the distribution is centered around zero with a standard deviation of one. Feature selection was conducted using Sequential Forward Feature Selection (SFFS), a method that sequentially chooses the most suitable features for a machine learning model. Five different classifiers—Random Forest, SVM, KNN, Gradient Boosting, and XGBoost—were utilized. For each model, GridSearchCV was employed to tune hyperparameters via cross-validation using 10 folds. The best hyperparameters for each model were identified, and their performance was evaluated on the test set using accuracy, precision, recall, F1-score, and AUC. Following the selection of the model that performed the best, eXplainable Artificial Intelligence (XAI) techniques like SHAP and LIME were used to examine the model's decision-making process in more detail. This entire process is illustrated in Figure 1.

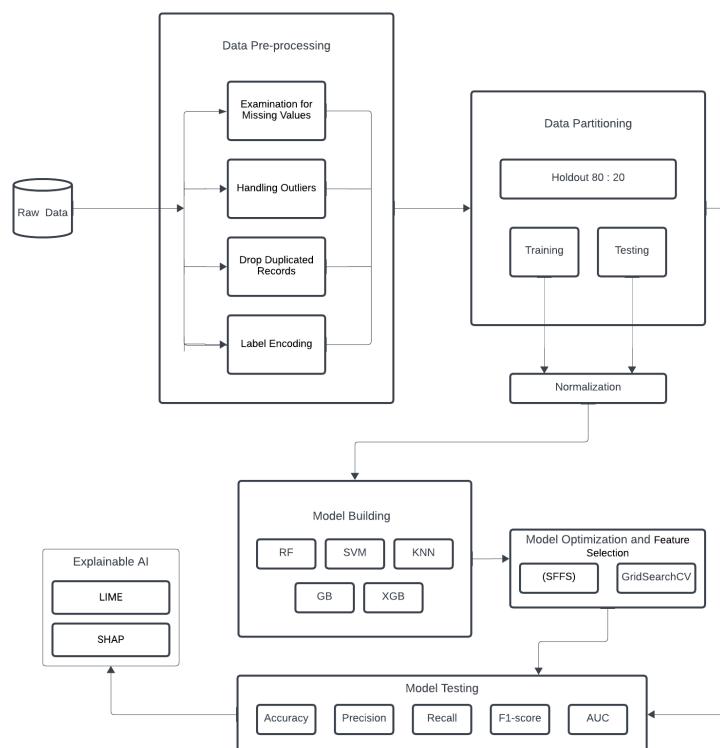


Figure 152 The proposed framework for the pre-emptive diagnosis of Osteoporosis and osteopenia

### 3.1. Data Description

The dataset utilized in this research originates from a health camp organized by the Unani and Panchkarma Hospital in Srinagar, Jammu \& Kashmir, India, held from December 21 to December 31, 2019 [11]. The dataset consists of 240 entries, each corresponding to an individual patient and encompassing 26 laboratory biomarkers and demographic features. Noteworthy features include Joint Pain, Gender, Age, Menopause Age, Height (Meter), Weight (KG), Smoking and Alcohol habits, Diabetes, Hypothyroidism, Number of Pregnancies, Seizure Disorder, Estrogen Use, Occupation, History of Fracture, Dialysis, Family History of Osteoporosis, Maximum Walking Distance (km), Daily Eating Habits, Medical History, BMI, Site of examination, Obesity status, and Diagnosis, considering each participant's T-score and Z-score data from their knee X-ray and Quantitative Ultrasound System. All features are illustrated in Table 2 along with

their respective types. This dataset provides a comprehensive snapshot of patients' health conditions, facilitating in-depth exploration and analysis for research purposes.

Feature	Type
Joint Pain	Categorical
Gender	Categorical
Age	Integer
Menopause Age	Integer
Height (Meter)	Float
Weight (KG)	Integer
Smoker	Categorical
Alcoholic	Categorical
Diabetes	Categorical
Hypothyroidism	Categorical
Number of Pregnancies	Integer
Seizure Disorder	Categorical
Estrogen Use	Categorical
Occupation	Categorical
History of Fracture	Categorical
Dialysis	Categorical
Family History of Osteoporosis	Categorical
Maximum Walking Distance (km)	Float
Daily Eating Habits	Categorical
Medical History	Categorical
T-score Value	Float
Z-Score Value	Float
BMI	Float
Site of examination	Categorical
Obesity status	Categorical
Diagnosis	Categorical

Table 155 Features' description.

### 3.2. Statistical Analysis

This section delves into a comprehensive statistical analysis pivotal for uncovering intrinsic data patterns and determining requisite preprocessing techniques for optimal model preparation. The dataset under scrutiny presents a mix of numerical and categorical features, each bearing significance in the subsequent modeling phase.

The numerical attributes underwent thorough scrutiny through various statistical metrics. These metrics provided profound insights into the inherent characteristics of the data. A detailed statistical breakdown showcasing the properties and attributes of the numerical features is presented in Table 3, offering a profound understanding of the dataset's numeric aspects.

Feature	Mean	Standard deviation	Min	25 <sup>th</sup> quartile	50 <sup>th</sup> quartile	75 <sup>th</sup> quartile	Max	Missing value counts
Age	51.13	13.23	17.00	44.50	50.00	60.00	107.00	0
weight	69.05	9.88	39.00	63.00	69.00	74.50	98.00	0

height	1.58	0.09	1.37	1.52	1.57	1.65	1.82	0
Maximum walking distance (km)	1.94	1.98	0.10	0.50	1.00	3.00	10.00	3
BMI	27.58	4.05	16.13	24.94	27.26	30.21	42.75	3

Table 156 Statistical analysis of numerical features

The categorical features as shown in Figure 2 across the dataset shed light on various patient attributes. The distribution of the target feature indicates that among 240 patients, 154 have osteopenia, 49 have osteoporosis, and 37 are considered normal. Gender distribution showcases 132 female and 108 male patients. Alcohol consumption shows that all the patients do not consume Alcohol. Smoking habits vary, with 199 non-smokers and 41 smokers. 12 patients are diabetic and 228 do not suffer from diabetes. Estrogen intake showcases 229 patients not taking supplements and 11 who do. Joint pain involves 2 patients who do not have any joint pain and 238 patients who suffer from joint pain. 34 patients have Hypothyroidism and 206 do not. Family history of osteoporosis comprises 174 patients with no family history and 66 with a positive history. 8 patients have Seizer Disorder and 232 do not. One patient is on dialysis and 339 are not. Menopause Age, Number of Pregnancies, Occupation, History of Fracture, Maximum Walking distance, Daily Eating habits, Medical History, Site, and Obesity complete the categorical feature analysis, providing a comprehensive view of patient demographics and health-related behaviors.

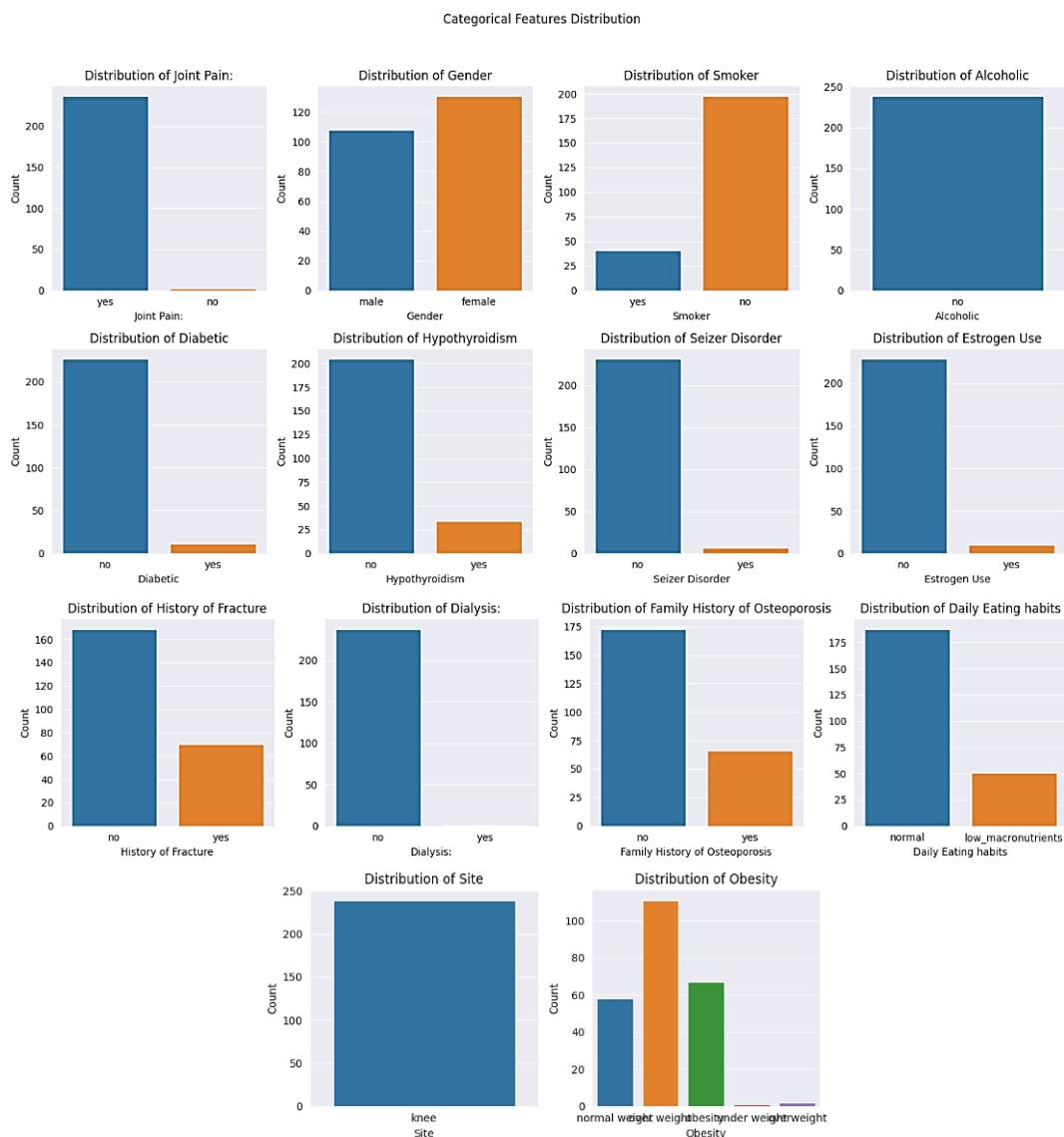


Figure 153 description of the categorical features

### 3.3 Data Pre-processing

The idea of data preprocessing is to transform unprocessed data into a clean data set. Before the dataset gets transmitted to the algorithm, it is preprocessed to look for missing values, noisy data, and other inconsistencies. Various Python libraries were used for data analysis and pre-processing in the current study. Initially, the data set contained 27 features and 240 records; T-Score values and Z-Score values weren't included in the training set since the DXA findings were used as the target. Columns that contain a significant number of null values were dropped. Moreover, cleaning the data involves removing duplication using the Pandas duplicated() method. Following preprocessing, the number of features was reduced to 20, and instances were limited to 239. Furthermore, identify and transform categorical values into numerical values using the LabelEncoder() method.

Identifying and managing missing values is essential; it is a challenge that arises during the data analysis phase and may produce misleading or inaccurate conclusions. Therefore, addressing this problem is crucial for preserving the quality of the data and for better model performance. Missing values can be handled in many ways, but the most common ones are the removal of incomplete entries or the imputing of reasonable values to fill in the gaps. Deletion and imputation are the most widely used approaches for dealing with missing data [29]. In this study, by examining the distribution of numerical null variables, if a variable has a normal distribution, utilize the mean value, as shown in Equation 1, where n represents the total number of values in a column and X represents a single data point [30]. If not, replace missing values in skewed data sets with the median value, as shown in Equation 2, where n refers to the total number of observations.

$$\text{Mean} = \frac{\sum X}{n} \quad (1)$$

$$\text{Median} = \frac{(n+1)}{2} \quad (2)$$

After that, due to the Data Imbalanced oversampling and undersampling techniques were used to balance the distribution of the classes. Using the SMOTETomek method it performs an oversampling technique using SMOTE followed by the undersampling technique using Tomek Links. SMOTE is done by identifying data from the minority class then finding its K nearest neighbor, creating new Synthetic data between the point and its neighbor, repeat until the minority and majority classes are balanced as desired. Tomek Links are points from different classes that are close to one another, and for each pair, the points from the majority class will be removed to eliminate any points that are unclearly near the class decision boundary. This technique will improve model performance and decision-making process [31]. Following that StandardScaler was used to scale down the training and testing features with a zero mean and unit variance, Equation 3 shows the formula used for StandardScaler, by subtracting the mean of the feature value and dividing the result by the standard deviation [32].

$$Z = \frac{X-\mu}{\sigma} \quad (3)$$

### 3.4 Description of Utilized Machine Learning Algorithms

#### 3.4.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) was proposed by Cortes and Vapnik in 1990 and has since gained popularity within the machine learning community [33]. Operating as a supervised learning algorithm, SVM addresses both classification and regression problems, with a primary focus on binary classification [34]. SVM establishes a hyperplane in the feature space to separate different classes, ensuring a maximum margin between them [35]. During testing, data points are mapped onto the feature space and categorized based on their position relative to the margin, showcasing SVM's effectiveness in creating robust decision boundaries.

### **3.4.2 Random Forest (RF)**

Random Forest, introduced by Leo Breiman in 2001, revolutionized ensemble learning through the concept of bagging or "bootstrap aggregation [36]." Comprising multiple Decision Trees (DTs), the RF classifier avoids overfitting by aggregating predictions through a majority vote mechanism. Instead of relying on a single DT, RF leverages the collective insights from various trees, enhancing predictive accuracy. This approach, rooted in diversity, makes Random Forest a powerful tool for classification tasks [37].

### **3.4.3 k-Nearest Neighbors (KNN)**

k-Nearest Neighbors (KNN) is a straightforward yet effective supervised machine learning algorithm that emerged in the field of pattern recognition. The KNN algorithm was introduced by Evelyn Fix and Joseph Hodges in 1951 as a non-parametric method for pattern recognition. It gained further prominence and formalization in the 1960s and 1970s [38]. Operating on the principle of proximity in feature space, KNN classifies new data points based on the majority vote of their  $k$ -nearest neighbors. While intuitive and easy to understand, KNN may face challenges in high-dimensional spaces due to the "curse of dimensionality," where the effectiveness of the algorithm decreases as the number of features increases. Despite this limitation, its strength lies in its adaptability to various datasets and simplicity in implementation, making it a valuable tool in many machine-learning applications [39].

### **3.4.4 Extreme Gradient Boosting (XGBoost)**

Extreme Gradient Boosting (XGBoost) emerged in 2011, introduced by Carlos Guestrin and Tianqi Chen. Continuously optimized for modern data science challenges, XGBoost is a boosting tree-based learning framework renowned for its scalability, parallelizability, and superior performance. By combining multiple models, XGBoost creates a robust ensemble that frequently outperforms competing algorithms [40]. The regularization techniques employed in XGBoost effectively control overfitting, contributing to its high level of performance [41].

### **3.4.5 Gradient Boosting (GBoost)**

Gradient Boosting, a powerful ensemble learning technique, was introduced by Jerome Friedman in 1999. The specific algorithm, known as the Gradient Boosting Machine (GBM), was later developed by Friedman, Trevor Hastie, and Robert Tibshirani. GBM constructs a predictive model through an ensemble of weak learners, typically decision trees. The algorithm sequentially adds trees to correct errors made by the existing ensemble, resulting in a strong predictive model. Gradient Boosting excels in capturing complex relationships within the data, making it suitable for regression and classification tasks. Its ability to handle diverse datasets and improve predictive accuracy over iterations has contributed to its widespread use in various machine-learning applications [42].

## **3.5 performance measure**

Performance evaluation for the model involves assessing the model's ability to correctly identify the patient state from the dataset. Thus, a range of performance metrics were used in this study to evaluate the models' performance. Including Accuracy, Precision, Recall, and F1-score. Confusion matrices, which contain True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), were utilized for further model evaluation.

- **TP:** specifies the patients who were accurately classified as osteoporosis, osteopenia (low bone mineral density), or normal.
- **TN:** specifies the patients who were accurately classified as not belonging to one of the three classes.
- **FP:** specifies the patients who were classified as osteoporosis, osteopenia (low bone mineral density), or normal but actually belong to another class.

- **FN:** specifies the patients who were classified as not being part of osteoporosis, osteopenia (low bone mineral density), or normal class but actually belong to that class.

Accuracy measures the correctly predicted patients' state (osteoporosis, osteopenia low, or normal) over the total number of patients in the dataset. Equation 4 provides a mathematical representation of it.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Precision for a given class is defined as the ratio of the number of patients accurately classified to that class (TP), to the sum of positive instances in that class (TP and FP). Equation 5 provides a mathematical representation of it.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Recall for a given class is defined as the ratio of the number of patients accurately classified to that class (TP), to the sum of TP and the number of patients who were from this class but misclassified to another (FN). Equation 6 provides a mathematical representation of it.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

F1-score is a combination of both Precision and Recall. Equation 7 provides a mathematical representation of it.

$$F1 - score = \frac{2(Precision \times Recall)}{Precision + Recall} \quad (7)$$

### 3.6 Optimization strategy

To get the best possible performance for the models, hyperparameters were utilized. Proper hyperparameter tuning is often a critical step in achieving optimal model performance. To get the suitable hyperparameter for each algorithm the GridSearchCV method was employed. The method was given a set of different values for each hyperparameter, it then generates the best combination of these values by using the training set. Table 4 illustrates the algorithms and their best hyperparameter with the original data.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	10
	Min_samples_leaf	1
	Min_samples_split	5
	N_estimators	100
SVM	Kernel	linear
	C	1
	gamma	scale
K-NN	N_neighbors	7
	algorithm	auto
	weights	uniform
Gboost	Learning_rate	0.01
	Max_depth	3
	N_estimators	200
XGboost	Gamma	0.1
	Learning_rate	0.01
	Max_depth	3
	N_estimators	100

Table 157 the optimal hyperparameter for each classifier in the original data

Table 5 represents the algorithms with their best hyperparameters in oversampled data and 8 features selected.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	none
	Min_samples_leaf	1
	Min_samples_split	2
	N_estimator	100
SVM	Kernel	linear
	C	10
	gamma	scale
K-NN	N_neighbors	7
	algorithm	auto
	weights	distance
Gboost	Learning_rate	0.1
	Max_depth	7
	N_estimator	100
XGboost	Gamma	0
	Learning_rate	0.1
	Max_depth	7
	N_estimator	200

Table 158 the optimal hyperparameter for each classifier in over-sampled data with 8 features selected

Table 6 represents the algorithms with their best hyperparameters in oversampled data and the number of features that have been selected is 10.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	none
	Min_samples_leaf	1
	Min_samples_split	5
	N_estimator	200
SVM	Kernel	linear
	C	10
	gamma	scale
K-NN	N_neighbors	7
	algorithm	auto
	weights	distance
Gboost	Learning_rate	0.1
	Max_depth	7
	N_estimator	150
XGboost	Gamma	0
	Learning_rate	0.1
	Max_depth	7
	N_estimator	150

Table 159 the optimal hyperparameter for each classifier in over-sampled data with 10 features selected

Table 7 represents the algorithms with their best hyperparameters in oversampled data and 15 features have been selected.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	none
	Min_samples_leaf	1
	Min_samples_split	5

	N_estimator	100
SVM	Kernel	rbf
	C	10
	gamma	scale
K-NN	N_neighbors	7
	algorithm	auto
	weights	distance
Gboost	Learning_rate	0.1
	Max_depth	7
	N_estimator	200
XGboost	Gamma	0
	Learning_rate	0.1
	Max_depth	5
	N_estimator	100

Table 160 the optimal hyperparameter for each classifier in over-sampled data with 15 features selected

Table 8 represents the algorithms with their best hyperparameters in oversampled data and using all the features.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	none
	Min_samples_leaf	1
	Min_samples_split	5
SVM	N_estimator	150
	Kernel	rbf
	C	10
K-NN	gamma	scale
	N_neighbors	5
	algorithm	auto
Gboost	weights	distance
	Learning_rate	0.1
	Max_depth	5
XGboost	N_estimator	150
	Gamma	0
	Learning_rate	0.1
	Max_depth	3
	N_estimator	200

Table 161 the optimal hyperparameter for each classifier in over-sampled data with all the features selected

#### 4. Empirical Results

Within this section, we present the outcomes derived from the developed models after the implementation of GridSearchCV on both the original dataset and the sampled data using SMOTETomek, a method used to balance the class distribution within the dataset by oversampling the minority class and undersampling the majority class. Following the acquisition of optimal hyperparameters and model training through stratified 10-fold cross-validation. Utilizing all the 20 features. The ensuing section delineates a comprehensive examination of the results obtained, as presented in Table 9.

Classifier	Dataset	Testing Accuracy	Precision	Recall	F1-Score
Random Forest	Original	70.83%	71.00%	71.00%	70.00%
	Using SMOTETomek	<b>91.11%</b>	<b>92%</b>	<b>91%</b>	<b>91%</b>
	SMOTETomek				
Gradient Boosting	Original	73.61%	67.00%	69.00%	68.00%
	Using SMOTETomek	87.77%	88.00%	88.00%	88.00%
	SMOTETomek				
SVM	Original	73.61%	76.00%	74.00%	74.00%
	Using SMOTETomek	87.77%	88.00%	88.00%	87.00%
	SMOTETomek				
XGBoost	Original	68.05%	74.00%	68.00%	67.00%
	Using SMOTETomek	86.66%	87.00%	87.00%	86.00%
	SMOTETomek				
K-NN	Original	69.44%	67.00%	67.00%	68.00%
	Using SMOTETomek	71.11%	71.00%	71.00%	68.00%
	SMOTETomek				

Table 162 The results of the proposed models before and after sampling was applied

The original results highlighted the need to address the class imbalance, as demonstrated by the enhanced performance of all models following SMOTETomek's implementation. However, substantial improvements were observed in all algorithms after using SMOTETomek. Testing accuracies soared, demonstrating a notable improvement in predictive power. Moreover, the use of hyperparameter tuning and feature selection contributed to more robust models. Random Forest emerged as the algorithm with the highest testing accuracy, achieving 91.11%, precision, recall and F1 values are 92%, 91%, and 91% respectively. Gradient Boosting and SVM both followed with a testing accuracy of 87.77%. Gradient Boosting with precision, recall and F1 values are all 88%. SVM with precision, recall and F1 values are 88%, 88%, and 87% respectively. These models have shown significant improvements in predictive accuracy, demonstrating the usefulness of SMOTETomek in managing class imbalance and boosting the models' capacity to predict osteoporosis and osteopenia, which may facilitate early detection and intervention. Figure 3 shows the number of samples of each class before and after applying SMOTETomek.

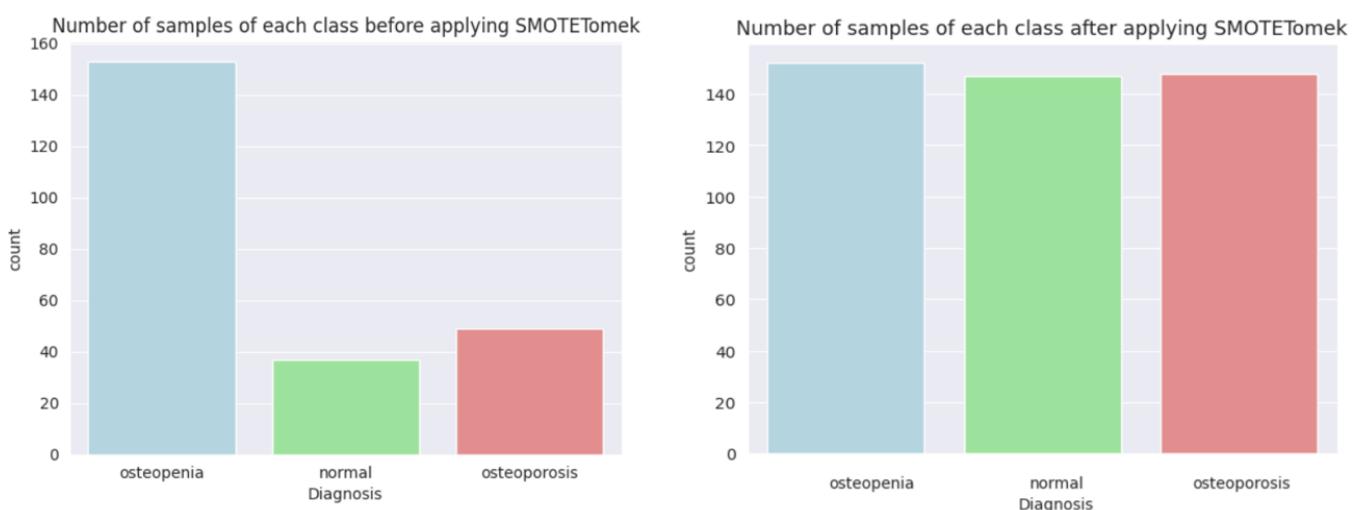


Figure 154 Number of samples of each class before and after applying SMOTETomek

## 4.1 Results with Feature Selection

Sequential Forward Feature Selection (SFFS) represents a strategic methodology employed to optimize the performance of machine learning models by iteratively selecting informative subsets of features. In this study, the SFFS technique was applied to a dataset comprised of 20 features extracted from clinical data gathered during the BMD camp conducted by the Unani and Panchkarma Hospital in Srinagar, J&K, India [11]. The objective was to ascertain the impact of varying feature subsets on model accuracy across different machine-learning algorithms. Sequential Forward Feature Selection (SFFS) is an iterative technique used in feature selection, initiates with an empty set, and systematically enriches subsets of features by adding one feature at a time, guided by their impact on boosting the model's predictive accuracy [43]. Using a selected metric, SFFS assesses various feature combinations at each iteration, retaining the combination that exhibits the greatest improvement. This process keeps going until it reaches a predetermined end point, like reaching a certain number of features or reaching the point where performance improvement is no longer possible. By choosing a subset that maximizes prediction accuracy, SFFS seeks to discover and incorporate the most valuable features, improving the model's predictive power and reducing processing requirements. Table 10 presents several feature subsets identified through SFFS. These subsets are examined to gauge their influence on model performance when paired with optimal hyperparameters and the SMOTETomek technique.

Feature Subsets	RF	SVM	KNN	GB	XGBoost	Features selected
<b>8</b>	87.77 %	86.66%	85.55%	88.88 %	87.77%	Joint Pain, Gender, Age, Diabetic, Seizer Disorder, Dialysis, Site, age_group
<b>10</b>	87.77 %	88.88%	85.55%	87.77 %	87.77%	Joint Pain, Gender, Age, Alcoholic, Diabetic, Seizer D isorder, Estrogen Use, Dialysis, Site, age_group
<b>15</b>	87.77 %	88.88%	81.11%	85.55 %	88.88%	Joint Pain, Gender, Age, Weight, Smoker, Alcoholic, Diabetic, Hypothyroidism, Seizer Disorder , Estrogen Use, Dialysis, Daily Eating habits, Site, Obesity,age_group
<b>All feature</b>	91.11 %	87.77%	71.11%	87.77 %	86.66%	Joint Pain, Gender, Age, Height,Weight, Smoker, Alcoholic, Diabetic, Hypothyroidism, Seizer Disorder , Estrogen Use, History of Fracture, Maximum Walki ng distance, Dialysis, Family History of Osteoporosis, Daily Eating habits, BMI, Site, Obesity,age_group
<b>s(20)</b>						

Table 163 Testing Accuracy for Different Feature Subsets

The results showed varying performances of all the five machine learning algorithms, with different feature subsets. Random Forest achieve its highest accuracy using all 20 features with a testing accuracy of 91.11%, SVM, Gradient Boosting, and XGBoost maintains stable accuracy across different feature subsets. Additionally, K-NN shows a decrease in accuracy as the number of features increases.

## 4.2 Further Discussion of the Results

RandomForest's confusion matrix showcases a robust performance across all classes as shown in Table 11, with 30 instances correctly predicted as osteopenia, 24 instances as normal, and 28 instances as osteoporosis, demonstrating fewer misclassifications compared to other models. Following that, SVM in Table 12 demonstrates a strong performance with 30 instances correctly predicted as osteopenia, 22 instances as normal, and 27 instances as osteoporosis, displaying fewer misclassifications overall, particularly between class normal and class osteoporosis. Table 13 with GradientBoosting exhibits a reasonably balanced performance, with 29 instances correctly predicted as osteopenia, 24 instances as normal, and 26 instances as osteoporosis. XGBoost in Table 14 displays a similar prediction to

GradientBoosting, correctly predicting 29 instances as osteopenia, 23 instances as normal, and 26 instances as osteoporosis. However, KNN demonstrates comparatively more misclassifications across all classes as shown in Table 15, correctly predicting 26 instances as osteopenia, 11 instances as normal, and 27 instances as osteoporosis, highlighting its relatively weaker performance among the models considered.

RandomForest		Predicted		
		osteopenia	normal	osteoporosis
Actual	osteopenia	30 (TP)	0 (FN)	0 (FN)
	normal	3 (FP)	24 (TN)	4 (FN)
	osteoporosis	0 (FP)	1 (FP)	28 (TN)

Table 164 RandomForest Confusion

GradientBoosting		Predicted		
		osteopenia	normal	osteoporosis
Actual	osteopenia	29 (TP)	1 (FN)	0 (FN)
	normal	4 (FP)	24 (TN)	3 (FN)
	osteoporosis	0 (FP)	3 (FP)	26 (TN)

Table 13 GradientBoosting Confusion Matrix

SVM		Predicted		
		osteopenia	normal	osteoporosis
Actual	osteopenia	30 (TP)	0 (FN)	0 (FN)
	normal	4 (FP)	22 (TN)	5 (FN)
	osteoporosis	0 (FP)	2 (FP)	27 (TN)

Table 12 SVM Confusion Matrix

KNN		Predicted		
		osteopenia	normal	osteoporosis
Actual	osteopenia	26 (TP)	4 (FN)	0 (FN)
	normal	7 (FP)	11 (TN)	13 (FN)
	osteoporosis	1 (FP)	1 (FP)	27 (TN)

Table 15 KNN Confusion Matrix

XGBoost		Predicted		
		osteopenia	normal	osteoporosis
Actual	osteopenia	29 (TP)	1 (FN)	0 (FN)
	normal	3 (FP)	25 (TN)	5 (FN)
	osteoporosis	0 (FP)	3 (FP)	26 (TN)

Table 14 XGBoost Confusion Matrix

Figure 4 shows the Receiver Operating Characteristics (ROC) curve analysis for RF, GradientBoosting, SVM, KNN and XGBoost, which is a valuable tool for evaluating the performance of classification models across different classes. It illustrates the trade-off between a true positive rate (sensitivity) and a false positive rate (1-specificity) across different thresholds. RandomForest and GradientBoosting both indicate strong discrimination ability across all classes, with Class 0, Class 1, and Class 2 showcasing AUC values of 0.98, 0.91, and 0.96 respectively. These results underline the effective discrimination capabilities of both GradientBoosting and RandomForest models across different classes, showcasing their aptitude in classification tasks. SVM showcases strong discrimination for Class 0 (0.97) and Class 2 (0.94) but slightly lower for Class 1 (0.86). KNN demonstrates moderate discrimination, with Class 0 (0.94) having the highest AUC, followed by Class 2 (0.89) and Class 1 (0.71). Moreover, XGBoost displays high discrimination ability across all classes (Class 0: 0.97, Class 1: 0.90, Class 2: 0.95). These results outline the varying discrimination capacities of each model across different classes in classification tasks.

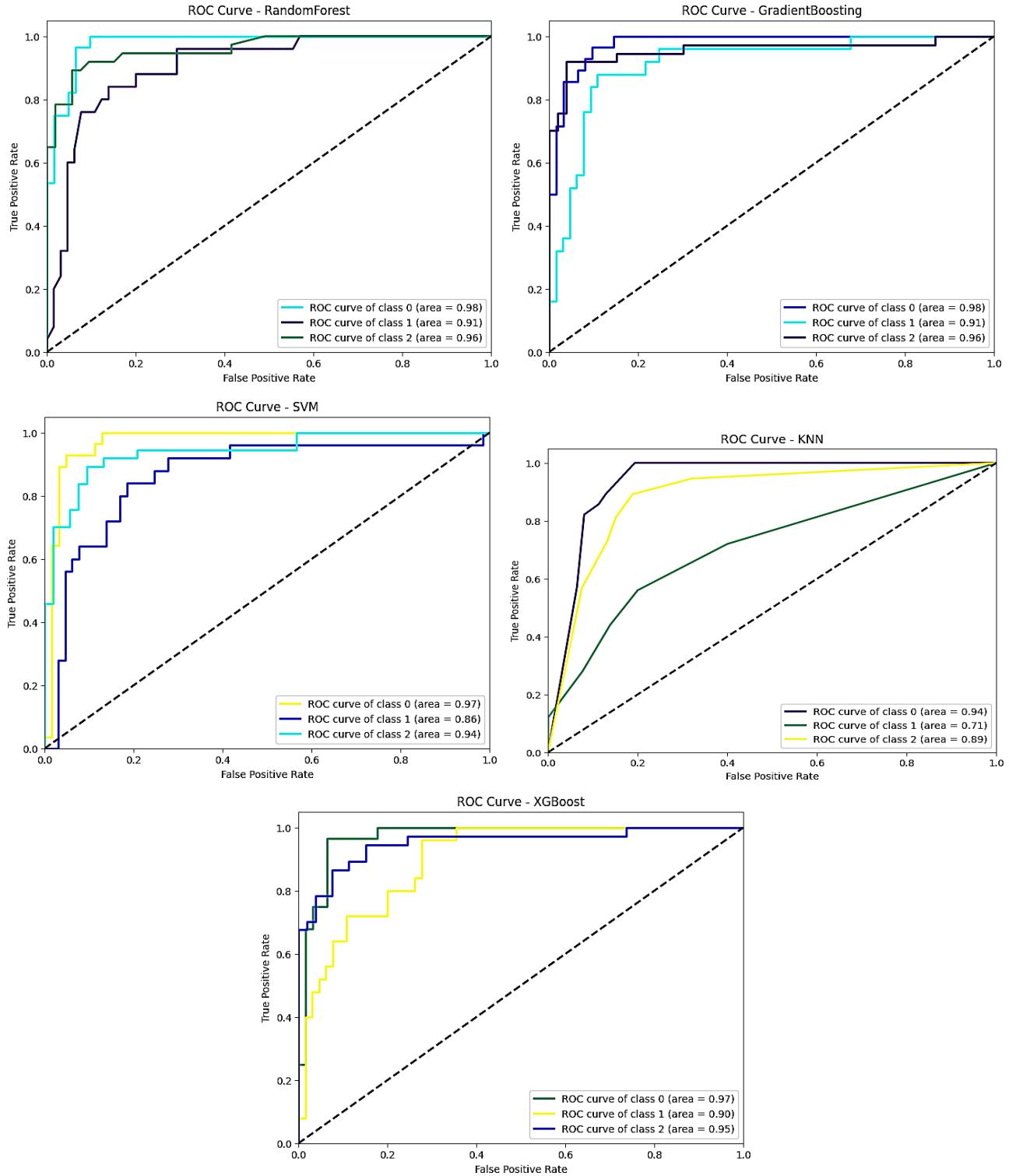


Figure 155 RF, GradientBoosting, SVM, KNN and XGBoost ROC Curve

## 5.1 Interpretation of the Final Recommended Model

AI finds widespread application in sophisticated fields, but the ambiguous nature of many AI models presents challenges in terms of understanding and trust. The results often resemble black boxes, which necessitates a clear understanding of the rationale behind the decision-making processes in AI models. Hence, the integration of explainable artificial intelligence (XAI) methods has gained importance to enhance trust and transparency in interpretations of AI models. [44] Interpreting the final model recommended in machine learning osteoporosis prediction systems becomes particularly crucial in this

context. In healthcare, where early and accurate detection is vital, XAI serves as a critical mechanism to elucidate factors that influence predictions, enhancing understanding among healthcare practitioners and patients regarding risk factors and key features that shape model outcomes. In this work, two XAI techniques were used which are LIME and SHAP.

### 5.1.1. Shapley Additive Explanation (SHAP)

The Shapley Additive Explanations methodology serves as a valuable tool in machine learning (ML) that addresses the black-box nature of many models, and it is commonly used in the domain of healthcare. [45] It evaluates feature importance by calculating the influence of each attribute on predictions and assigning SHAP values accordingly. The procedure assigns a score to each feature in each data point, indicating its contribution to the model's prediction, using Shapley values from cooperative game theory, which assigns collective effort value to each participant. [46] Consequently, this approach enhances interpretability in ML by offering a more detailed understanding of how each feature affects model predictions. The strength of the SHAP lies in its consistency and ability to provide a global view. It not only explains individual predictions but also provides information about the model as a whole. Fig. 5 shows the SHAP values from all classes are combined to create a summary plot that demonstrates how important it is of each of the 20 features, (a) shows the plot according to the mean SHAP values, where the labels 0, 1, and 2 represent osteopenia, normal, and osteoporosis. The feature importance plot is made by stacking the effects of a feature on the classes. The multiclass classification summary plot demonstrates what the machine was able to infer from the features. As observed, Age is the most important feature that contributed to the prediction of osteoporosis with the longest bar. Followed by Obesity and Family History of Osteoporosis. (b) shows the importance of the features and how they affect the model. It arranges features according to the total magnitudes of SHAP values across all samples. Additionally, SHAP values are used to display the distribution of each feature's impacts. Red denotes a high feature value and blue denotes a low feature value. As shown, Age has the highest impact on the model Followed by Family History of Osteoporosis and Site.

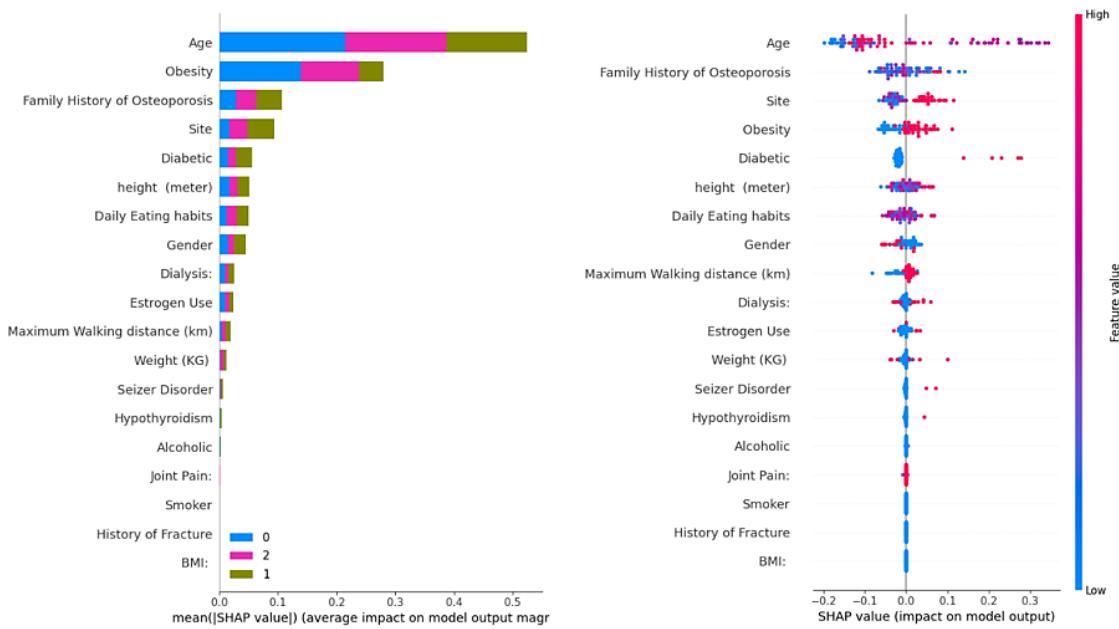
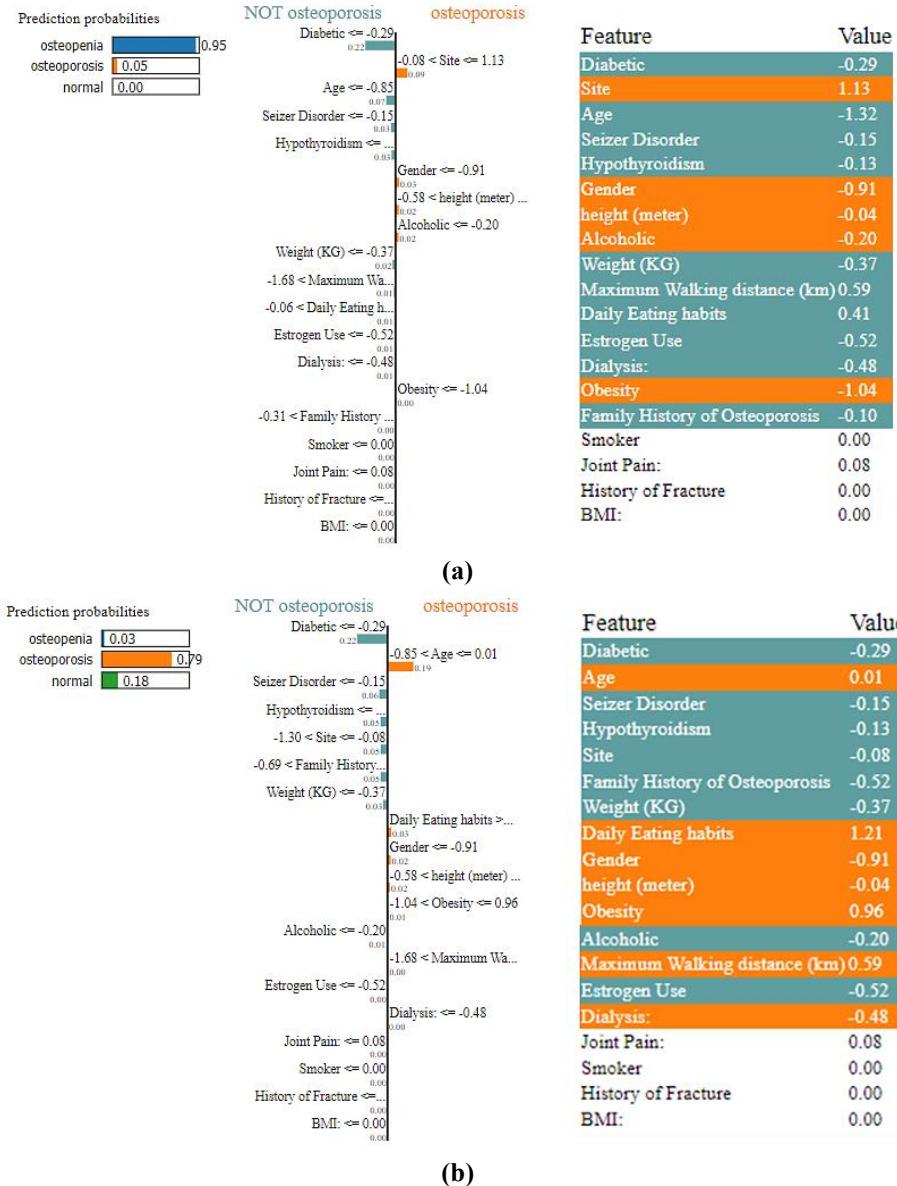


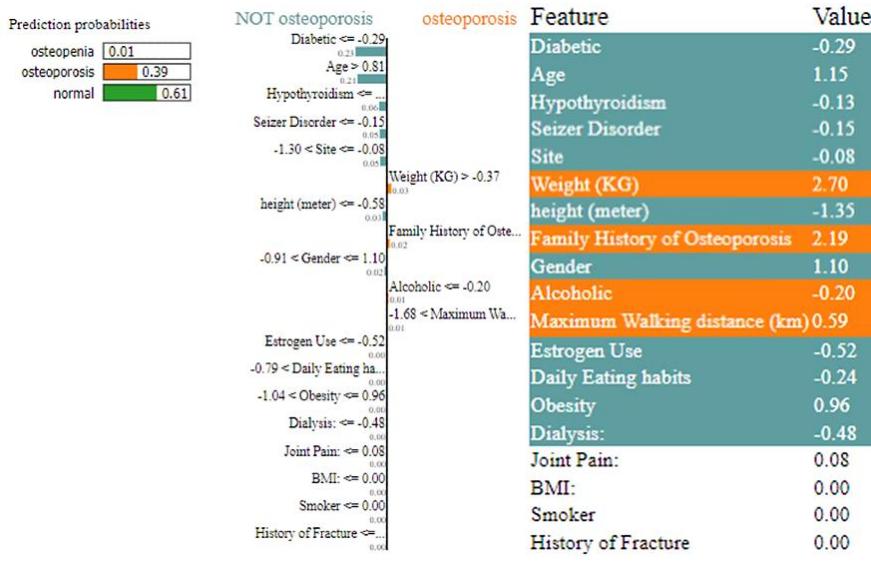
Figure 156 Shapely values using the Gradient Boosting model (a): SHAP bar plot, (b) SHAP beeswarm plot.

### 5.1.2. Local Interpretable Model-Agnostic Explanations (LIME)

LIME is a technique that uniquely addresses the challenge of explaining predictions from complex black box models by generating local and understandable explanations for individual cases. [47] It provides insight into the model's decision-making process by altering the input data, creating an alternative model, and assessing the relevance of features via random perturbation. LIME's model-agnostic nature makes it

suitable for various models after training, contributing to AI application transparency and confidence. [48] Figure 7 shows the LIME plot that gives local prediction probabilities for osteopenia, normal, and osteoporosis using the Gradient Boosting Classifier. It calculates the prediction by contrasting the probability values with the target variable. Each characteristic in the feature value table has a different color code to indicate how much it contributes to the prediction; for example, osteopenia is colored blue, while osteoporosis is colored orange, and normal is colored green. Using various weighted feature values, the features and associated values that contributed to the model's prediction are also emphasized.





(c)

Figure 157 Lime prediction probability for the Random Forest model a) osteopenia b) osteoporosis c) normal.

## 6. Discussion

The rapid strides in technology over the last few decades have heralded a promising revolution in healthcare, with emerging technologies, notably Machine Learning (ML), playing a pivotal role. ML's true value becomes evident in its adeptness at interpreting the vast troves of healthcare data generated daily through electronic health records. This capability empowers healthcare providers to not only enhance but also expedite care delivery by scrutinizing a broader spectrum of data [49]. The deployment of ML models within healthcare systems has demonstrated significant potential in automating primary and tertiary healthcare processes. This automation, coupled with the introduction of intelligent decision-making techniques, holds the promise of reducing medical testing costs and time, thereby optimizing resource utilization. Moreover, the ripple effects of these technological advancements extend to improvements in average life expectancy. As ML algorithms contribute to more precise diagnostics, personalized treatment plans, and efficient preventive measures, the overall health outcomes for individuals and communities stand to improve. The evolving landscape of healthcare, driven by ML innovations, invites us to envision a future where the amalgamation of data-driven insights and intelligent automation continues to elevate the standards of care delivery [50].

The rise in the prevalence of osteoporosis over the last two decades constitutes a concern to public health, particularly in Saudi Arabia, where the condition is present in 37.8% of men and 28.2% of women over the age of 50 [3]. Since osteoporosis can cause disability and death, society needs to give more consideration to it to avoid unintended consequences. Consequently, pre-emptive diagnosis of osteoporosis using ML may help to reduce the possible risks.

Many studies have been conducted related to the pre-emptive diagnosis of osteoporosis using ML techniques, some of these studies relied in their research on a database of images [19,20,22,23,25,26,27], which gave satisfactory results. However, the high cost of diagnostic imaging equipment constitutes an obstacle to the growth of the market in Saudi Arabia which is valued at USD 399.92 million in 2024 and is expected to reach USD 483.06 million by 2029 [51]. Therefore, taking advantage of clinical data becomes extremely important considering the increase in chronic diseases in said Arabia by developing a diagnostic model to detect osteoporosis using clinical data.

This study focuses on improving early detection of osteoporosis and osteoporosis using machine learning methods that simply rely on clinical data, thus eliminating the necessity of bone mineral density (BMD) testing. This paper investigates the performance of five machine learning algorithms using a dataset from a BMD camp in Srinagar, India, which included 240 patients. The Random Forest method outperformed

the other algorithms by using 20 features and achieving 91.11% accuracy, 91% recall, 92% precision, and 91% F1 score.

Osteoporosis can be caused by various features such as Gender, Age, smoking, and Body mass index. In our study Age, is the most important feature as well as in other studies. According to [14], Age is one of the main factors in the prediction of osteoporosis. The study showed that the third decade of life is when an individual reaches their maximal bone mass. Bone mineral density (BMD) generally decreases after a stability phase; around age 50, the reduction is 0.5–1% annually. As a person ages, the risk of developing osteoporosis and osteopenia increases, as bones weaken with increasing age. Another important feature is whether the patient have diabetes or not. Diabetes causes patients many health problems, and one of the most prominent of these diseases is the problem of osteoporosis. Based on [28] findings diabetes is an important feature of the predictors of osteoporosis, and it has been linked to low bone density.

This study concentrates on using medical dataset in the process of predicting osteoporosis disease. This gives a great benefit in facilitating the detection process and reducing costs. Most studies focused on using image-based datasets, using images of bone density to detect the disease might not be efficient since these kinds of medical examinations are not available everywhere in addition to being expensive.

## 7. Conclusion

Osteoporosis is a chronic disease defined by a loss of bone mass and mineral density, which makes bones weaker and vulnerable to fractures. Even though osteoporosis is frequently asymptomatic in its early stages, people may only become aware of it when they break a bone because of its "silent" nature. Even though several studies were carried out to detect osteoporosis early on, the majority of them relied on imaging technology to gather information that not all hospitals would have access to. This research attempts to solve these constraints by training and evaluating a range of machine learning algorithms for the proactive detection of osteoporosis using freely available clinical data. Random Forest achieved the highest result of 91.11% accuracy, 91% recall, 92% precision, and 91% F1 score with 20 features using the SMOTETomek sampling technique. Moreover, XAI is used to interpret the decision-making process of the final recommended model by employing SHAP and LIME. The most important feature of the Random Forest model is "Age" followed by "Obesity" and "Family History of Osteoporosis", which is the result of SHAP.

In future work, the model will be tested in a new dataset or new patients, additionally exploring new classifiers and methods like deep learning to make the accuracy higher. We recommend using the model in patients by using Longitudinal Data Analysis where we collect the patient's data continuously.

## References

- [1] "Non communicable diseases," World Health Organization, <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases> (accessed Oct. 8, 2023).
- [2] M. Heine and S. Hanekom, "Chronic disease in low-resource settings: Prevention and management throughout the continuum of care-a call for papers," International journal of environmental research and public health, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9966920/#:~:text=It%20is%20well%20described%20that,individuals%20with%20multimorbidity%20%5B7%5D> (accessed Oct. 8, 2023).
- [3] "National Plan for Osteoporosis Prevention and Management in the Kingdom of Saudi Arabia," Ministry of Health, <https://www.moh.gov.sa/en/Ministry/MediaCenter/Publications/Documents/NPOP-2018.pdf> (accessed Oct. 8, 2023).
- [4] "Osteoporosis," National Institute on Aging, <https://www.nia.nih.gov/health/osteoporosis> (accessed Oct. 8, 2023).
- [5] "Osteoporosis," National Institute of Arthritis and Musculoskeletal and Skin Diseases, <https://www.niams.nih.gov/health-topics/osteoporosis#:~:text=Osteoporosis%20in%20Men->

- ,Osteoporosis%20is%20a%20bone%20disease%20that%20develops%20when%20bone%20mineral,Pre gnancy%2C%20Breastfeeding%2C%20and%20Bone%20Health (accessed Oct. 8, 2023).
- [6] K.N. Haseltine, T. Chukir, P.J. Smith, J.T. Jacob, J.P. Bilezikian, A. Farooki, Bone mineral density: clinical relevance and quantitative assessment, *J. Nucl. Med.* .62 ,454–446 (2021) <https://doi.org/10.2967/jnumed.120.256180>
- [7] E. M. Lewiecki, Osteoporosis: Clinical Evaluation. MDText.com, Inc., 2018. Available: <https://www.ncbi.nlm.nih.gov/sites/books/NBK279049/>
- [8] “T-Score Vs. Z-Score for Osteoporosis: What the Results Mean,” Healthline, Feb. 21, 2023. <https://www.healthline.com/health/t-score-vs-z-score-osteoporosis#what-are-they>
- [9] J. Peng, E. C. Jury, P. Dönnes, and C. Ciurtin, “Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: Applications and challenges,” *Frontiers*, <https://www.frontiersin.org/articles/10.3389/fphar.2021.720694/full> (accessed Oct. 8, 2023).
- [10] S. Ganiger and R. K M M, “Chronic Diseases Diagnosis using Machine Learning,” ResearchGate, [https://www.researchgate.net/publication/335576533\\_Chronic\\_Diseases\\_Diagnosis\\_using\\_Machine\\_Learning](https://www.researchgate.net/publication/335576533_Chronic_Diseases_Diagnosis_using_Machine_Learning) (accessed Oct. 8, 2023).
- [11] Majeed Wani, Insha ; Arora, Sakshi (2021), “Knee X-ray Osteoporosis Database”, Mendeley Data, V2, doi: 10.17632/fxjm8fb6mw.2
- [12] What is explainable ai? (no date) IBM. Available at: <https://www.ibm.com/topics/explainable-ai> (Accessed: 22 January 2024).
- [13] G. A. Albuquerque et al., “Osteoporosis screening using machine learning and Electromagnetic Waves,” *Nature News*, <https://www.nature.com/articles/s41598-023-40104-w> (accessed Oct. 8, 2023).
- [14] X. Wu et al., “Development of machine learning models for predicting osteoporosis in patients with type 2 diabetes mellitus—a preliminary study,” *Diabetes, Metabolic Syndrome and Obesity*, vol. Volume 16, pp. 1987–2003, 2023. doi:10.2147/dmso.s406695
- [15] X. Wu and S. Park, “A prediction model for osteoporosis risk using a machine-learning approach and its validation in a large cohort,” *Journal of Korean Medical Science*, vol. 38, no. 21, 2023. doi:10.3346/jkms.2023.38.e162
- [16] Inui, A. et al. (2023) ‘Screening for osteoporosis from blood test data in elderly women using a machine learning approach’, *Bioengineering*, 10(3), p. 277. doi:10.3390/bioengineering10030277.
- [17] Y. Ma, Q. Lu, F. Yuan, and H. Chen, “Comparison of the effectiveness of different machine learning algorithms in predicting new fractures after PKP for osteoporotic vertebral compression fractures,” *Journal of Orthopaedic Surgery and Research*, vol. 18, no. 1, 2023. doi:10.1186/s13018-023-03551-9
- [18] Wani, I. M., & Arora, S. (2023). Osteoporosis diagnosis in knee X-rays by transfer learning based on convolution neural network. *Multimedia Tools and Applications*, 82(9), 14193–14217. <https://doi.org/10.1007/s11042-022-13911-y>
- [19] L. Fasihi, B. Tartibian, R. Eslami, and H. Fasihi, “Artificial intelligence used to diagnose osteoporosis from risk factors in clinical data and proposing sports protocols,” *Nature News*, <https://www.nature.com/articles/s41598-022-23184-y> (accessed Oct. 8, 2023).
- [20] R. Dzierżak and Z. Omiotek, “Application of deep convolutional neural networks in the diagnosis of osteoporosis,” *Sensors*, vol. 22, no. 21, p. 8189, 2022. doi:10.3390/s22218189
- [21] Kwon, Y. et al. (2022) ‘Osteoporosis pre-screening using ensemble machine learning in postmenopausal Korean women’, *Healthcare*, 10(6), p. 1107. doi:10.3390/healthcare10061107.
- [22] M. Jang et al., “Opportunistic osteoporosis screening using chest radiographs with Deep Learning: Development and external validation with a cohort dataset,” *Journal of Bone and Mineral Research*, vol. 37, no. 2, pp. 369–377, 2021. doi:10.1002/jbmr.4477
- [23] Jang, R. et al. (2021) ‘Prediction of osteoporosis from simple hip radiography using deep learning algorithm’, *Scientific Reports*, 11(1). doi :10.1038/s41598-021-99549-6.
- [24] Ou Yang, W.-Y. et al. (2021) ‘Development of machine learning models for prediction of osteoporosis from Clinical Health Examination Data’, *International Journal of Environmental Research and Public Health*, 18(14), p. 7635. doi:10.3390/ijerph18147635.
- [25] M. Anam et al., “Osteoporosis prediction for trabecular bone using Machine Learning: A Review,” SSRN, [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3786263](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3786263) (accessed Oct. 8, 2023).

- [26] Yamamoto, N. et al. (2020) ‘Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates’, *Biomolecules*, 10(11), p. 1534. doi:10.3390/biom10111534.
- [27] X. Yu, C. Ye, and L. Xiang, “Application of artificial neural network in the diagnostic system of osteoporosis,” *Neurocomputing*, <https://www.sciencedirect.com/science/article/abs/pii/S0925231216306610> (accessed Oct. 8, 2023).
- [28] Yoo, T.K. et al. (2013) ‘Osteoporosis risk prediction for bone mineral density assessment of postmenopausal women using machine learning’, *Yonsei Medical Journal*, 54(6), p. 1321. doi:10.3349/ymj.2013.54.6.1321.
- [29] L. Ren, T. Wang, Aicha Sekhari Seklouli, H. Zhang, and A. Bouras, “A review on missing values for main challenges and methods,” *Information Systems*, vol. 119, pp. 102268–102268, Oct. 2023, doi: <https://doi.org/10.1016/j.is.2023.102268>.
- [30] S. Bai, J. Z. Kolter, and V. Koltun, “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling,” Mar. 2018, Accessed: Dec. 14, 2022. [Online]. Available: <http://arxiv.org/abs/1803.01271>
- [31] R. A. A. Viadinugroho, “Imbalanced Classification in Python: SMOTE-Tomek Links Method,” *Medium*, Apr. 18, 2021. <https://towardsdatascience.com/imbalanced-classification-in-python-smote-tomek-links-method-6e48dfe69bbc>
- [32] Scikit-Learn, “`sklearn.preprocessing.StandardScaler` — scikit-learn 0.21.2 documentation,” *Scikit-learn.org*, 2019. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
- [33] R. G. Brereton and G. R. Lloyd, “Support Vector Machines for classification and regression,” *Analyst*, <https://pubs.rsc.org/en/content/articlelanding/2010/an/b918972f> (accessed Jan. 21, 2024).
- [34] J. M. Moguerza and A. Muñoz, “Support Vector Machines with applications,” *Project Euclid*, <https://projecteuclid.org/journals/statistical-science/volume-21/issue-3/Support-Vector-Machines-with-Applications/10.1214/088342306000000493.full> (accessed Jan. 21, 2024).
- [35] G. Battineni, N. Chintalapudi, and F. Amenta, “Machine learning in medicine: Performance calculation of dementia ...,” *Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM)*, [https://www.researchgate.net/publication/334054467\\_Machine\\_learning\\_in\\_medicine\\_Performance\\_calculation\\_of\\_dementia\\_prediction\\_by\\_support\\_vector\\_machines\\_SVM](https://www.researchgate.net/publication/334054467_Machine_learning_in_medicine_Performance_calculation_of_dementia_prediction_by_support_vector_machines_SVM) (accessed Jan. 6, 2024).
- [36] Author links open overlay panel Pall Oskar Gislason et al., “Random forests for land cover classification,” *Pattern Recognition Letters*, <https://www.sciencedirect.com/science/article/abs/pii/S0167865505002242?via%3Dihub> (accessed Jan. 21, 2024).
- [37] R. Krishnamoorthi et al., “A novel diabetes healthcare disease prediction framework using machine learning techniques,” *Journal of Healthcare Engineering*, <https://www.hindawi.com/journals/jhe/2022/1684017/> (accessed Jan. 21, 2024).
- [38] “What is the K-nearest neighbors algorithm?,” IBM, <https://www.ibm.com/topics/knn> (accessed Jan. 21, 2024).
- [39] M. A. Vahedifar, A. Akhtarshenas, M. Sabbaghian, and M. Rafatpanah, “Information Modified K-Nearest Neighbor.” Tehran, Dec. 4, 2023.
- [40] W. Li, Y. Yin, X. Quan, and H. Zhang, “Gene expression value prediction based on XGBoost algorithm,” *Frontiers*, <https://www.frontiersin.org/articles/10.3389/fgene.2019.01077/full> (accessed Jan. 21, 2024).
- [41] N. K. Sinha, M. Khulal, M. Gurung, and A. Lal, “Developing a web based system for breast cancer prediction using XGboost classifier,” *International Journal of Engineering Research & Technology*, <https://www.ijert.org/developing-a-web-based-system-for-breast-cancer-prediction-using-xgboost-classifier> (accessed Jan. 21, 2024).
- [42] B. B. & B. Greenwell, “Hands-on machine learning with R,” Chapter 12 Gradient Boosting, <https://bradleyboehmke.github.io/HOML/gbm.html> (accessed Jan. 21, 2024).
- [43] Marcano-Cedeño, A. et al. (2010) Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network, *researchgate*. Available at:

[https://www.researchgate.net/publication/224207758\\_Feature\\_selection\\_using\\_Sequential\\_Forward\\_Selection\\_and\\_classification\\_applying\\_Artificial\\_Metaplasticity\\_Neural\\_Network](https://www.researchgate.net/publication/224207758_Feature_selection_using_Sequential_Forward_Selection_and_classification_applying_Artificial_Metaplasticity_Neural_Network) (Accessed: 23 January 2024).

- [44] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99. <https://doi.org/10.1016/j.inffus.2023.101805>
- [45] Rodríguez-Pérez, R., & Bajorath, J. (2020). Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design*, 34(10), 1013–1026. <https://doi.org/10.1007/s10822-020-00314-0>
- [46] van Zyl, C., Ye, X., & Naidoo, R. (2024). Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP. *Applied Energy*, 353. <https://doi.org/10.1016/j.apenergy.2023.122079>
- [47] Gabbay, F., Bar-Lev, S., Montano, O., & Hadad, N. (2021). A lime-based explainable machine learning model for predicting the severity level of covid-19 diagnosed patients. *Applied Sciences (Switzerland)*, 11(21). <https://doi.org/10.3390/app112110417>
- [48] Viswan, V., Shaffi, N., Mahmud, M., Subramanian, K., & Hajamohideen, F. (2023). Explainable Artificial Intelligence in Alzheimer's Disease Classification: A Systematic Review. In *Cognitive Computation*. Springer. <https://doi.org/10.1007/s12559-023-10192-x>
- [49] A. S. Ahuja, “The impact of artificial intelligence in medicine on the future role of the physician,” NIH, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6779111/> (accessed Jan. 24, 2024).
- [50] M. Javaid et al., “Significance of machine learning in Healthcare: Features, pillars and applications,” *International Journal of Intelligent Networks*, <https://www.sciencedirect.com/science/article/pii/S2666603022000069> (accessed Jan. 24, 2024).
- [51] Saudi Arabia Diagnostic Imaging Market Size & Share Analysis - Industry Research Report - Growth Trends, <https://mordorintelligence.com/industry-reports/saudi-arabia-diagnostic-imaging-equipment-market-industry> (accessed Jan. 23, 2024).

## Appendix D.2: Osteoporosis Conference Paper

# Pre-emptive Diagnosis of Osteoporosis and Osteopenia using Clinical Data

Sunday O. Olatunji  
*College of Computer Science and Information Technology*  
*Imam Abdulrahman Bin Faisal University*  
Dammam, Saudi Arabia  
[osunday@iau.edu.sa](mailto:osunday@iau.edu.sa)

Razan Alghammary  
*College of Computer Science and Information Technology*  
*Imam Abdulrahman Bin Faisal University*  
Dammam, Saudi Arabia  
[2200004035@iau.edu.sa](mailto:2200004035@iau.edu.sa)

Mohammad Aftab Alam Khan  
*College of Computer Science and Information Technology*  
*Imam Abdulrahman Bin Faisal University*  
Dammam, Saudi Arabia  
[mkhan@iau.edu.sa](mailto:mkhan@iau.edu.sa)

Shahad Alghamdi  
*College of Computer Science and Information Technology*  
*Imam Abdulrahman Bin Faisal University*  
Dammam, Saudi Arabia  
[2200003434@iau.edu.sa](mailto:2200003434@iau.edu.sa)

Fai Saleh Alanazi  
*College of Computer Science and Information Technology*  
*Imam Abdulrahman Bin Faisal University*  
Dammam, Saudi Arabia  
[220000931@iau.edu.sa](mailto:220000931@iau.edu.sa)

Fatimah Abbas Alkhatim  
*College of Computer Science and Information Technology*  
*Imam Abdulrahman Bin Faisal University*  
Dammam, Saudi Arabia  
[2200001977@iau.edu.sa](mailto:2200001977@iau.edu.sa)

Rahaf Yaanallah  
*College of Computer Science and Information Technology*  
*Imam Abdulrahman Bin Faisal University*  
Dammam, Saudi Arabia  
[2200003935@iau.edu.sa](mailto:2200003935@iau.edu.sa)

Mehwash Farooqui  
*College of Computer Science and Information Technology*  
*Imam Abdulrahman Bin Faisal University*  
Dammam, Saudi Arabia  
[mfarooqui@iau.edu.sa](mailto:mfarooqui@iau.edu.sa)

Mohammed Imran Basheer Ahmed  
*College of Computer Science and Information Technology*  
*Imam Abdulrahman Bin Faisal University*  
Dammam, Saudi Arabia  
[mbahmed@iau.edu.sa](mailto:mbahmed@iau.edu.sa)

**Abstract**—Osteoporosis and osteopenia are bone diseases characterized by low bone density and structural deterioration to an extent that makes the bones brittle and prone to fractures. The World Health Organization (WHO) differentiates osteopenia as a milder form from osteoporosis as a severe form associated with bone fractures based on statistical values. The fact that this disease is often silent, meaning it is not accompanied by noticeable symptoms, typically reveals itself when a bone fracture occurs. By then, the bones have undergone years of severe deterioration in strength and structure, which is why Bone Mineral Density (BMD) testing is essential for identifying osteoporosis and osteopenia and estimating the likelihood of future fractures. BMD is determined via a bone density test that uses dual-energy X-ray absorptiometry (DXA) to quantify the amount of minerals, such as calcium and phosphorus, inside the bones. Nevertheless, this method may be costly, which limits its accessibility. Therefore, this paper investigates the pre-emptive diagnosis of osteoporosis and osteopenia using machine learning techniques, exclusively utilizing clinical

data without relying on Bone Mass Density (BMD) measurements. The dataset, which consists of 240 patients, originates from the BMD camp organized by the Unani and Panchkarma Hospital in Srinagar, J&K, India. Employing three machine learning algorithms—Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (GBoost)—this study leverages GridSearchCV with 10-folds for model optimization. Central to the methodology is the utilization of Sequential Forward Feature Selection (SFFS) to reduce the number of features and enhance predictive accuracy. Significantly, employing the SMOTE-Tomek approach revealed that Random Forest outpaced other methodologies, using 20 features, where it achieved an accuracy of 91.11% with precision, recall, and F1-score values of 92%, 91%, and 91% respectively.

**Index Terms**—Osteoporosis, Osteopenia, Bone Mineral Density (BMD) testing, Dual-energy X-ray absorptiometry (DXA), Pre-emptive diagnosis, Machine learning techniques, Clinical data.

## I. INTRODUCTION

Chronic diseases, a result of genetic, environmental, and lifestyle factors, are a global challenge, with over 31.4 million deaths occurring in regions with limited economic resources [1]. These diseases affect people of all ages and regions, with 17 million deaths occurring before 70. Rapid urbanization, globalization, and an aging population contribute to their rise. They're closely tied to poverty, hindering efforts to reduce it, as chronic diseases bring hefty healthcare costs. Millions fall into poverty every year as a result of expensive medical care and lost wages [2]. In Saudi Arabia, chronic diseases like osteoporosis and osteopenia are prevalent, with the population expected to rise due to demographic shifts and increased life expectancy [3].

Osteoporosis is a condition that arises when there's a decline in bone mineral density and mass [4]. This weakening of bones can significantly elevate the risk of fractures. On the other hand, osteopenia is less severe than osteoporosis; it can be considered as the precursor of osteoporosis. It simply indicates that there is less bone mineral density than usual, but it is not yet seriously affecting the person. Additionally, osteoporosis and osteopenia typically progress without noticeable symptoms. This disease is a primary cause of fractures in older men and postmenopausal women, and affecting individuals from

all backgrounds [5]. While the risk of developing osteoporosis increases with age, it can manifest at any stage of life.

According to the World Health Organization (WHO), Dual-energy X-ray absorptiometry (DXA) is a globally recognized standard for diagnosing osteoporosis and osteopenia through the measurement of bone mineral density (BMD) testing [6] using the T-score as shown in Table I. The T-score is determined by taking the patient's bone mineral density (BMD) and subtracting the mean BMD (in g/cm<sup>2</sup>) of a reference population of young adults. This difference is then divided by the standard deviation (SD) of the young-adult reference group [7].

TABLE I  
WHO DIAGNOSIS OF OSTEOPOROSIS AND OSTEOPENIA BASED ON BMD MEASUREMENTS FROM DXA

Diagnosis	Bone Mass Density Measurement	T-Score Range
Normal	$\mu : \sigma < 1$	$\geq -1$
Osteopenia	$\mu : 1 \leq \sigma < 2.5$	$-1 < -2.5$
Osteoporosis	$\mu : \sigma \geq 2.5$	$\leq -2.5$

Table I shows the classification of patients with the corresponding T score values. Normal samples have T-scores greater than -1, while those between -1 and -2.5 are diagnosed as osteopenia. T-scores equal to or below -2.5 are osteoporosis. Bone mineral density test results are reported in two forms: T-score and Z-score. T-score represents bone density compared to healthy adults of the same sex, and Z-score represents bone density compared to healthy adults of the same age and sex [8].

In the last decade, ML has significantly impacted healthcare, especially in personalized medicine [9]. This has opened new possibilities in clinical studies and offers promising avenues for precision medicine in research and practical healthcare settings. Motivated by ML's potential to transform early disease detection, we are driven to explore new horizons in our work [10]. Many studies have contributed to the identification and prediction of this medical disease, but most of them depended on specialized imaging technologies that limit the general application. Additionally, earlier studies focused mostly on diagnosing the illness after the patient had experienced symptoms. As a response to these limitations, our main contribution to this paper is to address these issues by employing available clinical data to train and evaluate several types of machine learning algorithms for the preemptive detection of osteoporosis and osteopenia without relying on bone mineral density test results that usually depend on using DXA scans.

The Knee X-ray Osteoporosis Database from Shri Mata Vaishno Devi University is the dataset used in this study, which was obtained from the Mendeley Data website. The dataset includes the T-score values from each participant's knee X-ray and Quantitative Ultrasound System [11], together with clinical factors responsible for

osteoporosis and osteopenia. The utilized ML algorithms are Random Forest (RF), Support Vector Machines (SVM), and Gradient Boosting (GBoost) which have been chosen after an extensive literature survey. Using 20 features only and SMOTETomek sampling, Random Forest outperformed with an accuracy of 91.11% with precision, recall, and F1-score values of 92%, 91%, and 91% respectively.

## II. RELATED WORK

Albuquerque et al. [12] proposed a study to develop an automated system for early detection of osteoporosis using bone mineral density data from 505 patients. Approximately 21.8% are healthy, while the majority, 78.2%, have osteoporosis. They trained a machine learning model on DXA scan measurements, using a 5-fold cross-validation with 20% of the data for testing. By incorporating electromagnetic wave analysis, Random Forest is the best model, with a sensitivity of 0.853, a specificity of 0.879, and an F1 score of 0.859. The findings emphasize age, BMI, and Osseus' signal attenuation as the most critical criteria in the categorization of osteoporosis.

Additionally, Wu and Park [13] conducted a study aiming to predict osteoporosis risk using machine learning (ML) in adults over 40 in the Ansan/Anseong cohort and to analyze its association with fractures in the HEXA cohort. This data was split into 80% training (7,074 samples) and 20% testing (1,768 samples). Using 109 standardized variables, they select LoR, SVM, XGBoost, DT, RF, KNN, and DNN for predicting metabolic status models to improve ROC area, accuracy, and K-fold performance in testing. XGBoost, DNN, and RF produced a highly accurate prediction model with an AUC of 0.86 in the ROC using 10, 15, and 20 features. The top two models for accuracy were XGBoost and RF with 15 features, achieving 0.902 and 0.903, respectively. The study has some limitations, such as its cross-sectional design and less precise BMD measurements compared to DEXA. The large sample size helps in clinical osteoporosis risk assessment.

Moreover, Fasihi et al. [14] conducted a study that aims to explore the use of artificial intelligence to diagnose osteoporosis and suggest exercise protocols to mitigate its effects. They analyzed clinical data from three Tehran hospitals, comprising 1224 individuals. They used various machine learning models, such as RF, KNN, SVM, DT, AB, GB, ET, and MLP analysis, to predict osteoporosis and suggest suitable exercise programs for treatment. The dataset was separated into training (80%) and test (20%) sets. FR yielded the most accurate predictions for men (AUROC 0.91), while GB performed best for women (AUROC 0.95). RF excelled in recommending exercises for healthy individuals and those with osteopenia and osteoporosis, achieving AUROCs of 0.96 and 0.99 for women and men, respectively.

Another study by Kwon et al. [15] aimed to develop a pre-screening method for early osteoporosis diagnosis in post-menopausal Korean women using machine learning. They utilized data from the Korea National Health and Nutrition Examination Surveys, comprising 1431 women aged 40–69. Feature selection identified 20 relevant features. Three machine learning algorithms (AdaBoost, Gradient Boosting,

and Random Forest) were trained on three models. Model C, incorporating both checkup and survey features, yielded the highest accuracy rates: 84.9%, 82.9%, and 83.2% for AdaBoost, GBM, and RF, respectively, with AUROC scores of 92.1%, 90.8%, and 91.9%. Limitations include data from cross-sectional observational studies and focusing on a specific age group, limiting generalizability.

A review of the literature on early osteoporosis and osteopenia prediction indicated that most previous studies on the subject have concentrated on imaging datasets while ignoring the potential of clinical data. Moreover, earlier studies frequently encountered constraints, such as limited sample sizes and a lack of adequate clinical data. Additionally, it is notable that the majority of these studies have focused on the detection of osteoporosis, frequently ignoring osteopenia. Our paper addresses this gap by focusing on the prediction of both osteopenia and osteoporosis using clinical data to construct a machine-learning model for reliable and more accurate osteoporosis and osteopenia prediction.

## III. MATERIALS AND METHODS

This study developed a pre-emptive model for diagnosing osteoporosis and osteopenia using Python programming language. All operations were conducted with a fixed seed value of 0 to ensure repeatability of results and consistency across runs. The dataset went through a number of crucial pre-processing steps before modeling to ensure data integrity and feature quality. These steps included handling outliers, checking for duplication, and checking for missing values across all features. After that, LabelEncoder was used to convert the values of categorical columns into numerical representations for better model interpretability. Moreover, to address the class imbalance, the SMOTETomek technique was used which is a hybrid sampling method that combines the advantages of SMOTE (Synthetic Minority Over-sampling Technique) and Tomek links. With the use of Tomek links and SMOTE, it attempts to rectify class imbalance by undersampling the majority class and oversampling the minority class. The dataset was then divided into training and testing sets, using 80% for training and 20% for testing, and standardized using StandardScaler for optimal model performance. This method assumes that data is normally distributed and will scale them such that the distribution is centered around zero with a standard deviation of one. Feature selection was conducted using Sequential Forward Feature Selection (SFFS), a method that sequentially chooses the most suitable features for a machine learning model. Three different classifiers—Random Forest, SVM, and Gradient Boosting—were utilized. For each model, GridSearchCV was employed to tune hyperparameters via cross-validation using 10 folds. The best hyperparameters for each model were identified, and their performance was evaluated on the

test set using accuracy, precision, recall, F1-score, and AUC. This entire process is illustrated in Figure 1.

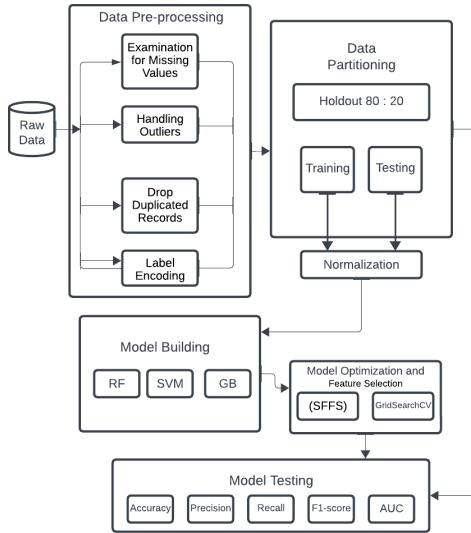


Fig. 1. The proposed framework for the pre-emptive diagnosis of Osteoporosis and osteopenia.

#### A. Data Description

TABLE II  
FEATURES' DESCRIPTION.

Feature	Type
Joint Pain	Categorical
Gender	Categorical
Age	Integer
Menopause Age	Integer
Height (Meter)	Integer
Weight (KG)	Integer
Smoker	Categorical
Alcoholic	Categorical
Diabetes	Categorical
Hypothyroidism	Categorical
Number of Pregnancies	Integer
Estrogen Use	Categorical
Occupation	Categorical
History of Fracture	Categorical
Dialysis	Categorical
Family History of Osteoporosis	Categorical
Maximum Walking Distance (km)	Float
Daily Eating Habits	Categorical
Medical History	Categorical
T-score Value	Float
Z-Score Value	Float
BMI	Float
Site of examination	Categorical
Obesity status	Categorical
Diagnosis	Categorical

The dataset utilized in this research originates from a health camp organized by the Unani and Panchkarma Hospital in Srinagar, Jammu & Kashmir, India, held from December 21 to December 31, 2019 [11]. The dataset named Knee X-ray Osteoporosis Database compromises both images and clinical records, but for our study only the clinical data is used. The dataset consists of 240 entries, and encompassing 26

laboratory biomarkers and demographic features. Noteworthy features are illustrated in Table II along with their respective types, considering each participant's T-score and Z-score data from their knee X-ray and quantitative ultrasound system. For the purpose of model training and evaluation, the dataset was initially split into 80% for training, comprising 167 samples, and 20% for testing, comprising 72 samples. Subsequently, to address potential class imbalance and improve the model's performance, an oversampling technique was applied to the entire dataset. After oversampling, the training set consisted of 357 samples, while the testing set remained at 90 samples. This oversampling technique enhances the model's ability to learn from minority classes, providing a more balanced and reliable evaluation of the study findings.

#### B. Statistical Analysis

This section conducts a thorough statistical analysis to identify data patterns and identify appropriate preprocessing techniques for optimal model preparation using a dataset with numerical and categorical features. The numerical attributes underwent thorough scrutiny through various statistical metrics. These metrics provided profound insights into the inherent characteristics of the data. A detailed statistical breakdown showcasing the properties and attributes of the numerical features is presented in Table III, offering a profound understanding of the dataset's numeric aspects.

TABLE III  
DESCRIPTIVE STATISTICS OF PARTICIPANT CHARACTERISTICS

	Age	Weight	Height	MWD	BMI
Mean	51.13	69.05	1.58	1.94	27.58
Standard deviation	13.23	9.88	0.09	1.98	4.05
Min	17.00	39.00	1.37	0.10	16.13
25th quartile	44.50	63.00	1.52	0.50	24.94
50th quartile	50.00	69.00	1.57	1.00	27.26
75th quartile	60.00	74.50	1.65	3.00	30.21
Max	107.00	98.00	1.82	10.00	42.75
Missing values	0	0	0	1	0

The categorical features as shown in Figure 2 across the dataset shed light on various patient attributes. The distribution of the target feature indicates that among 240 patients, 154 have osteopenia, 49 have osteoporosis, and 37 are considered normal. The gender distribution showcases 132 female and 108 male patients. Alcohol consumption shows that all the patients do not consume alcohol. Smoking habits vary, with 199 non-smokers and 41 smokers. 12 patients are diabetic, and 228 do not suffer from diabetes. Estrogen intake shows 229 patients not taking supplements and 11 who do. Joint pain involves 2 patients who do not have any joint pain and 238 patients who suffer from joint pain. 34 patients have

hypothyroidism, and 206 do not. The family history of osteoporosis comprises 174 patients with no family history and 66 with a positive history. 8 patients have seizure disorder, and 232 do not. One patient is on dialysis, and 339 are not. Menopause Age, Number of Pregnancies, Occupation, History of Fracture, Maximum Walking Distance (MWD), Daily Eating Habits, Medical History, Site, and Obesity complete the categorical feature analysis, providing a comprehensive view of patient demographics and health-related behaviors.

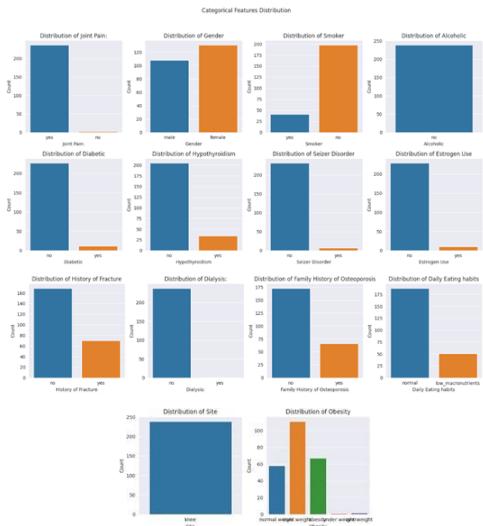


Fig. 2. description of the categorical features

### C. Data Pre-processing

Before the dataset gets transmitted to the algorithm, it is preprocessed to look for missing values, noisy data, and other inconsistencies. Various Python libraries were used for data analysis and pre-processing in the current study. Initially, the data set contained 27 features and 240 records; T-Score values and Z-Score values weren't included in the training set since the DXA findings were used as the target. Columns that contain a significant number of null values were dropped. Moreover, cleaning the data involves removing duplication using the Pandas duplicated() method. Following preprocessing, the number of features was reduced to 20, and instances were limited to 239. Furthermore, identify and transform categorical values into numerical values using the LabelEncoder() method. Missing values are crucial for data quality and model performance. Common methods include removing incomplete entries or imposing reasonable values, with deletion and imputation being the most widely used approaches [16]. In this study, by examining the distribution of numerical null variables, if a variable has a normal distribution, utilize the mean value. If not, replace missing values in skewed data sets with the median value.

The SMOTETomek method is used to balance class distribution by performing an oversampling technique using SMOTE and an undersampling technique using Tomek Links. SMOTE identifies minority class data, creates synthetic data, and repeats until balanced. Tomek Links are points from different classes that are close to one another, and for each pair, the

points from the majority class will be removed to eliminate any points that are unclearly near the class decision boundary. This technique will improve model performance and the decision-making process [17]. Following that, StandardScaler was used to scale down the training and testing features with a zero mean and unit variance.

### C. D. Description of Utilized Machine Learning Algorithms

1) *Random Forest (RF)*: Introduced by Leo Breiman in 2001 [18], RF revolutionized ensemble learning with bagging or "bootstrap aggregation." It uses multiple Decision Trees to avoid overfitting, aggregating predictions through a majority vote mechanism. By leveraging insights from various trees, Random Forest enhances predictive accuracy, making it a powerful tool for classification tasks [19].

2) *Gradient Boosting (GBoost)*: GBoost, introduced by Jerome Friedman in 1999, is a potent ensemble learning technique. The Gradient Boosting Machine (GBM), refined by Friedman, Hastie, and Tibshirani, constructs a predictive model through the sequential addition of weak learners, usually decision trees. Excelling in regression and classification tasks, GBM captures intricate data relationships. Its adaptability to diverse datasets and iterative accuracy improvement make it widely used in machine-learning applications. [20].

3) *Support Vector Machine (SVM)*: Proposed by Cortes and Vapnik in 1990, SVM has gained popularity in machine learning [21]. It functions as a supervised learning algorithm, addressing classification and regression, with a focus on binary classification. SVM establishes a hyperplane in the feature space for maximum margin between classes [22]. During testing, data points are mapped, allowing SVM to create robust decision boundaries based on their position relative to the margin.

### D. E. Performance Measure

Performance evaluation for the model involves assessing the model's ability to correctly identify the patient's state from the dataset. Thus, a range of performance metrics were used in this study to evaluate the models' performance. Including Accuracy, Precision, Recall, and F1-score. Confusion matrices, which contain True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), were utilized for further model evaluation.

### F. Optimization Strategy

To get the best possible performance for the models, hyperparameters were utilized. Proper hyperparameter tuning is often a critical step in achieving optimal model performance. To get the suitable hyperparameter for each algorithm, the GridSearchCV method was employed. The method was given a set of different values for each

hyperparameter, and it then generated the best combination of these values by using the training set. Table IV illustrates the algorithms and their best hyperparameters with the original data. Table V represents the algorithms with their best hyperparameters in oversampled data and using all the features.

TABLE IV  
THE OPTIMAL HYPERPARAMETER FOR EACH CLASSIFIER IN THE ORIGINAL DATA

Algorithm	Hyperparameter	Best Hyperparameter
RF	Max depth	none
	Min samples leaf	1
	Min samples split	5
	N estimator	200
Gboost	learning rate	0.1
	Max depth	7
	N estimator	100
SVM	kernal	linear
	C	1
	Gamma	scale

TABLE V

THE OPTIMAL HYPERPARAMETER FOR EACH CLASSIFIER IN OVER-SAMPLED DATA WITH ALL THE FEATURES SELECTED

Algorithm	Hyperparameter	Best Hyperparameter
RF	Max depth	none
	Min samples leaf	1
	Min samples split	5
	N estimator	150
Gboost	learning rate	0.1
	Max depth	5
	N estimator	150
SVM	kernal	Rbf
	C	10
	Gamma	scale

## I. RESULTS AND DISCUSSION

Within this section, we present the outcomes derived from the developed models after the implementation of Grid- SearchCV on both the original dataset and the sampled data using SMOTETomek, a method used to balance the class distribution within the dataset by oversampling the minority class and undersampling the majority class. Following the acquisition of optimal hyperparameters and model training through stratified 10-fold cross-validation. Utilizing all 20 features. The ensuing section delineates a comprehensive examination of the results obtained, as presented in Table VI. The original results highlighted the need to address the class imbalance, as demonstrated by the enhanced performance of all models following SMOTETomek's implementation. However, substantial improvements were observed in all algorithms after using SMOTETomek.. Testing accuracies soared, demonstrating a notable improvement in predictive power. Moreover, the use of hyperparameter tuning and feature selection contributed to more robust models. Random Forest emerged as the algorithm with the highest testing accuracy, achieving 91.11%, precision, recall, and F1 values of 92%, 91%, and 91%, respectively. Gradient Boosting and SVM both followed with a testing accuracy of 87.77%. Gradient Boosting with precision, recall, and F1 values are all 88%. SVM with precision, recall, and F1 values are 88%, 88%, and 87%, respectively. These models

have shown significant improvements in predictive accuracy, demonstrating the usefulness of SMOTETomek in managing class imbalance and boosting the models' capacity to predict osteoporosis and osteopenia, which may facilitate early detection and intervention.

TABLE VI  
RESULTS OF TESTING ACCURACY, PRECISION, RECALL, AND F1 ON ORIGINAL DATA AND USING SMOTETOMEK

Original Dataset	RF	GBoost	SVM
Testing Accuracy	70.83%	73.61%	73.61%
Precision	71.00%	67.00%	76.00%
Recall	71.00%	69.00%	74.00%
F1-Score	70.00%	68.00%	74.00%
Using SMOTETomek			
Testing Accuracy	91.11%	87.77%	87.77%
Precision	92.00%	88.00%	88.00%
Recall	91.00%	88.00%	88.00%
F1-Score	91.00%	88.00%	87.00%

### A. Further Discussion of the Results

RandomForest's confusion matrix showcases a robust performance across all classes, as shown in the tables below, with 30 instances correctly predicted as osteopenia, 24 instances as normal, and 28 instances as osteoporosis, demonstrating fewer misclassifications compared to other models. Following that, SVM demonstrates a strong performance with 30 instances correctly predicted as osteopenia, 22 instances as normal, and 27 instances as osteoporosis, displaying fewer misclassifications overall, particularly between class normal and class osteoporosis. GradientBoosting exhibits a reasonably balanced performance, with 29 instances correctly predicted as osteopenia, 24 instances as normal, and 26 instances as osteoporosis.

TABLE VII  
CONFUSION MATRIX OF RANDOMFOREST

Actual	Predicted		
	Osteopenia	Normal	Osteoporosis
Osteopenia	30 (TP)	0 (FN)	0 (FN)
Normal	3 (FP)	24 (TN)	4 (FN)
Osteoporosis	0 (FP)	1 (FP)	28 (TN)

TABLE VIII  
CONFUSION MATRIX OF GRADIENTBOOSTING

Actual	Predicted		
	Osteopenia	Normal	Osteoporosis
Osteopenia	29 (TP)	1 (FN)	0 (FN)
Normal	4 (FP)	24 (TN)	3 (FN)
Osteoporosis	0 (FP)	3 (FP)	26 (TN)

TABLE IX  
CONFUSION MATRIX OF SVM

Actual	Predicted		
	Osteopenia	Normal	Osteoporosis
Osteopenia	30 (TP)	0 (FN)	0 (FN)
Normal	4 (FP)	22 (TN)	5 (FN)
Osteoporosis	0 (FP)	2 (FP)	27 (TN)

### I. CONCLUSION

Osteoporosis and osteopenia are chronic diseases defined by a loss of bone mass and mineral density, which makes bones weaker and more vulnerable to fractures. Even though osteoporosis is frequently asymptomatic in its early stages, people may only become aware of it when they break a bone because of its "silent" nature. Even though several studies were carried out to detect osteoporosis early on, the majority of them relied on imaging technology to gather information that not all hospitals would have access to. This research attempts to solve these constraints by training and evaluating a range of machine learning algorithms for the proactive detection of osteoporosis using freely available clinical data. Gradient Boosting achieved the highest result of 91.11% accuracy, 91% recall, 92% precision, and 91% F1 score with 20 features using the SMOTETomek sampling technique. In future work, the model will be tested on a new dataset or with new patients, additionally exploring new classifiers and methods like deep learning to increase accuracy. We recommend using the model with patients by using Longitudinal Data Analysis where we collect the patient's data continuously.

### REFERENCES

- [1] World Health Organization. Non communicable diseases. Available online: [link] (accessed Oct. 8, 2023).
- [2] Heine, M.; Hanekom, S. Chronic disease in low-resource settings: Prevention and management throughout the continuum of care-a call for papers. *International Journal of Environmental Research and Public Health* Available online: [link] (accessed Oct. 8, 2023).
- [3] Ministry of Health. National Plan for Osteoporosis Prevention and Management in the Kingdom of Saudi Arabia. Available online: [link] (accessed Oct. 8, 2023).
- [4] National Institute on Aging. Osteoporosis. Available online: [link] (accessed Oct. 8, 2023).
- [5] National Institute of Arthritis and Musculoskeletal and Skin Diseases. Osteoporosis. Available online: [link] (accessed Oct. 8, 2023).
- [6] Haseltine, K.N.; Chukir, T.; Smith, P.J.; Jacob, J.T.; Bilezikian, J.P.; Farooki, A. Bone mineral density: clinical relevance and quantitative assessment. *J. Nucl. Med.* **2021**, *62*, 446–454. [link].
- [7] E. M. Lewiecki, Osteoporosis: Clinical Evaluation. MDText.com, Inc., **2018**.
- [8] “T-Score Vs. Z-Score for Osteoporosis: What the Results Mean,” Healthline, Feb. 21, **2023**. [link].
- [9] J. Peng, E. C. Jury, P. Dönnes, and C. Ciurtin, “Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: Applications and challenges,” *Frontiers*, [link].
- [10] S. Ganiger and R. K M M, “Chronic Diseases Diagnosis using Machine Learning,” ResearchGate, [link].
- [11] Majeed Wani, Insha ; Arora, Sakshi **2021**, “Knee X-ray Osteoporosis Database”, MendeleyData, V2, doi: 10.17632/fxjm8fb6mw.2.
- [12] G. A. Abuquerque et al., “Osteoporosis screening using machine learning and Electro-magnetic Waves,” *Nature News*, <https://www.nature.com/articles/s41598-023-40104-w>.
- [13] X. Wu and S. Park, “A prediction model for osteoporosis risk using a machine-learning approach and its validation in a large cohort,” *Journal of Korean Medical Science*, vol. 38,no. 21, **2023**. doi:10.3346/jkms.2023.38.e162 .
- [14] L. Fasih, B. Tartibian, R. Eslami, and H. Fasih, “Artificial intelligence used to diagnose osteoporosis from risk factors in clinical data and proposing sports protocols,” *NatureNews*, <https://www.nature.com/articles/s41598-022-23184-y> .
- [15] Kwon, Y. et al. (2022) ‘Osteoporosis pre-screening using ensemble machine learning in postmenopausal Korean women’, *Healthcare*, **10**(6), p. 1107. doi:10.3390/healthcare10061107.
- [16] L. Ren, T. Wang, Aicha Sekhari Seklouli, H. Zhang, and A. Bouras, “A review on missing values for main challenges and methods,” *Information Systems*, vol. 119, pp.102268–102268, Oct. 2023, [link].
- [17] R. A. A. Viadinugroho, “Imbalanced Classification in Python: SMOTE-Tomek Links Method,” Medium, Apr. 18, **2021**. [link]
- [18] “Random forests for land cover classification,” *Pattern Recognition Letters*, [link].
- [19] R. Krishnamoorthi et al., “A novel diabetes healthcare disease prediction framework using machine learning techniques,” *Journal of Healthcare Engineering*, [link].
- [20] B. & B. Greenwell,“Hands-on machine learning with R,” Chapter 12 Gradient Boosting, [link].
- [21] R. G. Brereton and G. R. Lloyd, “Support Vector Machines for classification and regression,” *Analyst*, [link].
- [22] G. Battineni, N. Chintalapudi, and F. Amenta, “Machine learning in medicine: Performance calculation of dementia ,” *Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM)*, [link](accessed Jan. 6, 2024).

## Appendix E.1: Epileptic Seizure Journal Paper

# Comprehensible Machine Learning-Based Models for the Pre-emptive Diagnosis of epileptic seizure using Clinical Data

Sunday O. Olatunji<sup>1\*</sup>, Mohammad Aftab Alam Khan<sup>1\*</sup>, Fai Alanazi<sup>1\*</sup>, Rahaf yaan allah<sup>1\*</sup>, Shahad alghamdi<sup>1\*</sup>, Razan Alshammari<sup>1\*</sup>, Fatimah Alkhatim<sup>1\*</sup>, Mehwash Farooqui<sup>1\*</sup> and Mohammed Imran Basheer Ahmed<sup>1\*</sup>

<sup>1</sup> College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia.

\* Correspondence: <sup>1</sup> [osunday@iau.edu.sa](mailto:osunday@iau.edu.sa) <sup>2</sup> [mkhan@iau.edu.sa](mailto:mkhan@iau.edu.sa) <sup>3</sup> [220000931@iau.edu.sa](mailto:220000931@iau.edu.sa) <sup>4</sup> [2200003935@iau.edu.sa](mailto:2200003935@iau.edu.sa)  
<sup>5</sup> [2200003434@iau.edu.sa](mailto:2200003434@iau.edu.sa) <sup>6</sup> [2200004035@iau.edu.sa](mailto:2200004035@iau.edu.sa) <sup>7</sup> [2200001977@iau.edu.sa](mailto:2200001977@iau.edu.sa) <sup>8</sup> [mfarooqui@iau.edu.sa](mailto:mfarooqui@iau.edu.sa) <sup>9</sup> [mbahmed@iau.edu.sa](mailto:mbahmed@iau.edu.sa)

## Abstract

Chronic diseases are a major burden on the world's healthcare systems since they require ongoing care and have a negative effect on the quality of life of those who are afflicted. Among these, epilepsy is most notable as a common neurological condition with repeated seizures that affects more than 50 million individuals globally. Epilepsy diagnosis can be challenging, especially in healthcare settings with limited resources. Nevertheless, early detection of epileptic seizures is essential for optimal management. The paper investigates the use of machine learning algorithms for the preemptive diagnosis of epileptic seizures by utilizing diagnostic, clinical, and demographic data. To distinguish between seizures that are epileptic and those that are not, eight machine learning methods are used: K-Nearest Neighbors, Gradient Boosting, eXtreme Gradient Boosting, Random Forest, Support Vector Machine, Logistic Regression, Decision Tree, and AdaBoost. The dataset from patients with epileptic seizures (ES) or patients with non-epileptic seizures (NES) was gathered between January 1, 2017, and May 15, 2019, and it includes a variety of clinical and demographic information. This study optimizes the model using 10-folds GridSearchCV. The use of SelectKBest to decrease the number of features and improve predicted accuracy is a key component of the process. The results showed that Logistic Regression outpaced other methodologies, using 36 features, where it achieved an accuracy of 87.67% with precision, recall, and F1 values all of 88%. To gain a deeper understanding of the top-performing AI model, the study employed Explainable Artificial Intelligence (XAI) approaches such as Feature importance and Local Interpretable Model-Agnostic Explanations (LIME).

Keywords: Pre-emptive Diagnosis; Chronic diseases; Epilepsy; epileptic seizures; Machine Learning Algorithms; Early Detection.

## 1. Introduction

Chronic diseases are considered one of the greatest health challenges for individuals and communities. They require constant medical care, and those affected often suffer from persistent or recurring symptoms. These conditions can affect the patient's quality of life and cause a decrease in productivity, making certain tasks or activities challenging to perform. Genetic, environmental, and behavioral factors combine to cause chronic diseases, which can have a long duration. With over 75% of the 31.4 million deaths linked to chronic diseases occurring in areas experiencing financial difficulties, chronic diseases represent a major global health concern [1]. Hence, early detection of chronic diseases such as epilepsy is crucial. These illnesses persist throughout an individual's life since those suffering from it are rarely cured and rarely get better on their own. Furthermore, people with chronic illnesses typically face significant problems, which can cause major

disruptions, such as impairments in body processes. It should be noted that some chronic illnesses can be so serious that they may immediately endanger the Patient's life.

Epilepsy, known as a seizure disorder, is a neurological condition characterized by repeated seizures. It is a common chronic disease that affect the nervous system, which arises due to a change in the electrical activity of the brain. Epilepsy is one of the most widespread neurological disorders, affecting approximately 50 million people globally, according to the World Health Organization. Epilepsy affects people of all races, ethnic backgrounds, and ages [2]. Moreover, symptoms of epileptic seizures can differ among individuals. Some individuals may lose consciousness during an epileptic seizure while others remain fully conscious. Certain individuals may gaze into space for several seconds while having an epileptic seizure. Others may have convulsions, which are characterized by repeated twitches of arms or legs.

Seizures are the main symptoms of epilepsy, but these seizures and their accompanying symptoms vary depending on their type. Partial seizures, this type of epilepsy manifests itself in a specific area of the brain and does not involve the entire brain, have two main types: simple partial seizures, which do not involve loss of consciousness, but may present symptoms like dizziness, tingling sensations in limbs, and changes in taste, smell, vision, hearing, or touch; and complex partial seizures, which result in loss of consciousness, inability to respond, and repetitive motions. Then, generalized seizures, which affect the entire brain and come in six categories: absence seizures, which cause brief lapses in consciousness and repetitive movements like eye blinking; atonic seizures, which result in a loss of muscle control and an increased risk of abrupt falls; tonic seizures, which cause muscular stiffness; clonic seizures, characterized by repetitive, jerking movements in facial, neck, and arm muscles; myoclonic seizures, featuring rapid, involuntary movements in arms and legs; and tonic-clonic seizures, which include loss of consciousness, tongue biting, body stiffness, loss of bladder and bowel control [3].

Diagnosing epilepsy is the most important stage of treatment for this disease. Ignoring seizures and not treating them may lead to a significant deterioration of the patient's condition, and more importantly, neglecting the symptoms of epilepsy may cause catastrophic results, up to the point of sudden death [2]. The electroencephalogram (EEG) is a well-known diagnostic tool for epilepsy. It offers information about the electrical activity of the brain during the analysis. The patient's brain activity changes during an epileptic episode. This procedure involves placing a cap on the patient's head with electrodes attached to the scalp. These electrodes monitor brain activity, which is then magnified and displayed as a graph on a computer screen. Abnormal brain activity patterns indicate the presence of epilepsy [4]. But getting a diagnosis early can be challenging, particularly in health care facilities with low resources where access to advanced diagnostic technologies like electroencephalography (EEG) may be restricted and expensive. Therefor, to address the limitations associated with EEG, we've developed an approach for achieving preemptive diagnose of epilepsy using machine learning techniques with clinical data. Our focus is on distinguishing between patients who suffer from epileptic seizures, that indicates the presence of epilepsy, and those with non-epileptic seizures, which appear as a result of different emotional or psychological reasons.

Moreover, machine learning is a subset of artificial intelligence which enables systems to identify patterns and learn from data without significant human engagement, revolutionizing healthcare in the long term. It is anticipated that technology will continue to advance and have a significant impact on the diagnosis and treatment of diseases [5]. Numerous areas of medicine, including infection prevention, disease diagnosis and prognosis, and improving individualized treatment, can benefit from the application of machine learning. Moreover, machine learning algorithms aid in understanding of the huge amounts of healthcare data entered into digital health records on a regular basis. This helps uncover patterns and insights in medical data that could be challenging or take more time to identify manually. With the use of machine learning algorithms, healthcare professionals can make accurate diagnostic decisions without the need for expensive tests or specialized equipment by successfully analyzing clinical characteristics such patient demographics, medical history, and presenting symptoms.

Non-epileptic seizures are sometimes used to describe seizures that are not caused by abnormal electrical activity in the brain. They may be caused by something physiological, such high blood pressure, or they may have something to do with the functioning of the heart. Alternatively, there can be a psychological reason behind them, like depression or anxiety [6]. To provide patients with the best possible medical care and outcomes, and to early diagnose epilepsy, it is critical to be able to differentiate between these two types of seizures. It is not always easy to distinguish between seizures that are epileptic and those that are non-epileptic, in many cases, it requires a thorough assessment and the integration of diverse data sources. However, machine learning algorithms have the potential to discriminate between epileptic and non-epileptic seizures with great accuracy and efficiency because of their ability to evaluate vast volumes of information and spot intricate patterns. By using diverse demographic, clinical, and laboratory characteristics of patients, these algorithms show potential for diagnosing epileptic seizures without relying on expensive techniques such as electroencephalography (EEG).

Therefore, this paper investigates the preemptive diagnosis of epileptic seizures using machine learning techniques, utilizing patients' demographic, clinical and diagnostic data. Using eight machine learning methods are used: K-Nearest Neighbors, Gradient Boosting, eXtreme Gradient Boosting, Random Forest, Support Vector Machine, Logistic Regression, Decision Tree, and AdaBoost. Furthermore, the dataset utilized in this study comprises individuals referred to the Functional Neurological Disorders Clinic and the University of Colorado Epilepsy Monitoring Unit between January 1, 2017, and May 15, 2019 [7]. These patients were diagnosed either with non-epileptic seizures (NES) or epileptic seizures (ES) based on video-electroencephalography (vEEG) assessments. The dataset includes various demographic, clinical, and diagnostic variables such as gender, seizure characteristics, medical history. The results showed that Logistic Regression outpaced other methodologies, using 36 features, where it achieved an accuracy of 87.67% with precision, recall, and F1 values all of 88%.

The following portions of this paper are organized as follows: A full overview of the literature is included in Section 2. The materials and methods used are described in depth in Section 3, including the dataset description, statistical analysis, an explanation of the ML algorithms utilized, and the performance metrics used to evaluate the produced models, and optimization strategy selected. Section 4 provides an explanation of the results and the feature selection technique used, and Section 5 provides a description of the models' outcomes using XAI techniques. Discussion is included in section 6. Lastly, Section 7 concludes with a recommendation for further work and a conclusion.

## 2.Literature Review

Chen et al. [8] developed an automated method for detecting and classifying epileptic seizures using electroencephalogram (EEG) signals. They employed a feature fusion and selection approach to extract mixed features such as Fuzzy Entropy (FuzzyEn), Sample Entropy (SampEn), Approximate Entropy (ApEn), and Standard Deviation (STD) from subbands obtained through Discrete Wavelet Transform (DWT) decomposition. Feature selection was performed using the RF algorithm, and the classification of epilepsy EEG signals was accomplished using CNN. The suggested approach was tested on benchmark datasets involving the EEG datasets from Bonn and the New Delhi datasets. The results showed high accuracy and performance, with the model achieving 99.9% accuracy for the Bonn datasets and 100% classification accuracy for the New Delhi datasets.

A study conducted by Ode et al. [9] addresses the critical need for early detection of focal epileptic seizures by leveraging alterations in heart rate variability (HRV) derived from electrocardiogram (ECG) data, which can be indicative of autonomic nervous system (ANS) disturbances occurring between 15- and 20-minutes preceding seizure onset. The major goal is to create a machine learning algorithm capable of real-time monitoring of R-R interval (RRI) data for seizure prediction using SA-AE algorithm. The application of the

developed technique to clinical data indicates its effectiveness across a majority of patients but has a false positive (FP) rate of 0.85. Moreover, the dataset collection involved 39 individuals with focal epilepsy who were hospitalized to Tokyo Medical and Dental University's (TMDU) Medical Hospital. Additionally, the SA-AE model had a sensitivity of 74%, a precision of 0.35, a false positive rate of 0.85 times/h, and an (AUC) of 0.97. Notably, 29 patients achieved a 100% sensitivity, with eight of them experiencing no false positives.

A paper introduced by Nazari et al. [10] proposes a novel approach for epileptic seizure prediction, focusing on patients with late-onset seizures, where recording preictal signals proves challenging. The suggested method employs a convolutional neural network (CNN) and leverages few-shot learning, requiring minimal data for training. The method is evaluated on EEG data from three cases from the CHB-MIT dataset. The evaluation focuses on a 10-minute SPH, which is a seizure prediction horizon, and a 20-minute SOP, which is a seizure occurrence period, yielding an average sensitivity of 95.70% and a false prediction rate (FPR) of 0.057/h. While these results are promising, it is important to acknowledge limitations, including the small sample size and the need for validation across a more diverse patient population. Further research should focus on expanding the dataset and validating the approach across diverse patient populations to establish its robustness and applicability in clinical settings.

A study by Tawfik et al. [11], This paper aims to explore the use of deep learning and ML for epileptic seizure detection, comparing their effectiveness and trying to enhance existing detection methods. The EEG dataset from Bonn University is available on the UCI website which includes 400 individuals where 200 have epileptic and 200 do not. Epilepsy in individuals can be identified by analyzing EEG signals using various ML methods like SVM, LoR, KNN, DT, RF, XGboost, Catboost, and others. The EEG dataset performed very well with the CNN algorithm, achieving 99.2% accuracy, 99.3% specificity, and 98.7% sensitivity. The hybrid DNN (CNN with LSTM) achieved 98.7% accuracy.

The goal of the study proposed by Hilal et al. [12] was to create an intelligent model employing electroencephalogram (EEG) signals to identify and categorize epileptic seizures. The authors present a deep canonical sparse autoencoder-based epileptic seizure detection and classification (DCSAE-ESDC) model that includes two primary processes: feature selection and classification. utilizing coyote optimization algorithm (COA) for feature selection and krill herd algorithm (KHA) for tuning the model's parameters. A benchmark dataset for the detection of epileptic seizures from the UCI repository was used for the investigation. The analytical findings demonstrate that the DCSAE-ESDC approach surpasses existing techniques in binary and multi- classification, with an accuracy rating of 98.67% and 98.73%, respectively.

A study written by Jemal et al. [13], the author's aim of the study is to examine a comprehensible Deep-learning (DL) classifier to predict seizures by the EEG data. The publicly available dataset includes a recording of continuous multi-channel scalp EEG for 940 hours from 23 patients between the ages of 1.5 to 19 years at Boston Children's Hospital. The architecture of the DNN employs the algorithm Filter Bank Common Spatial Pattern (FBCSP) for EEG data analysis. It includes temporal and depth-wise convolutions, batch normalization, and feature extraction for long continuous EEG data analysis. They reached an accuracy of 90.9% and compared to other studies they achieved the highest sensitivity which is 96.1%.

Ouichka et. al. [14] have conducted a study regarding the automatic epileptic seizure's prediction. The study focuses on improving the accuracy of epileptic seizure prediction and early detection using intracranial electroencephalogram (iEEG) datasets. Five deep learning models were presented for automated seizure prediction, together with Convolutional Neural Networks (CNNs) and transfer learning with ResNet50. The 3-CNN and 4-CNN models yielded the highest accuracy, according to the experimental data, achieving an

impressive 95%. The dataset utilized in this study originates from the American Epilepsy Society for Seizure Prediction. This dataset encompasses intracranial EEG signals (iEEG) obtained from a diverse sample pool, consisting of five (5) dogs and two human patients.

A study authored by Usman et. al. [15], the main focus was to develop an epileptic seizure prediction method utilizing EEG monitoring, with a specific focus on accurately predicting the preictal state preceding seizure onset. The methodology comprises three key steps: firstly, employing empirical model decomposition for noise reduction and utilizing Generative Adversarial Networks to address class imbalance by generating preictal samples during EEG signal preprocessing. Secondly, automated feature extraction is performed using a three-layer CNN. Lastly, Long Short-Term Memory units are employed in order to distinguish between preictal and interictal states. The CHBMIT dataset, which contains scalp EEG signals from 22 subjects, is utilized. The proposed method achieves high sensitivity (93%) and specificity (92.5%) with an average anticipation time of 32 minutes for predicting seizure onset.

A study written by Almustafa [16] aimed to find the best classification algorithm for an epileptic seizure dataset to determine whether a seizure was present or not by using various classification techniques, as well as to assess the behavior of the algorithm when parameters are changed. The dataset used in this study consisted of 11,500 samples, each with 178 features, and was categorized into five classes based on different conditions during the recording of EEG signals such as open eyes, closed eyes, and various stages of epileptic seizures. The ML algorithms applied were KNN, Naïve Bayes, RF, DT, LoR, J48, and Stochastic Gradient Descent. As a result, the RF classifier outperformed all others with an accuracy of 97.08%.

A study by Alqaseer et al. [17], the authors present the critical need for accurate and timely detection of epileptic seizures, using an algorithm that leverages Discrete Cosine Transformation (DCT) type II to transform EEG signals into the frequency-domain, extracting energy features from 16 sub-bands. Data is segmented into non-overlapping 1-second frames, and based on Euclidean distance, the K-Nearest Neighbor (KNN) model is utilized to recognize those frames that are either ictal (seizure) or interictal (non-seizure). Testing on 21 patients from the CHB-MIT dataset yields an impressive average F1-score of 93.12, with a low False-Positive Rate (FPR) average of 0.07.

A study by Savadkoohi et al. [18], the authors employed machine learning methods to detect brain electrical activities and patterns from an epileptic Electroencephalogram (EEG) that indicates an epileptic seizure is imminent. The dataset used a signal record from 5 healthy and 5 epileptic patient volunteers. To select the required feature from the dataset T-test and SFFS were applied. Additionally, the SVM and KNN algorithms are employed to differentiate between preprocessed EEG signals, with SVM exhibiting a slight performance advantage over KNN. SVM accuracy result is 100%, while KNN accuracy is 99.5%.

A study by Usman et al. [19], proposed a deep learning technique to predict epileptic seizures. Convolution neural networks (CNN) were applied as a feature extraction method, while Support vector machines (SVM) were utilized as the classification approach to differentiate between interictal and preictal states. The model is employed on a dataset of 24 subjects of scalp EEG. The result of the study was successful and accomplished a sensitivity of 92.7% and specificity 90.8%. Nonetheless, this model only tested for epileptic patients, it would improve the model to test it on non-epileptic patients that who experience these seizures.

LIU et al. [20] developed a research paper that aims to develop a prediction model for epileptic seizures using multi-view convolutional neural networks (CNNs). By leveraging the power of CNNs in processing spatial and temporal features. The researchers performed experiments on kaggle dataset to enhance the accuracy as

well as the reliability of seizure prediction. This paper has been achieved with several key steps. Firstly, the researchers preprocessed the EEG signals, applying techniques such as data augmentation, PCA and FFT. Subsequently, they divided the preprocessed data into multiple views to capture different aspects of the EEG signal, possibly focusing on different EEG channels or spectral representations. Next, a multi-view CNN architecture was developed and trained on the dataset to predict epileptic seizures. Furthermore, on two subjects from the CHB-MIT scalp EEG dataset, the suggested model attained (AUC) values of 0.82 and 0.89.

A study written by Usman et. al. [21], the study introduces a robust machine learning model for predicting epileptic seizures, focusing on effective EEG signal preprocessing and feature extraction. The approach involves converting multiple EEG channels into a surrogate signal and applying empirical mode decomposition (EMD) to enhance signal quality. Various features, including approximate entropy, entropy, spectral, Hjorth parameters, and statistical moments, are extracted, providing valuable discriminative information. For distinguishing between preictal and interictal states, the SVM classifier is employed. Moreover, the results on the CHBMIT dataset, which contains scalp EEG signals from 22 subjects, demonstrate a notably improved true positive rate of 92.23% for detecting the preictal state, surpassing conventional methods.

Kornek et al. [22] proposed a research paper aiming to provide a solution for predicting epileptic seizures, which would enhance patient care and allow for timely intervention, using deep learning and big data techniques. The researchers used iEEG data from ten patients obtained from a seizure advisory system. The researchers first trained a deep learning classifier is used to discriminate between preictal (seizure-prone) and interictal (non-seizure) data. They next evaluated the classifier's performance using held-out iEEG data from all patients, comparing it to a random predictor. Additionally, the prediction system was fine-tuned to prioritize either sensitivity (the ability to detect seizures) or time in warning. The researchers proved that the prediction system may be deployed on an ultra-low-power neuromorphic chip for autonomous operation in a wearable device. The results revealed that the prediction system outperformed a random predictor by 42% for all patients, with a mean sensitivity of 69% and a mean time in warning of 27%.

The review of the literature on the prediction of early epileptic seizures revealed that the majority of earlier research on the topic focused on imaging datasets containing EEG signals, largely ignoring the possibilities of clinical data. Furthermore, previous research faced limitations, including small sample sizes and insufficient clinical data. Consequently, the goal of this research is to reduce this gap by building a machine-learning model for more accurate and dependable prediction of epileptic seizures using clinical data. In addition, the research will utilize explainable AI (XAI) methodologies to establish dependability and transparency in the prediction model, hence guaranteeing its successful implementation by healthcare professionals in diverse environments.

### 3.Materials and Methods

This study developed a pre-emptive model for diagnosing Epilepsy by detecting epileptic seizures using Python programming language. The materials and methods utilized in this study involved several key steps. The dataset was first imported using the Pandas package, and then its original contents and structure were analyzed. Preprocessing steps included handling duplicate entries and assessing missing values. Visualization techniques such as heatmaps were employed to visualize the distribution of missing values. Furthermore, specific features were explored using histograms and box plots to identify outliers and understand their distribution. For features exhibiting outliers, the median was selected for imputation to mitigate the influence of outliers. Categorical features with missing values were imputed using the most frequent strategy. Additionally, certain categorical features underwent data cleaning to standardize values. Following

preprocessing, label encoding was applied to convert categorical features into numerical representations. The dataset was split into 70% training and 30% testing sets, and feature scaling was performed using standardization. Eight machine learning methods are used: K-Nearest Neighbors, Gradient Boosting, eXtreme Gradient Boosting, Random Forest, Support Vector Machine, Logistic Regression, Decision Tree, and AdaBoost, were implemented using GridSearchCV for hyperparameter tuning. The model performance was evaluated using accuracy scores, confusion matrices, and classification reports to assess predictive performance on both training and testing datasets. Once the top-performing model was chosen, eXplainable Artificial Intelligence (XAI) methods such as Feature Importance and LIME were employed to investigate the model's decision-making mechanism in greater depth. Figure 1 shows this complete procedure.

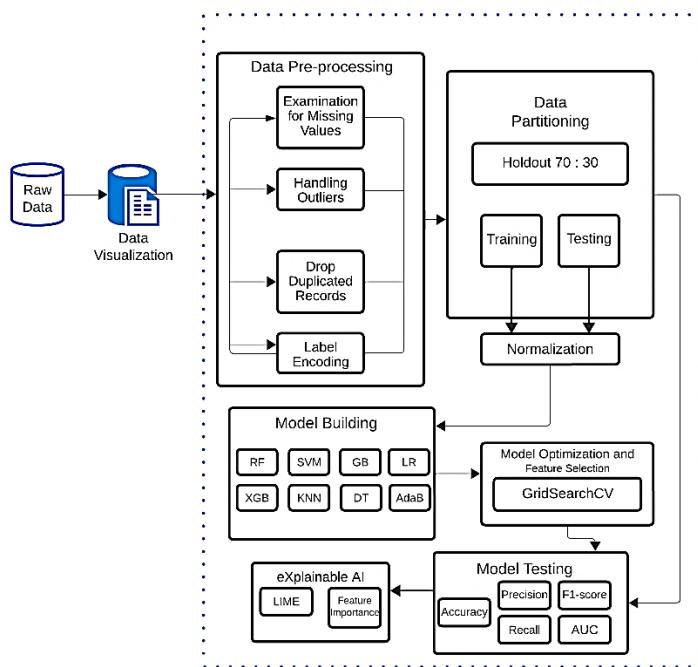


Figure 158 The proposed framework for the pre-emptive diagnosis of Epileptic Seizures

### 3.1 Data Description

The dataset comprised individuals sourced from a continuous stream of referrals to both the Functional Neurological Disorders Clinic and the University of Colorado (CU) Epilepsy Monitoring Unit (EMU) during the period spanning January 1, 2017, to May 15, 2019 [7]. These patients were mostly identified by a confirmed video-electroencephalography (vEEG) diagnosis of epilepsy or a history of seizures (DS) which is referred to dissociative seizures or non-epileptic seizures, with vEEG evaluations performed both inside and outside the medical system. Patients were randomly chosen to form two groups, comprising individuals with dissociative seizures (DS) and epileptic seizures (ES). Additionally, patients with both DS and ES diagnoses were included due to their clinical importance and limited data availability.

Numerous demographic, clinical, and diagnostic variables are included in the dataset. These include gender, seizure characteristics, clinical and medical history, records of traumas or traumatic occurrences, and other relevant information. Table 1 lists all the features and their corresponding types. For research purposes, this dataset offers a thorough overview of the health status of the patients, enabling in-depth examination and analysis.

Feature	Type
RRID	int64
Diagnosis	object
Sex	object

#non psych comorbidities	int64
# Prior AEDs	int64
Asthma	int64
Migraine	int64
Chronic Pain	int64
Diabetes	int64
non metastatic cancer	int64
Number of non-seizure non psych medication	int64
# Current AEDs	int64
Date Onset	object
Date Admit	object
Baseline (sz freq)	float64
Median duration of seizures	float64
# of seizure types	Int64
Injury with seizure	object
Catamenial	object
Trigger of sleep deprivation	object
Aura	object
Ictal Eye Closure	object
Ictal hallucinations	object
Oral automatisms	object
Incontinence	object
Limb automatisms	Object
Ictal tonic-clonic	object
Muscle twitching	Object
Hip thrusting	Object
Post-ictal fatigue	Object
Any head injury	Object
Psych traumatic events	Object
Concussion w/o LOC	Object
Concussion w/LOC	Object
Severe TBI (LOC>30min)	Object
Opioids? Yes/no	Int64
Sex abuse	Object
Physical Abuse	Object
Rape	Object

Table 165 Features' description.

### 3.2 Statistical Analysis

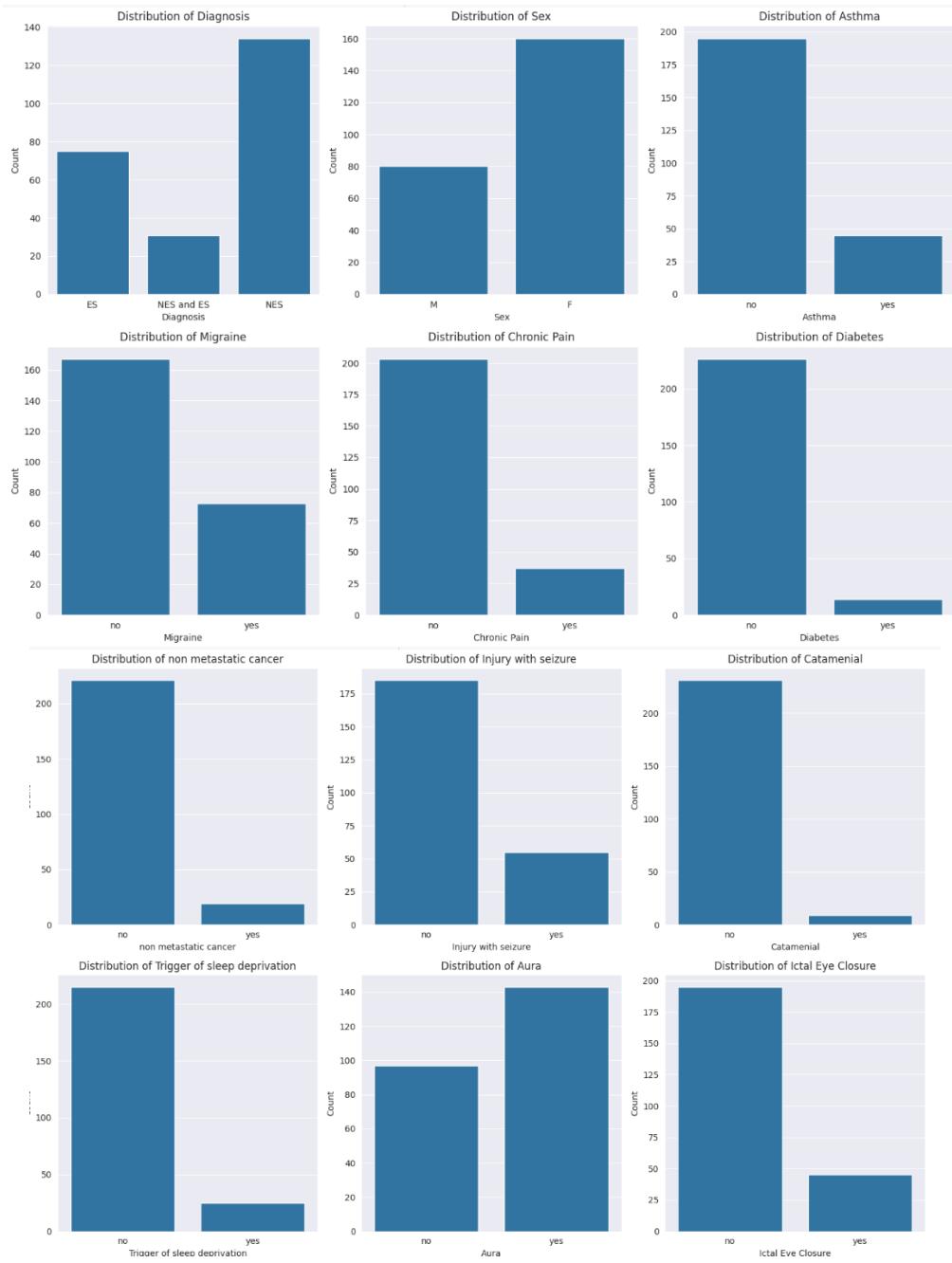
This section illustrates the characteristics of each attribute in the dataset. This step is essential to determine the most suitable pre-processing techniques for optimal modeling methods. Both categorical and numerical attributes will be explored as both hold a significant value for forthcoming modeling stages.

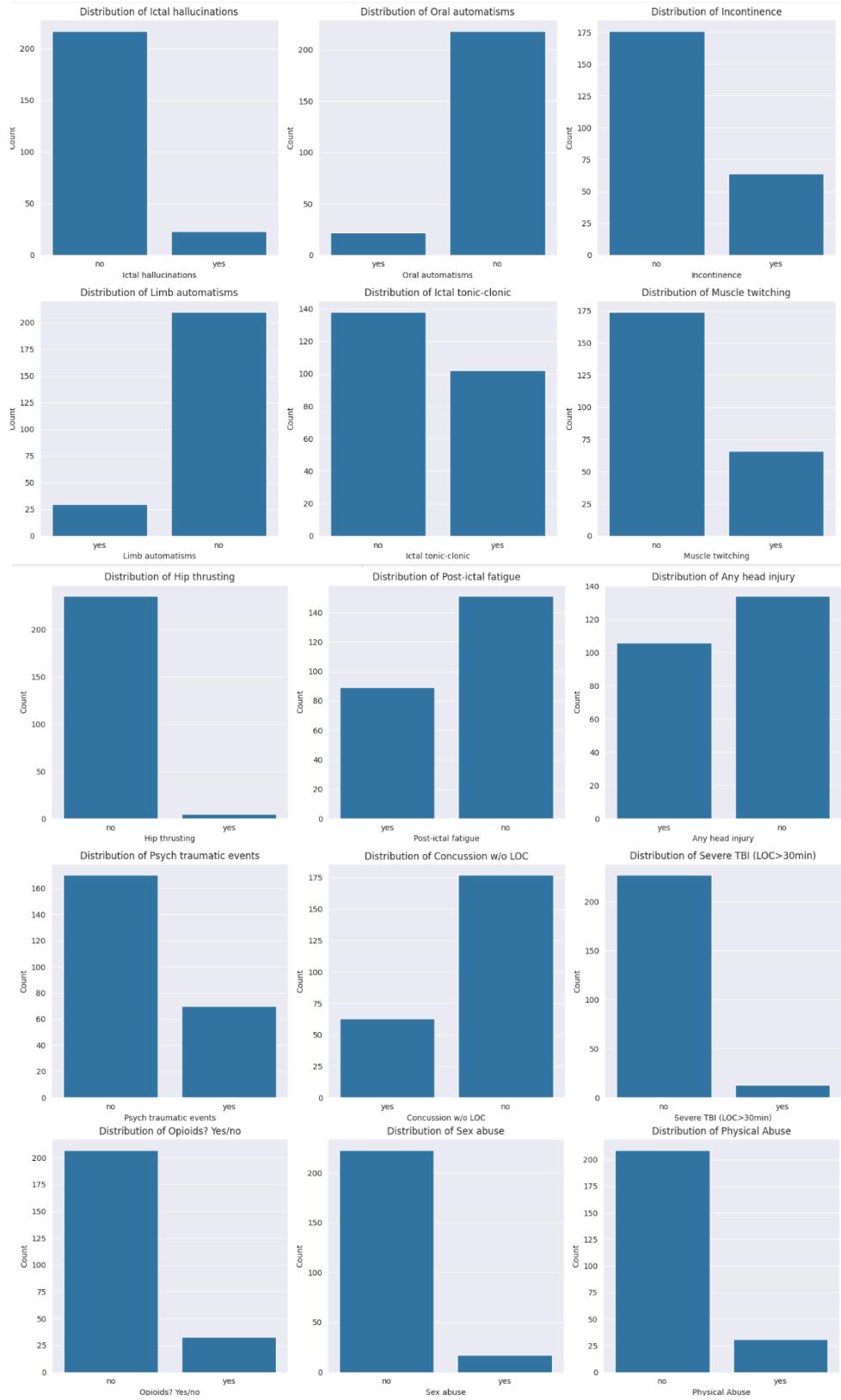
Various statistical matrices were used for numerical attributes. These metrics offer deep insights into the inherent characteristics of the data. Table 2 represents a detailed statistic for numerical features.

Feature	Mean	Standard deviation	Min	25 <sup>th</sup> quartile	50 <sup>th</sup> quartile	75 <sup>th</sup> quartile	Max	Missing value counts
Number of non-psych comorbidities	4.58	4.98	0.00	1.00	3.00	6.00	31.00	0
Number of Prior AEDs	1.71	2.32	0.00	0.00	1.00	3.00	14.00	0
Number of non-seizure non psych medication	5.41	6.20	0.00	1.00	3.00	8.00	44.00	0
Number of Current AEDs	1.77	1.34	0.00	1.00	2.00	3.00	6.00	0
Baseline (sz freq)	24.95	41.16	0.00	3.00	10.00	30.00	308.00	5
Median duration of seizures	5.17	10.50	0.05	1.00	2.00	5.00	65.00	12
Number of seizure types	2.19	1.35	1.00	1.00	2.00	3.00	10.00	0

Table 166 Statistical analysis of numerical features

Figure 2 shows the distribution of various categorical features of patients. The distribution of the class named diagnosis which is the target class shows that 134 patients have non-epileptic seizures, 75 have epileptic seizures, and 31 have been diagnosed with both. The dataset contains 160 female patients and 80 male patients. 45 patients have asthma while 195 are not infected with it. A total of 73 patients suffered from migraine pain and 167 did not. 37 patients experienced chronic pain and 203 did not. The patients who have diabetes are 14 and 226 have not been diagnosed with it. Patients who have non metastatic cancer are 19 while 221 do not have this type of cancer. The seizures cause injury for 55 patients, on the other hand, 182 did not face such a problem. A majority of 230 patients haven't experience catamenial epilepsy only 9 patients have experience it. 213 have no sleep deprivation while 25 have trigger sleep deprivation. The patients who feel the aura before the seizure start are 143 but 97 have not feel the aura. 195 patients haven't close their eyes during the seizure, 45 have closed their eyes. 23 patients who experienced seizures did suffer from hallucinations during the seizure while 217 have not experienced it. Through the seizure 22 have oral automatisms and 218 have not. A total of 64 patients has Incontinence and 176 have not. 30 have their limb automatisms while 210 have normal limb. During the seizure 102 experience tonic-colonic phase and 138 did not. 66 patients have their muscle twitching and 174 have not. The majority of the patients did not have their hip thrusting with total of 235 while 5 did have. After the seizure 89 felt fatigue while 151 did not feel fatigue. 106 have an injury on their head while 134 have not. 170 patients have not experienced any Psychological traumatic event while 70 have experienced. A total of 48 patients has experience concussion with loss of consciousness and 192 have not. 13 patients have lost their consciousness for more than 30 minutes while 227 have not experienced such a condition. 33 patients have used Opioids while 207 have not. 17 and 31 patients have experienced sex or physical abuse respectfully while 223 and 209 have not. 7 patients have been raped and 209 did not. the patients who have experienced concussion without losing their consciousness are 63 while 177 have loss it.





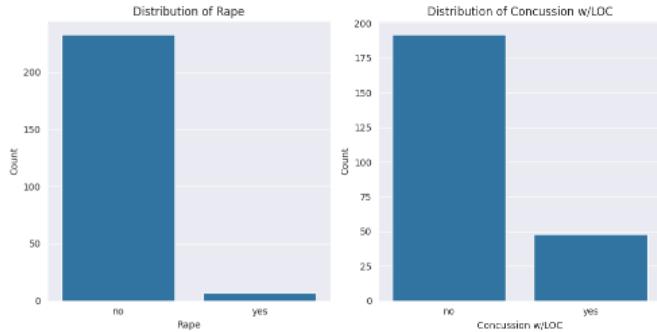


Figure 159 distribution of categorical features

### 3.3 Data Pre-processing

Transforming raw data into clean and usable data is a significant step before starting modeling. Raw data usually contains missing values, outliers, inconsistent values, and noise. Thus, discarding these problems is essential to enhance the quality and performance of the model [23]. To achieve this goal, several libraries were used. In the beginning, the dataset contains 39 features including the target class and 241 instances. Columns with date and ID values were dropped using the function `drop()`. Duplicates and rows with lots of null values were also dropped. After these steps, the dataset contains 36 features and 240 instances.

Missing data is one of the most common problems with datasets, it happens during the collection of data for various reasons. Missing values cause the size of the dataset to be smaller. To deal with this issue there are several solutions such as ignoring it or filling in the data manually [24]. In this study, there were two types of missing values, numerical and categorical. Each type was dealt with differently, for the numerical values median value was used because the distribution of the numerical data was skewed. Equation 1 shows a mathematical representation of the median where  $n$  is the total number of instances in the column that have missing values. On the other hand, missing categorical values were replaced by the most frequent value in that column.

$$\text{Median} = \frac{(n+1)}{2} \quad (1)$$

For further preprocessing and due to the imbalance in the target class, the SMOTE-Tomek method was used to resolve this issue. This sampling technique is a combination of two methods called SMOTE and Tomeklink, it was developed to avoid the drawbacks of these two techniques. First SMOTE will create new data in the minority class, and then Tomeklink will follow by eliminating instances that are close to the minority class from the majority class. Using this method can enhance the accuracy of the model [25]. Lastly, StandardScaler was utilized to standardize features, this works by removing the mean and scaling to unit variance. Equation 2 shows the formula for calculating the standard score for sample  $x$ , where  $u$  is the mean and  $s$  is representing the standard deviation [26].

$$Z = \frac{(x-u)}{s} \quad (2)$$

### 3.4 Description of Utilized Machine Learning Algorithms

#### 3.4.1 K-Nearest Neighbors (KNN):

KNN is a simple, instance-based learning algorithm used for classification and regression tasks. It classifies data points based on the majority class of their  $k$ -nearest neighbors in the feature space. It's non-parametric and effective for small datasets [27].

### **3.4.2 Gradient Boosting:**

Gradient Boosting is an ensemble learning technique that builds decision trees sequentially, each correcting the errors of its predecessor. It minimizes a loss function by greedily adding weak learners. It's widely used in both regression and classification problems due to its high predictive accuracy [28].

### **3.4.3 eXtreme Gradient Boosting (XGBoost):**

XGBoost is an optimized version of gradient boosting, known for its scalability and speed. It employs a more regularized model formalization to control overfitting and provides better performance. It's highly efficient for large datasets and has become a popular choice in various machine learning competitions [29].

### **3.4.4 Random Forest:**

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of individual trees. It offers robustness to overfitting and high accuracy, making it suitable for a wide range of tasks [30].

### **3.4.5 Support Vector Machine (SVM):**

SVM is a supervised learning algorithm that analyzes data and recognizes patterns, used for classification and regression analysis. It works by finding the hyperplane that best separates different classes in the feature space. SVM is effective in high-dimensional spaces and in cases where the number of dimensions exceeds the number of samples [31].

### **3.4.6 Logistic Regression:**

Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. It's primarily used for binary classification problems and estimates the probability that a given input belongs to a particular class. Despite its name, it's a classification algorithm rather than a regression algorithm [32].

### **3.4.7 Decision Tree:**

Decision Tree is a supervised learning algorithm used for classification and regression tasks. It partitions the data into subsets based on the value of the features and makes predictions by following the tree from the root to the leaf nodes. It's intuitive, easy to understand, and capable of handling both numerical and categorical data [33].

### **3.4.8 AdaBoost:**

AdaBoost, short for Adaptive Boosting, is an ensemble learning method that combines multiple weak learners to create a strong classifier. It assigns higher weights to misclassified data points, allowing subsequent weak learners to focus more on difficult cases. It's particularly effective in improving the performance of classifiers in situations where simple models perform poorly [34].

## **3.5 Performance measure**

Performance measure is an essential step to evaluate the model's ability to differentiate between the patient's state and which type of seizure they have. To achieve this goal, various types of methods were used. Including accuracy, recall, Precision, and F1-score. Furthermore, confusion matrices were utilized for more evaluation. This matrix contains True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP: indicates the patients who were accurately classified as diagnosed with epileptic seizure (ES), non-epileptic seizure (NES), or both epileptic seizure and non-epileptic seizure (ES and NES).

TN: indicates the patients who were correctly classified as not belonging to one of the three classes.

FP: indicates the patients who were classified as diagnosed with ES, NES, or ES and NES but belong to another class.

FN: indicates the patients who were classified as not being part of some class but belong to that class.

The accuracy used to measure how many patients have their seizure type (ES, NES, or ES and NES) classified correctly from the total number of patients. This is shown in equation 3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

The ratio of the number of patients correctly classified to some class (TP), to the sum of both TP and FP is called precision which is shown in equation 4.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

The ratio of the number of patients accurately classified to that class (TP), to the sum of TP and FN is called recall. Equation 5 provides a mathematical representation of it.

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

F1-score can be calculated by using precision and recall which is shown in equation 6.

$$F1 - score = \frac{2(Precision \times Recall)}{Precision + Recall} \quad (6)$$

### 3.6 Optimization strategy

Selecting the best hyperparameter for each algorithm has a significant impact on the algorithm's performance. Proper hyperparameter leads to the best possible performance and accuracy. To achieve this goal GridSearchCV method was utilized, this method works by giving a set of various values. The method will try and combine all the values until it finds the best combination. Table 3 below shows the best hyperparameter for each algorithm with all the features.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	None
	Min_samples_leaf	1
	Min_samples_split	2
	N_estimators	100
SVM	C	1
	Gamma	Scale
	kernel	Rbf
K-NN	algorithm	Auto
	N_neighbors	7
	weights	Distance
Gboost	Learning_rate	0.01
	Max_depth	5
	N_estimators	100
XGboost	Gamma	0
	Learning_rate	0.1
	Max_depth	7
Logistic Regression	N_estimators	100
	C	0.1
	Solver	Liblinear
Decision Tree	Max_depth	5
	Min_samplesleaf	2
	Min_samples_split	2
AdaBoost	Learning rate	0.01
	N_estimators	150

Table 167 the best hyperparameter selected on the original data

Table 4 shows the best hyperparameter with 10 features selected on the original data.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	10
	Min_samples_leaf	2
	Min_samples_split	2
	N_estimators	100
SVM	C	0.1
	Gamma	Scale
	kernel	linear
K-NN	algorithm	Auto
	N_neighbors	7
	weights	uniform
Gboost	Learning_rate	0.01
	Max_depth	3
	N_estimators	150
XGboost	Gamma	0
	Learning_rate	0.01
	Max_depth	3
	N_estimators	200
Logistic Regression	C	0.1
	Solver	Newton-cg
Decision Tree	Max_depth	5
	Min_samplesleaf	2
	Min_samples_split	2
AdaBoost	Learning rate	0.01
	N estimators	150

Table 168 the best hyperparameter selected on the original data with 10 features selected

Table 5 represent the best hyperparameter with 20 features selected on the original data.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	10
	Min_samples_leaf	1
	Min_samples_split	5
	N_estimators	200
SVM	C	1
	Gamma	Scale
	kernel	Rbf
K-NN	algorithm	Auto
	N_neighbors	5
	weights	Distance
Gboost	Learning_rate	0.01
	Max_depth	3
	N_estimators	200
XGboost	Gamma	0
	Learning_rate	0.1
	Max_depth	5
	N_estimators	150
Logistic Regression	C	0.1
	Solver	Liblinear
Decision Tree	Max_depth	5
	Min_samplesleaf	2

	Min_samples_split	5
AdaBoost	Learning rate	0.01
	N estimators	150

Table 169 the best hyperparameter selected on the original data with 20 features selected

Table 6 shows the best hyperparameter with 30 features selected on the original data.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	10
	Min_samples_leaf	2
	Min_samples_split	2
	N_estimator	200
SVM	C	1
	Gamma	Scale
	kernel	Rbf
K-NN	algorithm	Auto
	N_neighbors	5
	weights	uniform
Gboost	Learning_rate	0.01
	Max_depth	3
	N_estimator	200
XGboost	Gamma	0.3
	Learning_rate	0.1
	Max_depth	7
Logistic Regression	N_estimator	100
	C	0.1
	Solver	Liblinear
Decision Tree	Max_depth	5
	Min_samplesleaf	2
	Min_samples_split	2
AdaBoost	Learning rate	0.01
	N estimators	150

Table 170 the best hyperparameter selected on the original data with 30 features selected

Table 7 below shows the selected hyperparameters for each algorithm with oversampled data and 10 features selected.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	None
	Min_samples_leaf	1
	Min_samples_split	5
	N_estimator	150
SVM	C	10
	Gamma	Scale
	kernel	Rbf
K-NN	algorithm	Auto
	N_neighbors	5
	weights	Distance
Gboost	Learning_rate	0.1
	Max_depth	5
	N_estimator	150
XGboost	Gamma	0.3
	Learning_rate	0.1

	Max_depth	3
	N_estimator	100
Logistic Regression	C	1
	Solver	Liblinear
Decision Tree	Max_depth	5
	Min_samples_leaf	1
	Min_samples_split	2
AdaBoost	Learning rate	0.01
	N estimators	100

Table 171 the best hyperparameter selected on the oversampled data with 10 features selected

Table 8 represent the hyperparameters with oversampled data and 20 features selected.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	None
	Min_samples_leaf	1
	Min_samples_split	5
	N_estimator	100
SVM	C	10
	Gamma	Scale
	kernel	Rbf
K-NN	algorithm	Auto
	N_neighbors	3
	weights	Distance
Gboost	Learning_rate	0.1
	Max_depth	7
	N_estimator	200
XGboost	Gamma	0
	Learning_rate	0.1
	Max_depth	3
	N_estimator	200
Logistic Regression	C	0.1
	Solver	Liblinear
Decision Tree	Max_depth	5
	Min_samplesleaf	2
	Min_samples_split	5
AdaBoost	Learning rate	0.1
	N estimators	100

Table 172 the best hyperparameter selected on the oversampled data with 20 features selected

Table 9 illustrate the hyperparameters with oversampled data and 30 features selected.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	None
	Min_samples_leaf	1
	Min_samples_split	2
	N_estimator	150
SVM	C	10
	Gamma	Scale
	kernel	Rbf
K-NN	algorithm	Auto

	N_neighbors weights	5 Distance
Gboost	Learning_rate	0.1
	Max_depth	5
	N_estimator	200
XGboost	Gamma	0
	Learning_rate	0.1
	Max_depth	3
Logistic Regression	N_estimator	150
	C	0.1
	Solver	Newton-cg
Decision Tree	Max_depth	5
	Min_samplesleaf	2
	Min_samples split	2
AdaBoost	Learning rate	0.1
	N estimators	150

Table 173 the best hyperparameter selected on the oversampled data with 30 features selected

Table 10 shows the selected hyperparameters with oversampled data and all features.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	None
	Min_samples_leaf	1
	Min_samples_split	5
	N_estimator	100
SVM	C	10
	Gamma	Scale
	kernel	Rbf
K-NN	algorithm	Auto
	N_neighbors	3
	weights	Distance
Gboost	Learning_rate	0.1
	Max_depth	5
	N_estimator	100
XGboost	Gamma	0
	Learning_rate	0.1
	Max_depth	5
Logistic Regression	N_estimator	200
	C	0.1
	Solver	Liblinear
Decision Tree	Max_depth	10
	Min_samplesleaf	1
	Min_samples split	2
AdaBoost	Learning rate	1
	N estimators	150

Table 174 the best hyperparameter selected on the oversampled data with all features

#### 4. Empirical Results

Within this section, we present the outcomes derived from our developed models after the implementation of GridSearchCV on both the original dataset and the sampled data using SMOTETomek, which is a method

used to balance the class distribution within the dataset by oversampling the minority class and under sampling the majority class.[35] Following the acquisition of optimal hyperparameters and model training through stratified 10-fold cross-validation and utilizing all the 36 features, the following section delineates a comprehensive examination of the results obtained, as presented in Table 11.

<b>Classifier</b>	<b>Dataset</b>	<b>Testing Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>
<b>Logistic Regression</b>	<b>Original</b>	<b>87.67%</b>	<b>88.00%</b>	<b>88.00%</b>	<b>88.00%</b>
	<b>Using SMOTETomek</b>	<b>79.45%</b>	<b>85.00%</b>	<b>79.00%</b>	<b>81.00%</b>
	Original	80.82%	75.00%	81.00%	78.00%
Random Forest	Using	78.08%	79.00%	78.00%	79.00%
	SMOTETomek				
	Original	86.30%	87.00%	86.00%	85.00%
SVM	Using	78.08%	78.00%	78.00%	78.00%
	SMOTETomek				
	Original	80.82%	82.00%	81.00%	80.00%
KNN	Using	60.27%	70.00%	60.00%	62.00%
	SMOTETomek				
	Original	82.19%	82.00%	82.00%	82.00%
Gradient Boosting	Using	72.60%	77.00%	73.00%	75.00%
	SMOTETomek				
	Original	79.45%	80.00%	79.00%	80.00%
XGBoost	Using	78.08%	83.00%	78.00%	80.00%
	SMOTETomek				
	Original	76.71%	78.00%	77.00%	77.00%
Decision Tree	Using	57.53%	69.00%	58.00%	61.00%
	SMOTETomek				
	Original	76.71%	76.00%	77.00%	76.00%
AdaBoost	Using	78.08%	83.00%	78.00%	80.00%
	SMOTETomek				

Table 175 The results of the proposed models before and after sampling was applied.

The results show that, although the SMOTETomek oversampling strategy was applied in an attempt to improve class imbalance, classifier performance was not always improved by this method. SMOTETomek reduced the testing accuracy of several classifiers (Random Forest, SVM, KNN, Logistic Regression, Gradient Boosting, XGBoost, and Decision Tree) when compared to the original dataset. This finding indicates that, for this dataset, the original, non-oversampled results are considered more reliable. With the best testing accuracy of 87.67%, a precision of 88.00%, recall of 88.00%, and F1-score of 88.00% among the original results, Logistic Regression was the best option for modeling this dataset without using the SMOTETomek technique. Following that, SVM obtained 86.30% accuracy, 87.00% precision, 86.00% recall, and 85.00% F1-score. This analysis emphasizes how important it is to carefully assess the effects of the SMOTETomek technique and choose the best classifier when working with unbalanced datasets. Figure 3 shows the number

of samples of each class of the target feature before and after applying SMOTETomek. (Class 0 represents NES, class 1 represents NES, class 2 represents NES and ES.)

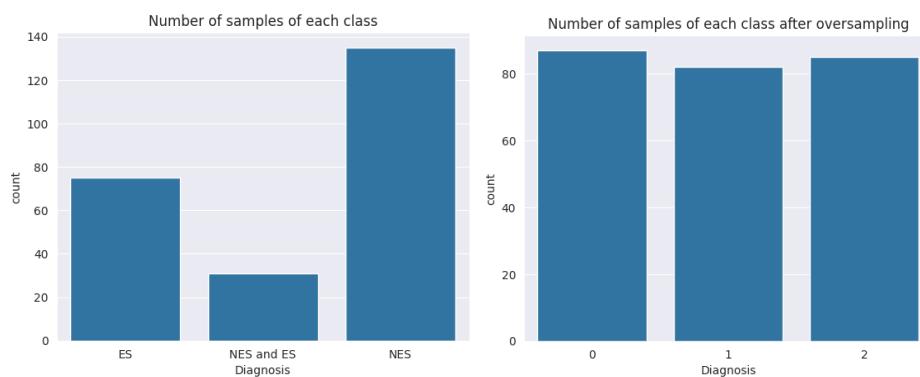


Figure 160 Number of samples of each class of the target feature before and after applying SMOTETomek.

#### 4.1 Results with Feature Selection

The feature selection utilized is based on the statistical test chi-squared to check if two categorical variables have significant relationship [36]. In this research, a dataset of 36 attributes that were taken from an ongoing stream of referrals to the FND Clinic and CU EMU was tested with the chi-squared test [7]. Comparing observed and expected frequencies inside a contingency table is the basis of the test. By examining the correlation between the components, the chi-square test addresses feature selection issues. The purpose of this analysis is to ascertain whether the correlation between two sample categorical variables accurately reflects the correlation between them in the population [36]. The following is the chi-squared equation 7, where O is the observed value and E is the expected value.

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \quad (7)$$

The SelectKBest is the methodology used in this study to select the best K number of features from the chi-squared test result. The chi-squared calculate the relationship between the target variable ‘Diagnosis’ and each feature in the dataset. The result is then sorted, so the SelectKBest take the highest K features in the selection process. Table 12 shows the result of the ML algorithms with feature selection.

Feature Subsets	10	20	30	All features (36)
<b>LoR</b>	80.82%	<b>86.30%</b>	<b>86.30%</b>	<b>87.67%</b>
<b>RF</b>	80.82%	83.56%	83.56%	80.82%
<b>SVM</b>	78.08%	84.93%	<b>86.30%</b>	86.30%
<b>KNN</b>	80.82%	79.45%	76.71%	80.82%
<b>GB</b>	<b>84.93%</b>	80.82%	82.19%	76.71%
<b>XGBoost</b>	80.82%	83.56%	82.19%	79.45%
<b>DT</b>	80.82%	71.23%	75.34%	71.23%
<b>AdaBoost</b>	76.71%	76.71%	76.71%	76.71%

Table 176 result with feature selection

#### 4.2 Further Discussion of the Results

The confusion matrices offer valuable insights into the performance of various machine learning models in classifying instances into three distinct categories: ES (epileptic seizure), NES (non-epileptic seizures), and a combined category of ES & NES.

Beginning with RandomForest Table 13, it demonstrates a relatively balanced performance, accurately classifying 19 instances of ES, 40 instances of NES, and 0 instance in the ES & NES category. However, there are instances where it struggles to distinguish between ES and ES & NES, as evidenced by the misclassifications present.

Moving to SVM Table 14, it shows a strong performance, correctly classifying 18 instances of ES, 43 instances of NES, and 2 instances in the ES & NES category. This model exhibits fewer misclassifications compared to RandomForest, particularly in the ES category, indicating its robustness in accurately identifying epileptic seizure instances.

KNN Table 15 displays a reasonably balanced performance, accurately classifying 19 instances of ES, 38 instances of NES, and 2 instances in the ES & NES category. However, it shows higher misclassification rates in distinguishing between ES and NES, suggesting potential challenges in generalizing across these categories. GradientBoosting Table 16 showcases competitive performance, accurately classifying 16 instances of ES, 42 instances of NES, and 2 instances in the ES & NES category. Similar to RandomForest and SVM, it faces challenges in distinguishing between ES and NES, leading to some misclassifications.

Similarly, XGBoost Table 17 provides results akin to GradientBoosting, accurately classifying 17 instances of ES, 39 instances of NES, and 2 instances in the ES & NES category. Yet, like GradientBoosting, it struggles to differentiate between ES and NES, resulting in misclassifications.

LogisticRegression Table 18 exhibits a balanced performance, accurately classifying 19 instances of ES, 42 instances of NES, and 3 instances in the ES & NES category. However, it shows slightly higher misclassification rates compared to SVM and RandomForest, particularly in distinguishing between NES and ES & NES.

Moving on to DecisionTree Table 19, it demonstrates varying performance across classes, correctly classifying 20 instances of ES, 35 instances of NES, and 1 instance in the ES & NES category. It shows higher misclassification rates, especially in the NES category, indicating potential limitations in accurately identifying non-epileptic seizure instances.

Finally, AdaBoost Table 20 provides relatively balanced results, accurately classifying 15 instances of ES, 40 instances of NES, and 3 instances in the ES & NES category. However, it faces challenges in distinguishing between ES and ES & NES, leading to some misclassifications.

RandomForest		Predicted		
		ES	NES	ES And NES
Actual	ES	19 (TP)	4 (FN)	0 (FN)
	NES	5 (FP)	40 (TN)	0 (FN)
	ES And NES	2 (FP)	3 (FP)	0 (TN)

Table 178 confusion matrix of SVM algorithm

SVM		Predicted		
		ES	NES	ES And NES
Actual	ES	18 (TP)	5 (FN)	0 (FN)
	NES	2 (FP)	43 (TN)	0 (FN)
	ES And NES	1 (FP)	2 (FP)	2 (TN)

Table 177 confusion matrix of random forest algorithm

KNN		Predicted		
		ES	NES	ES And NES
Actual	ES	19 (TP)	4 (FN)	0 (FN)
	NES	7 (FP)	38 (TN)	0 (FN)
	ES And NES	0 (FP)	3 (FP)	2 (TN)

Table 15 confusion matrix of KNN algorithm

Gradient Boosting		Predicted		
		ES	NES	ES And NES
Actual	ES	16 (TP)	4 (FN)	3 (FN)
	NES	3 (FP)	42 (TN)	0 (FN)
	ES And NES	2 (FP)	1 (FP)	2 (TN)

Table 16 confusion matrix of Gboost algorithm

XGBoost		Predicted		
		ES	NES	ES And NES
Actual	ES	17 (TP)	4 (FN)	2 (FN)
	NES	5 (FP)	39 (TN)	1 (FN)
	ES And NES	2 (FP)	1 (FP)	2 (TN)

Table 18 confusion matrix of Logistic regression algorithm

DecisionTree		Predicted		
		ES	NES	ES And NES
Actual	ES	20 (TP)	3 (FN)	0 (FN)
	NES	7 (FP)	35 (TN)	3 (FN)
	ES And NES	3 (FP)	1 (FP)	1 (TN)

Table 19 confusion matrix of Decision tree algorithm

Logistic Regression		Predicted		
		ES	NES	ES And NES
Actual	ES	19 (TP)	2 (FN)	2 (FN)
	NES	3 (FP)	42 (TN)	0 (FN)
	ES And NES	1 (FP)	1 (FP)	3 (TN)

Table 17 confusion matrix of XGboost algorithm

AdaBoost		Predicted		
		ES	NES	ES And NES
Actual	ES	15 (TP)	6 (FN)	2 (FN)
	NES	4 (FP)	40 (TN)	1 (FN)
	ES And NES	2 (FP)	2 (FP)	1 (TN)

Table 20 confusion matrix of AdaBoost algorithm

The displayed ROC curves in figure 4 represent the performance of various machine learning models applied in a study for the pre-emptive diagnosis of epileptic seizures using clinical data. Each curve plots the true positive rate against the false positive rate for three classes, labeled 0, 1, and 2, at various discrimination thresholds. A model with a curve closer to the top-left corner indicates a higher true positive rate and a lower false positive rate, showcasing better diagnostic accuracy. The area under the curve (AUC) serves as a summary metric, with a value of 1 representing perfect classification and 0.5 representing a random guess.

In these ROC curves, we see that the Support Vector Machine (SVM) model exhibits the most robust discrimination ability among the machine learning models evaluated. The SVM demonstrates high Area Under the Curve (AUC) values for all classes, 0.95 for Class 0, 0.96 for Class 1, and 0.96 for Class 2, implying a strong predictive performance with high true positive rates and low false positive rates. Logistic Regression follows closely, presenting AUCs of 0.95 for Class 0, 0.96 for Class 1, and 0.95 for Class 2, which suggests it is also highly effective at distinguishing between the classes. Random Forest ranks next with AUC values of 0.95, 0.93, and 0.93 for Classes 0, 1, and 2, respectively, showing its considerable predictive capabilities. The XGBoost model, while slightly lower, still shows impressive AUC measures of 0.91, 0.92, and 0.90 for the three classes, indicating a strong classification performance. Other models like Gradient Boosting and AdaBoost also show good performance but with slightly lower AUC values, while the Decision Tree and KNN models show moderate to lower AUCs, indicating more variability in their class distinction capability. These insights facilitate a clear comparison of model efficacy, with SVM and Logistic Regression leading in this context, thereby providing a valuable guide for researchers in selecting optimal models for epileptic seizure prediction.

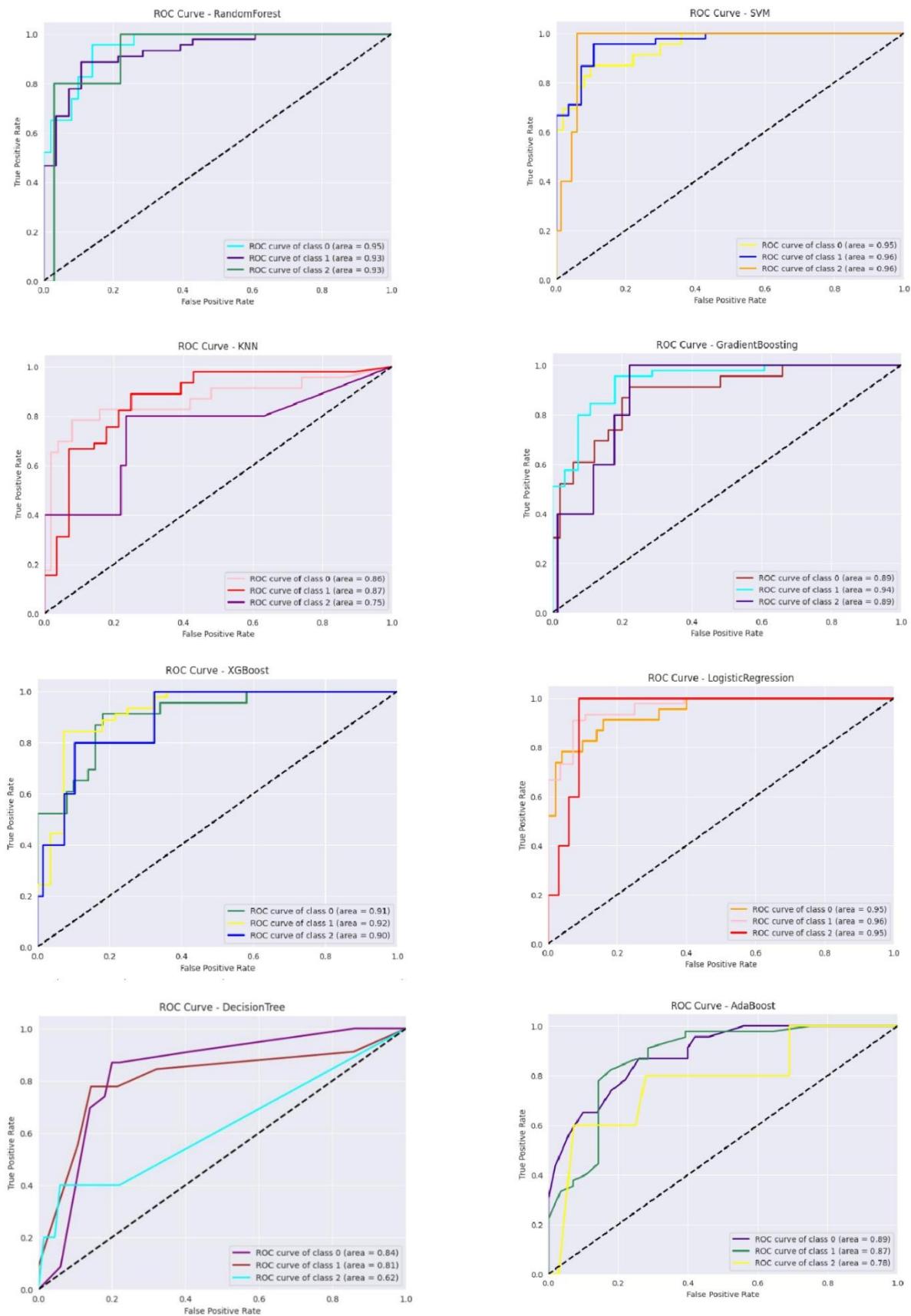


Figure 161 ROC curves of the eight models applied.

## 5. Interpretation of the Final Recommended Model

AI is increasingly used in various industries due to its predictive capabilities. Due to their widespread use, many AI and ML models' inherent ambiguity has become a major obstacle. Often acting as "black boxes," these models make it difficult to understand the logic underlying their predictions. To improve confidence and transparency in the interpretations of AI models, the incorporation of explainable artificial intelligence (XAI) methodologies has become increasingly important.[37] Giving light on the inner workings of artificial intelligence and machine learning models is the aim of XAI. XAI meets the urgent need for interpretability and transparency by providing explanations for model predictions. This enables stakeholders to understand AI-based insights and take confident action based on them. Particularly in the medical field, machine learning epilepsy prediction systems necessitate precise interpretation of the resulting model. XAI facilitates a better understanding of risk factors and important characteristics impacting model outcomes between patients and healthcare professionals. Two XAI techniques, Feature Importance and LIME, were used in this study.

### 5.1. Feature Importance

Feature importance, one of the fundamental approaches of eXplainable Artificial Intelligence (XAI), aims to explain the algorithm's selection of the most important features by giving values that reflect the importance of the input components in their contribution to that decision [38]. This method makes it possible to determine which features significantly influence the model's predictions or classifications, offering insightful knowledge about the inner workings of the model [39]. In the context of epileptic seizure detection, feature importance analysis could assist in identifying the most important patient characteristics for reliably predicting seizure events.

A logistic regression model's coefficients are used to create a feature importance graph, where the direction and magnitude of a feature's influence on the expected result are indicated by the bars' direction to zero. The predicted outcome is positively impacted by features whose bars extend to the right of zero, and negatively by features whose bars extend to the left. [40] The length of the bars indicates the effect's magnitude; The longer the bar, the greater the influence of the feature on the prediction. Figure 5 shows the feature importance plot, the "sex" feature shows the highest positive contribution to the prediction followed by "ictal tonic-clonic". In contrast, "Ictal Eye Closure" shows the highest negative contribution to the prediction followed by "Median duration of seizures".

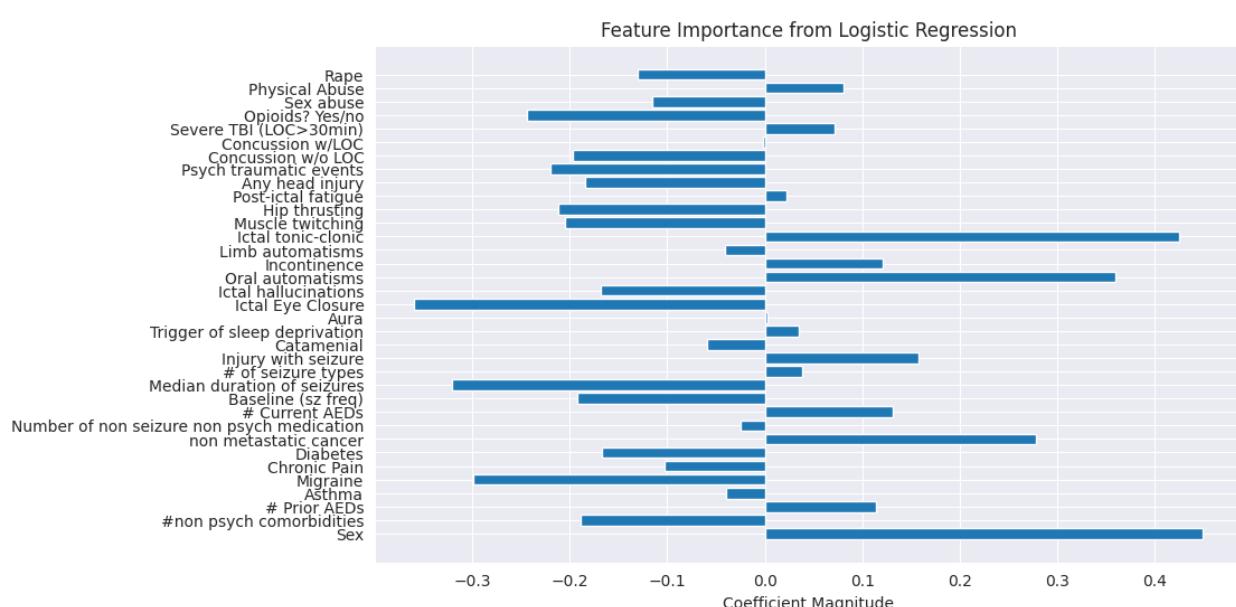
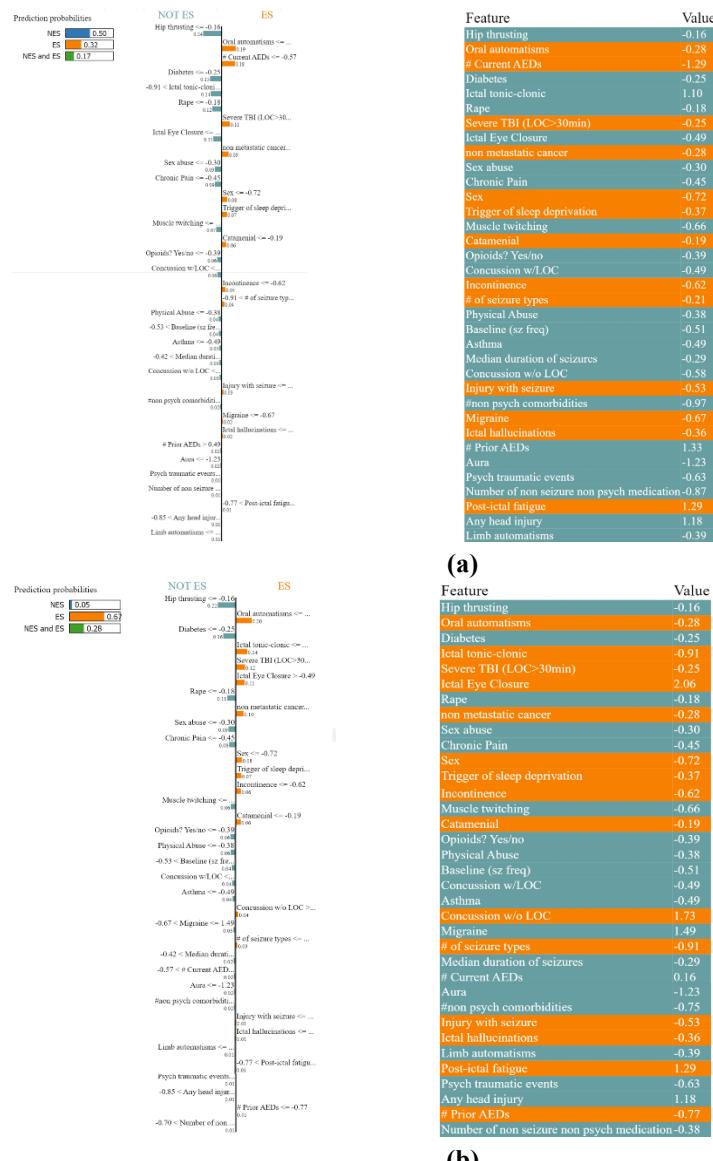


Figure 162 Feature importance of logistic regression.

## 5.2. Local Interpretable Model-Agnostic Explanations (LIME)

LIME is a method that produces local and intelligible explanations for particular situations, so offering a novel solution to the problem of explaining predictions from intricate black box models. [41] Through changing the input data, building a different model, and using random perturbation to determine the importance of features, it sheds light on the decision-making process of the model. Due to its model-agnostic nature, LIME can be used with various models after training, which increases the transparency and trustworthiness of AI applications. [42] Figure [6] shows the LIME plot that gives local prediction probabilities for ES, NES, and (NES and ES) using the Logistic Regression Classifier. The prediction is computed through a comparison of the target variable and the probability values. To show how much each characteristic contributes to the prediction, the feature value table assigns each one a different color code. For instance, ES is blue, NES is orange, and (NES and ES) is green. The features and related values that supported the model's prediction are also highlighted using different weighted feature values.



(b)

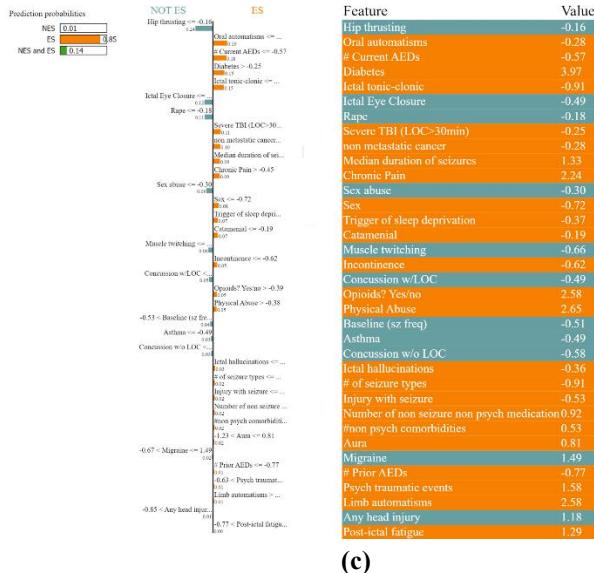


Figure 163 Lime prediction probability for the Logistic Regression model a) NES b) ES c) NES and ES.

## 6. Discussion

An optimism revolution in healthcare has been indicated by the tremendous advancements in technology over the past couple decades, with machine learning (ML) and other emerging technologies playing a key role. When it comes to understanding the enormous amounts of healthcare data that are produced on a daily basis by electronic health records, machine learning truly shines. By examining a wider range of data, this capability enables healthcare practitioners to improve and accelerate the delivery of care [43]. Automation of primary and tertiary healthcare processes has been shown to be significantly possible with the deployment of ML models into healthcare systems. The implementation of intelligent decision-making processes when combined with automation in medical testing has the potential to minimize expenses and time spent on testing, ultimately leading to optimal resource utilization. Furthermore, these technological developments have improved average life expectancy. Overall health outcomes for individuals and communities assure to improve as ML algorithms contribute to more accurate diagnosis, tailored treatment approaches, and effective preventative strategies. The way that machine learning (ML) is changing the healthcare industry allows us to imagine a time where intelligent automation and data-driven insights work together to continuously raise expectations for patient care [44].

In Arab countries, the median incidence of epilepsy is 89.5 per 100,000, while the median lifetime prevalence is 6.9 per 1000 [45]. Individuals with epilepsy frequently experience psychological disorders like anxiety and depression in addition to a higher frequency of physical issues including fractures and bruising from seizure-related traumas. In the same way, people with epilepsy have a three times greater chance of dying before their time than the general population, with low- and middle-income nations as well as rural locations having the highest rates of early mortality [46]. Consequently, pre-emptive diagnosis of epileptic seizures using ML may help to reduce the possible risks.

EEG is a medical procedure used to evaluate brain activity and diagnose various neurological disorders. This is done by recording electrical signals generated by brain cells using electrodes placed on the scalp. Although this process has been used for many years and is generally considered safe to evaluate, it may cause interference or problems in some cases [47]. One common problem is people with seizure disorders having seizures during an EEG. This effect is due to the electrical stimulation applied during the procedure, which may stimulate the nervous system in some people and lead to seizures. In addition, some external factors can affect the EEG results. For example, some medications may interfere with brain activity and affect the recorded electrical signals, making data analysis less accurate. Some of these medications include tranquilizers and antidepressants. Also, low blood sugar can affect brain activity and thus EEG results, making reading

more complicated. It is also known that external conditions such as oily hair or the presence of hair spray can cause interference in the electrical signals read by electromyography devices. As for the cost, the EEG machine is considered average, but not very expensive. However, appropriate training must be provided to technicians and health professionals who perform these tests [48]. Accurate diagnosis and interpretation of EEG data requires a deep understanding of the brain's anatomy and its electrical interactions, and this requires extensive training and specialized skills.

Using clinical and demographic data, such as age, gender, lifestyle behaviors, medical condition history, and patient symptoms, predictive models can identify potential disorders that a patient may suffer from. Although these alternative models do not completely rule out EEG, they may be a potential option, especially in cases where rapid diagnosis is required or in settings where EEG methods are not readily available.

The goal of this research is to enhance the early diagnosis of epileptic seizures by employing machine learning techniques that only require clinical data. This study uses a dataset from the FND Clinic and CU EMU, which comprised 242 patients, to examine the effectiveness of eight machine learning algorithms. Using all available features, the Logistic Regression technique outperformed other algorithms and achieved an accuracy of 87.67%, F1 score, precision, and recall of 88%.

Epileptic seizures can be caused by various features and in our study, "sex" and "ictal tonic-clonic" are most important feature where they have the highest contribution to the model. According to [45], the three risk features that were most commonly found in Arab countries were a history of prenatal infections/insults, paternal consanguinity, and a family history of epilepsy.

The focus of this study is on predicting epileptic seizures using clinical datasets. This has the major advantage in improving the detection process and minimizing the interference. The majority of research concentrated on employing EEG-based datasets, but these types of medical exams may be prone to interference and are not always easily accessible.

## 7. Conclusion

This study focuses on the preemptive diagnosis of epileptic seizures using machine learning algorithms, leveraging diagnostic, clinical, and demographic data. The research aims to address the challenges of early detection and diagnosis of epilepsy, particularly in healthcare settings with limited resources. Eight machine learning methods were employed, the findings indicate that for this specific dataset, the original, non-oversampled results are considered more reliable, with Logistic Regression achieving the best testing accuracy of 87.67% with precision, recall, and F1 values all of 88% without using the SMOTETomek technique. The study emphasizes the importance of carefully considering the impact of oversampling techniques and selecting the most appropriate classifier for handling imbalanced datasets. Additionally, Explainable Artificial Intelligence (XAI) approaches which are Feature importance and LIME were employed to gain a deeper understanding of the top-performing AI model. The "sex" element in the feature importance technique contributes most positively to the prediction, followed by "ictal tonic-clonic." On the other hand, "Ictal Eye Closure" had the most opposite impacts on the prediction followed by "Median duration of seizures".

In future work, the model could be tested in new datasets or with new patients, and the exploration of new classifiers and methods, such as deep learning, could be considered to further improve accuracy. Additionally, the recommendation is to utilize the model in patients by employing Longitudinal Data Analysis, where continuous patient data is collected, to enhance the model's applicability and effectiveness in real-world healthcare settings. This approach would enable the model to adapt and evolve based on ongoing patient data, potentially leading to improved diagnostic accuracy and proactive management of epileptic seizures.

## References

- [1] "Non communicable diseases," World Health Organization, <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases> (accessed Mar. 8, 2024).
- [2] World Health Organization, "Epilepsy," WHO, Feb. 07, 2024. <https://www.who.int/news-room/fact-sheets/detail/epilepsy> (accessed Mar. 1, 2024).
- [3] Klein. E., "Epilepsy: Definition, symptoms, treatment, causes, and more," medicalnewstoday. Nov. 21. 2019 <https://www.medicalnewstoday.com/articles/8947#symptoms> (accessed Mar. 1, 2024).
- [4] Sirven, J. and Schachter, S., "Electroencephalography (EEG)," Epilepsy Foundation. <https://www.epilepsy.com/diagnosis/eeg> (accessed Mar. 1, 2024).
- [5] An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023). "A Comprehensive Review on Machine Learning in Healthcare Industry...". Sensors (Basel, Switzerland), 23(9), 4178. <https://doi.org/10.3390/s23094178>.
- [6] "non-epileptic seizures and dissociative seizures | Epilepsy Society," epilepsysociety. (2023) <https://epilepsysociety.org.uk/about-epilepsy/what-epilepsy/non-epileptic-seizures> (accessed Mar. 1, 2024).
- [7] Lenio, Steven; Kerr, Wesley; Watson, Meagan; Baker, Sarah; Bush, Chad; Rajic, Alexander; Strom, Laura (2021), "Data & Code from Validation of a predictive calculator to distinguish between patients ...", Mendeley Data, V1, doi: 10.17632/cshccr8w3h.1 .
- [8] Chen, W., Wang, Y., Ren, Y., Jiang, H., Du, G., Zhang, J., & Li, J. (2023). An automated detection of epileptic seizures EEG using CNN classifier based on feature fusion with high accuracy. BMC Medical Informatics and Decision Making, 23(1). <https://doi.org/10.1186/s12911-023-02180-w>
- [9] Ode, R., Fujiwara, K., Miyajima, M. et al. "Development of an epileptic seizure prediction algorithm using R–R intervals with self-attentive autoencoder". Artif Life Robotics 28, 403–409 (2022). Doi: <https://doi.org/10.1007/s10015-022-00832-0>
- [10] J.Nazari , A. Nasrabadi , M. Menhaj, and S Raiesdana . "Epilepsy seizure prediction with few-shot learning method". Brain Inform. 2022;9(1):21. Published 2022 Sep 16. doi:10.1186/s40708-022-00170-8
- [11] M. Tawfik, E. Mahyoub, Z. A. Ahmed, N. M. Al-Zidi, and S. Nimbhore, "Classification of epileptic seizure using machine learning and deep learning based on electroencephalography (EEG)," Communication and Intelligent Systems, pp. 179–199, 2022. doi:10.1007/978-981-19-2130-8\_15
- [12] A. M. Hilal et al., "Intelligent Epileptic Seizure Detection and Classification Model Using Optimal Deep Canonical Sparse Autoencoder," Biology, vol. 11, no. 8, p. 1220, Aug. 2022, doi: 10.3390/biology11081220.
- [13] I. Jemal, N. Mezghani, L. Abou-Abbas, and A. Mitiche, "An interpretable deep learning classifier for epileptic seizure prediction using EEG Data," IEEE Access, vol. 10, pp. 60141–60150, 2022. doi:10.1109/access.2022.3176367
- [14] O. Ouichka, A. Echtioui, and H. Hamam, "Deep Learning Models for Predicting Epileptic Seizures Using iEEG Signals," Electronics, vol. 11, no. 4, p. 605, Feb. 2022, doi: 10.3390/electronics11040605.
- [15] S. Usman, S. Khalid, and Z. Bashir," Epileptic seizure prediction using scalp electroencephalogram signals" Sciencedirect, Vol. 41, no.1, p. 211-220. Jan.2021, doi: <https://doi.org/10.1016/j.bbe.2021.01.001>

- [16] Almustafa, K. M. (2020). Classification of epileptic seizure dataset using different machine learning algorithms. *Informatics in Medicine Unlocked*, 21, 100444. <https://doi.org/10.1016/J.IMU.2020.100444>
- [17] M. A. Jumaah, A.I.Shihab and A.A.Farhan, "Epileptic seizures detection using DCT-II and KNN classifier in long -Term EEG Signals"Researchgate, Feb. 2020, doi: [https://www.researchgate.net/publication/339302243\\_Epileptic\\_Seizures\\_Detection\\_Using\\_DCT-II\\_and\\_KNN\\_Classifier\\_in\\_Long-Term\\_EEG\\_Signals](https://www.researchgate.net/publication/339302243_Epileptic_Seizures_Detection_Using_DCT-II_and_KNN_Classifier_in_Long-Term_EEG_Signals)
- [18] M. Savadkoohi, T. Oladunni, and L. Thompson, "A machine learning approach to epileptic seizure prediction using electroencephalogram (EEG) signal," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 3, pp. 1328–1341, 2020. doi:10.1016/j.bbe.2020.07.004
- [19] S. Muhammad Usman, S. Khalid, and M. H. Aslam, "Epileptic seizures prediction using Deep Learning Techniques," *IEEE Access*, vol. 8, pp. 39998–40007, 2020. doi:10.1109/access.2020.2976866
- [20] C. L. Liu, B. Xiao, W. H. Hsiao, and V. S. Tseng, "Epileptic Seizure Prediction With Multi-View Convolutional Neural Networks," *IEEEExplore*, [https://ieeexplore.ieee.org/document/8910555?denied=\(accessed Oct. 8, 2023\)](https://ieeexplore.ieee.org/document/8910555?denied=(accessed Oct. 8, 2023))
- [21] S. Usman, M. Usman, and S. Fong, "Epileptic Seizures Prediction Using Machine Learning Methods", *Computational and Mathematical Methods in Medicine*, vol. 2017, Article ID 9074759, 10 pages, 2017. <https://doi.org/10.1155/2017/9074759>
- [22] I. K. Kornek et al., "Epileptic Seizure Prediction Using Big Data and Deep Learning: Toward a Mobile System." IBM, Australia, Dec. 12, 2017.
- [23] Review\_of\_Data\_Preprocessing\_Techniques. (n.d.).
- [24] Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. In *Korean Journal of Anesthesiology* (Vol. 70, Issue 4, pp. 407–411). Korean Society of Anesthesiologists. <https://doi.org/10.4097/kjae.2017.70.4.407>
- [25] Hairani, H., Anggrawan, A., & Priyanto, D. (n.d.). INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION journal homepage : [www.joiv.org/index.php/joiv](http://www.joiv.org/index.php/joiv) INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link. [www.joiv.org/index.php/joiv](http://www.joiv.org/index.php/joiv)
- [26] Scikit-Learn, "sklearn.preprocessing.StandardScaler — scikit-learn 0.21.2 documentation," *Scikit-learn.org*, 2019. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
- [27] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- [28] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5), 1189-1232.
- [29] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).
- [30] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
- [31] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273-297.
- [32] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). John Wiley & Sons.
- [33] Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1(1), 81-106.

- [34] Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1), 119-139.
- [35] R. A. A. Viadinugroho, “Imbalanced classification in Python: Smote-tomek links method,” Medium, <https://towardsdatascience.com/imbalance-classification-in-python-smote-tomek-links-method-6e48dfe69bbc> (accessed Mar. 28, 2024).
- [36] GfG (2023) Chi-square test in machine learning, GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/ml-chi-square-test-for-feature-selection/> (Accessed: 03 April 2024).
- [37] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99. <https://doi.org/10.1016/j.inffus.2023.101805>
- [38] Saarela, M., & Jauhainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, 3(2). <https://doi.org/10.1007/s42452-021-04148-9>
- [39] Alfeo, A. L., Zippo, A. G., Catrambone, V., Cimino, M. G. C. A., Toschi, N., & Valenza, G. (2023). From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks. *Computer Methods and Programs in Biomedicine*, 236. <https://doi.org/10.1016/j.cmpb.2023.107550>
- [40] Dr. T. Mitsa, “A guide to 21 feature importance methods and packages in machine learning,” Medium, <https://towardsdatascience.com/a-guide-to-21-feature-importance-methods-and-packages-in-machine-learning-with-code-85a841f8b319> (accessed Mar. 28, 2024).
- [41] Gabbay, F., Bar-Lev, S., Montano, O., & Hadad, N. (2021). A lime-based explainable machine learning model for predicting the severity level of covid-19 diagnosed patients. *Applied Sciences (Switzerland)*, 11(21). <https://doi.org/10.3390/app112110417>
- [42] Viswan, V., Shaffi, N., Mahmud, M., Subramanian, K., & Hajamohideen, F. (2023). Explainable Artificial Intelligence in Alzheimer’s Disease Classification: A Systematic Review. In *Cognitive Computation*. Springer. <https://doi.org/10.1007/s12559-023-10192-x>
- [43] Ahuja, Abhimanyu S. (2019) ‘The impact of Artificial Intelligence in medicine on the future role of the physician’, *PeerJ*, 7. doi:10.7717/peerj.7702.
- [44] Javaid, M. et al. (2022) ‘Significance of machine learning in Healthcare: Features, pillars and applications’, *International Journal of Intelligent Networks*, 3, pp. 58–73. doi:10.1016/j.ijin.2022.05.002.
- [45] Idris, A. et al. (2021) ‘Prevalence, incidence, and risk factors of epilepsy in Arab countries: A systematic review’, *Seizure*, 92, pp. 40–50. doi:10.1016/j.seizure.2021.07.031.
- [46] Epilepsy (no date) World Health Organization. Available at: <https://www.who.int/news-room/fact-sheets/detail/epilepsy> (Accessed: 13 April 2024) .
- [47] Blocka, K. (2021) EEG (electroencephalogram): Purpose, procedure, and risks, Healthline. Available at: <https://www.healthline.com/health/eeg> (Accessed: 12 April 2024).
- [48] Heart disease: Causes of a heart attack (2019) eMedicineHealth. Available at: [https://www.emedicinehealth.com/slideshow\\_pictures\\_visual\\_guide\\_to\\_heart\\_disease/article\\_em.htm](https://www.emedicinehealth.com/slideshow_pictures_visual_guide_to_heart_disease/article_em.htm) (Accessed: 12 April 2024).

## Appendix E.2: Epileptic Seizure Conference Paper

### Machine Learning-Based Models for the Pre-emptive Diagnosis of epileptic seizure using Clinical Data

Sunday O. Olatunji <sup>1\*</sup>, Mohammad Aftab Alam Khan<sup>1\*</sup>, Fai Alanazi<sup>1\*</sup>, Rahaf yaanallah<sup>1\*</sup>, Shahad alghamdi<sup>1\*</sup>, Razan Alshammari<sup>1\*</sup>, Fatimah Alkhatim<sup>1\*</sup>, Mehwash Farooqui <sup>1\*</sup>and Mohammed Imran Basheer Ahmed <sup>1\*</sup>

<sup>1</sup> College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia.

\* Correspondence: <sup>1</sup>[osunday@iau.edu.sa](mailto:sunday@iau.edu.sa) <sup>2</sup>[mkhan@iau.edu.sa](mailto:mkhan@iau.edu.sa)

<sup>3</sup>[220000931@iau.edu.sa](mailto:220000931@iau.edu.sa) <sup>4</sup>[2200003935@iau.edu.sa](mailto:2200003935@iau.edu.sa)

<sup>5</sup>[2200003434@iau.edu.sa](mailto:2200003434@iau.edu.sa) <sup>6</sup>[2200004035@iau.edu.sa](mailto:2200004035@iau.edu.sa)

<sup>7</sup>[2200001977@iau.edu.sa](mailto:2200001977@iau.edu.sa) <sup>8</sup>[mfarooqui@iau.edu.sa](mailto:mfarooqui@iau.edu.sa) <sup>9</sup>[mbahmed@iau.edu.sa](mailto:mbahmed@iau.edu.sa)

**Abstract.** Chronic diseases are a major burden on the world's healthcare systems since they require ongoing care and harm the quality of life of those who are afflicted. Among these, epilepsy is most notable as a common neurological condition with repeated seizures that affects more than 50 million individuals globally. Epilepsy diagnosis can be challenging, especially in healthcare settings with limited resources. Nevertheless, early detection of epileptic seizures is essential for optimal management. The paper investigates the use of machine learning algorithms for the preemptive diagnosis of epileptic seizures by utilizing diagnostic, clinical, and demographic data. To distinguish between seizures that are epileptic and those that are not, three machine learning methods are used: Gradient Boosting, Support Vector Machine, and Logistic Regression. The dataset from patients with epileptic seizures (ES) or patients with non-epileptic seizures (NES) was gathered between January 1, 2017, and May 15, 2019. This study optimizes the model using 10-fold GridSearchCV. The use of SelectKBest to decrease the number of features and improve predicted accuracy is a key component of the process. The results showed that Logistic Regression outpaced other methodologies, using 36 features, where it achieved an accuracy of 87.67% with precision, recall, and F1 values all of 88%.

**Keywords:** Pre-emptive Diagnosis; Chronic diseases; Epilepsy; epileptic seizures; Machine Learning Algorithms; Early Detection.

### 1. Introduction

Chronic diseases are a significant burden on global health, particularly impacting populations in economically poor areas. These conditions are responsible for over 75% of the 31.4 million deaths worldwide [1], necessitating continuous medical care and significantly reducing individuals' quality of life and productivity. Among these conditions, epilepsy stands out as a major chronic neurological disorder characterized by frequent seizures, affecting roughly 50 million people globally [2]. The manifestations of epilepsy vary significantly among patients; seizures can range from complete unconsciousness to subtle sensory disturbances, categorized into partial seizures affecting specific brain areas and generalized seizures involving the entire brain [3].

The diagnosis of epilepsy is crucial for effective management and prevention of severe complications, including sudden death. The standard diagnostic tool, the Electroencephalogram (EEG), monitors brain activity to detect abnormal patterns indicative of epilepsy [4]. However, in resource-limited settings, access to EEG is often challenging, underscoring the need for alternative diagnostic approaches. Recent advancements in machine learning, a branch of artificial intelligence, have introduced promising methods that utilize clinical data to enhance early detection and treatment of diseases. This technology has revolutionized healthcare, providing tools that can learn from vast amounts of data to aid in early diagnosis and tailored treatment plans [5].

Non-epileptic seizures can stem from physiological or psychological factors, not necessarily abnormal brain activity [6]. Distinguishing between epileptic and non-epileptic seizures is crucial for accurate diagnosis and optimal patient care, though it can be challenging. Machine learning algorithms offer promise in accurately discriminating between the two types of seizures by analyzing diverse patient data, potentially reducing reliance on costly diagnostic techniques like EEG.

This paper investigates the application of machine learning techniques for preemptive diagnosis of epileptic seizures. It specifically compares the effectiveness of three machine learning models: Gradient Boosting, Support Vector Machine, and Logistic Regression. These models were tested using a comprehensive dataset of patient data collected from the Functional Neurological Disorders Clinic and the University of Colorado Epilepsy Monitoring Unit between January 1, 2017, and May 15, 2019 [7]. Patients in this study, diagnosed with either epileptic or non-epileptic seizures, were assessed using video-electroencephalography (vEEG), which combines video recording with EEG to precisely monitor seizure activity. The comparative analysis revealed that Logistic Regression was particularly effective, utilizing 36 distinct features to achieve an accuracy of 87.67%, with precision, recall, and F1 scores all at 88%.

The remaining segment of this paper is structured as follows. Section 2 is focused on the exploration of literature relevant to the topic. Section 3 contains materials and methods demonstrating dataset, statistical analysis, data preprocessing, utilized ML algorithms, and optimization strategies. Section 4 involves the empirical results. Finally, the conclusion and future work are in section 5.

## 2. Literature Review

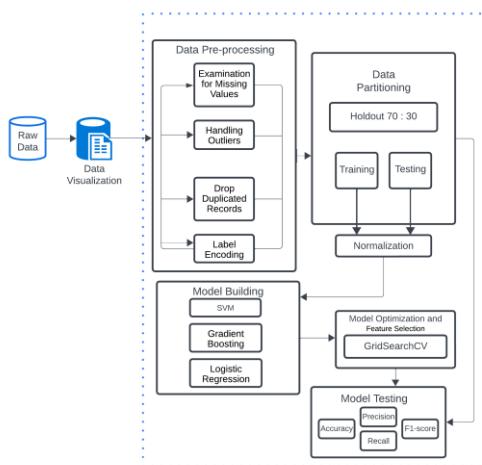
Chen et al. [8] introduced an automated method that combines feature fusion and selection approaches to classify EEG signals. By extracting mixed features such as Fuzzy Entropy (FuzzyEn), Sample Entropy (SampEn), Approximate Entropy (ApEn), and Standard Deviation (STD) from subbands obtained through Discrete Wavelet Transform (DWT) decomposition, their model achieved high accuracy in classifying epileptic EEG signals using Convolutional Neural Networks (CNN). Tawfik et al. [9] investigated the performance of deep learning and ML techniques for epileptic seizure detection, comparing the effectiveness of different algorithms. They found that CNN, especially when combined with Long Short-Term Memory (LSTM) networks, yielded promising results, achieving high accuracy, specificity, and sensitivity on the EEG dataset from Bonn University. Alqaseer et al. [10] emphasized the critical need for accurate and timely seizure detection, proposing an algorithm based on Discrete Cosine

Transformation (DCT) and K-Nearest Neighbor (KNN) classification. Their method, applied to EEG signals segmented into 1-second frames, demonstrated impressive performance in distinguishing between ictal (seizure) and interictal (non-seizure) states, with a high F1-score and low False-Positive Rate (FPR). Usman et al. [11] introduced a deep learning technique combining CNN for feature extraction and Support Vector Machines (SVM) for classification to predict epileptic seizures. Their model, trained on scalp EEG data, achieved notable sensitivity and specificity, highlighting the potential of such approaches in clinical settings.

However, despite these advancements, previous research predominantly focused on EEG imaging datasets, overlooking the valuable insights provided by clinical data. Moreover, limitations such as small sample sizes and inadequate clinical data hindered the generalizability and reliability of the models.

### 3. Materials and Methods

This study developed a Python-based model for early diagnosis of epilepsy by detecting epileptic seizures. The process as shown in figure 1, involved importing and analyzing the dataset using Pandas, handling duplicates and missing values, and visualizing data distribution. Outliers were identified and mitigated through median imputation, while categorical features were imputed and standardized. Label encoding converted categorical features into numerical representations. The dataset was split into training and testing sets, standardized, and subjected to three machine learning methods—Gradient Boosting, Support Vector Machine, and Logistic Regression—tuned via GridSearchCV for optimal performance. The model evaluation included accuracy scores, confusion matrices, and classification reports on both training and testing datasets.



**Fig 164.** The proposed framework for the pre-emptive diagnosis of Epileptic Seizures.

### 3.1 Data Description

The dataset comprised individuals sourced from a continuous stream of referrals to both the Functional Neurological Disorders Clinic and the University of Colorado (CU) Epilepsy Monitoring Unit (EMU) during the period spanning January 1, 2017, to May 15, 2019 [7]. These patients were mostly identified by a confirmed video-electroencephalography (vEEG) diagnosis of epilepsy or a history of seizures (DS) which is referred to as dissociative seizures or non-epileptic seizures, with vEEG evaluations performed both inside and outside the medical system. Patients were randomly chosen to form two groups, comprising individuals with dissociative seizures (DS) and epileptic seizures (ES). Table 1 lists all the features and their corresponding types. For research purposes, this dataset offers a thorough overview of the health status of the patients, enabling in-depth examination and analysis.

**Table 179.** Features' description.

Feature	Type
RRID	int64
Diagnosis	object
Sex	object
#nonpsych comorbidities	int64
# Prior AEDs	int64
Asthma	int64
Migraine	int64
Chronic Pain	int64
Diabetes	int64
non metastatic cancer	int64
Number of non-seizure nonpsych medication	int64
# Current AEDs	int64
Date Onset	object
Date Admit	object
Baseline (sz freq)	float64
Median duration of seizures	float64
# of seizure types	Int64
Injury with seizure	object
Catamenial	object
Trigger of sleep deprivation	object
Aura	object
Ictal Eye Closure	object
Ictal hallucinations	object
Oral automatisms	object
Incontinence	object
Limb automatisms	Object
Ictal tonic-clonic	object
Muscle twitching	Object
Hip thrusting	Object
Post-ictal fatigue	Object
Any head injury	Object
Psych traumatic events	Object
Concussion w/o LOC	Object

Concussion w/LOC	Object
Severe TBI (LOC>30min)	Object
Opioids? Yes/no	Int64
Sex abuse	Object
Physical Abuse	Object
Rape	Object

### 3.2 Statistical Analysis

This section illustrates the characteristics of each attribute in the dataset. This step is essential to determine the most suitable pre-processing techniques for optimal modeling methods. Both categorical and numerical attributes will be explored as both hold a significant value for forthcoming modeling stages. Table 2 represents a detailed statistic for numerical features.

**Table 180** Statistical analysis of numerical features

Feature	Mean	Standard deviation	Min	25 <sup>th</sup> quartile	50 <sup>th</sup> quartile	75 <sup>th</sup> quartile	Max	Missing value counts
Number of nonpsych comorbidities	4.58	4.98	0.00	1.00	3.00	6.00	31.00	0
Number of Prior AEDs	1.71	2.32	0.00	0.00	1.00	3.00	14.00	0
Number of non-seizure nonpsych medication	5.41	6.20	0.00	1.00	3.00	8.00	44.00	0
Number of Current AEDs	1.77	1.34	0.00	1.00	2.00	3.00	6.00	0
Baseline (sz freq)	24.95	41.16	0.00	3.00	10.00	30.00	308.00	5
Median duration of seizures	5.17	10.50	0.05	1.00	2.00	5.00	65.00	12
Number of seizure types	2.19	1.35	1.00	1.00	2.00	3.00	10.00	0

The dataset comprises patients with various medical conditions and symptoms, providing insights into the distribution of these attributes among the patients. Among the target class "diagnosis," 134 patients have non-epileptic seizures, 75 have epileptic seizures, and 31 have been diagnosed with both. In terms of gender, there are 160

female patients and 80 male patients. Additionally, 45 patients have asthma, while 195 are unaffected. Migraine pain is experienced by 73 patients, and 167 do not suffer from it. Chronic pain affects 37 patients, while 203 do not experience it. Other medical conditions include diabetes (14 patients), non-metastatic cancer (19 patients), and sleep deprivation triggered by seizures (25 patients). Various seizure-related symptoms and experiences are also recorded, such as feeling an aura before a seizure (143 patients), experiencing hallucinations during seizures (23 patients), and having muscle twitching (66 patients). Furthermore, post-seizure effects such as fatigue (89 patients) and head injuries (106 patients) are noted. The dataset also includes information on psychological trauma events, opioid use, and instances of abuse among the patients.

### 3.3 Data Pre-processing

The preprocessing phase is crucial for optimizing model performance by addressing issues such as missing values, outliers, and inconsistencies in the raw data. Initially, the dataset comprised 39 features and 241 instances, but after dropping date and ID columns, as well as duplicates and rows with excessive null values, it was refined to 36 features and 240 instances. Handling missing data is essential, and in this study, numerical missing values were replaced with the median due to skewed distributions, while categorical missing values were imputed with the most frequent value in each column. To address the class imbalance, the SMOTE-Tomek method was employed, combining SMOTE for generating synthetic minority class instances and Tomek links for removing near-borderline instances. This approach helps enhance model accuracy by balancing class distribution [12]. Lastly, StandardScaler was applied to standardize features by centering them around zero mean and scaling to unit variance.

### 3.4 Description of Utilized Machine Learning Algorithms

**Gradient Boosting:** Gradient Boosting is an ensemble learning technique that builds decision trees sequentially, each correcting the errors of its predecessor. It minimizes a loss function by greedily adding weak learners. It's widely used in both regression and classification problems due to its high predictive accuracy [13].

**Support Vector Machine (SVM):** SVM is a supervised learning algorithm that analyzes data and recognizes patterns, used for classification and regression analysis. It works by finding the hyperplane that best separates different classes in the feature space. SVM is effective in high-dimensional spaces and in cases where the number of dimensions exceeds the number of samples [14].

**Logistic Regression:** Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. It's primarily used for binary classification problems and estimates the probability that a given input belongs to a particular class. Despite its name, it's a classification algorithm rather than a regression algorithm [15].

### 3.5 Optimization strategy

Selecting the best hyperparameter for each algorithm has a significant impact on the algorithm's performance. Proper hyperparameter leads to the best possible performance and accuracy. To achieve this goal GridSearchCV method was utilized, this method works by giving a set of various values. The method will try and combine all the values until it finds the best combination. Tables 3 and 4 show the algorithms with their best hyperparameter values on both original dataset and oversampled dataset.

**Table 3.** The best hyperparameter was selected on the original data and all features.

algorithm	Hyperparameter	Best Hyperparameter
Gboost	Learning_rate	0.01
	Max_depth	5
	N_estimator	100
SVM	C	1
	Gamma	Scale
	kernel	Rbf
Logistic Regression	C	0.1
	Solver	Liblinear

**Table 4.** The best hyperparameter was selected on the oversampled data with all the features.

algorithm	Hyperparameter	Best Hyperparameter
Gboost	Learning_rate	0.1
	Max_depth	5
	N_estimator	100
SVM	C	10
	Gamma	Scale
	kernel	Rbf
Logistic Regression	C	0.1
	Solver	Liblinear

## 4. Empirical Results

Within this section, we present the outcomes derived from our developed models after the implementation of GridSearchCV on both the original dataset and the sampled data using SMOTETomek. Following the acquisition of optimal hyperparameters and model training through stratified 10-fold cross-validation and utilizing all 36 features, the following section delineates a comprehensive examination of the results obtained, as presented in Table 5.

**Table 5.** The results of the proposed models before and after sampling were applied.

Classifier	Dataset	Testing Accuracy	Precision	Recall	F1-Score

<b>Logistic Regression</b>	Original	<b>87.67%</b>	<b>88.00%</b>	<b>88.00%</b>	<b>88.00%</b>
	Using SMOTETomek	<b>79.45%</b>	<b>85.00%</b>	<b>79.00%</b>	<b>81.00%</b>
SVM	Original	86.30%	87.00%	86.00%	85.00%
Gradient Boosting	Original	78.08%	78.00%	78.00%	78.00%
	Using SMOTETomek	82.19%	82.00%	82.00%	82.00%
	Using SMOTETomek	72.60%	77.00%	73.00%	75.00%

The results show that, although the SMOTETomek oversampling strategy was applied in an attempt to improve class imbalance, classifier performance was not always improved by this method. SMOTETomek reduced the testing accuracy of all classifiers when compared to the original dataset. This finding indicates that, for this dataset, the original, non-oversampled results are considered more reliable. With the best testing accuracy of 87.67%, a precision of 88.00%, recall of 88.00%, and F1-score of 88.00% among the original results, Logistic Regression was the best option for modeling this dataset without using the SMOTETomek technique. Following that, SVM obtained 86.30% accuracy, 87.00% precision, 86.00% recall, and 85.00% F1-score. This analysis emphasizes how important it is to carefully assess the effects of the technique and choose the best classifier when working with unbalanced datasets.

#### 4.1 Results with Feature Selection

The feature selection utilized is based on the statistical test chi-squared to check if two categorical variables have a significant relationship [16]. In this research, a dataset of 36 attributes that were taken from an ongoing stream of referrals to the FND Clinic and CU EMU was tested with the chi-squared test [7]. Comparing observed and expected frequencies inside a contingency table is the basis of the test. By examining the correlation between the components, the chi-square test addresses feature selection issues. The purpose of this analysis is to ascertain whether the correlation between two sample categorical variables accurately reflects the correlation between them in the population [16]. The following is the chi-squared equation 1, where O is the observed value and E is the expected value.

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

The SelectKBest is the methodology used in this study to select the best K number of features from the chi-squared test result. The chi-squared calculates the relationship between the target variable ‘Diagnosis’ and each feature in the dataset. The result is then sorted, so the SelectKBest takes the highest K features in the selection process. Table 6 shows the result of the ML algorithms with feature selection.

**Table 6.** Result with feature selection.

Feature Subsets	20	All features (36)

<b>Logistic Regression</b>	<b>86.30%</b>	<b>87.67%</b>
SVM	84.93%	86.30%
GBoost	80.82%	76.71%

The confusion matrices offer valuable insights into the performance of various machine learning models in classifying instances into three distinct categories: ES (epileptic seizure), NES (non-epileptic seizures), and a combined category of ES & NES. Logistic Regression in Table 7, exhibits a balanced performance, accurately classifying 19 instances of ES, 42 instances of NES, and 3 instances in the ES & NES category.

**Table 7.** Confusion matrix of Logistic regression algorithm.

		<b>Predicted</b>		
		ES	NES	ES And NES
<b>Actual</b>		ES	19 (TP)	2 (FN)
5.	ES	3 (FP)	42 (TN)	0 (FN)
	NES	1 (FP)	1 (FP)	3 (TN)

## Conclusion

This study focuses on the preemptive diagnosis of epileptic seizures using machine learning algorithms, leveraging diagnostic, clinical, and demographic data. The research aims to address the challenges of early detection and diagnosis of epilepsy, particularly in healthcare settings with limited resources. Three machine learning methods were employed, the findings indicate that for this specific dataset, the original, non-oversampled results are considered more reliable, with Logistic Regression achieving the best testing accuracy of 87.67% with precision, recall, and F1 values all of 88% without using the SMOTETomek technique. The study emphasizes the importance of carefully considering the impact of oversampling techniques and selecting the most appropriate classifier for handling imbalanced datasets.

In future work, the model could be tested in new datasets or with new patients, and the exploration of new classifiers and methods, such as deep learning, could be considered to further improve accuracy. Additionally, the recommendation is to utilize the model in patients by employing Longitudinal Data Analysis, where continuous patient data is collected, to enhance the model's applicability and effectiveness in real-world healthcare settings. This approach would enable the model to adapt and evolve based on ongoing patient data, potentially leading to improved diagnostic accuracy and proactive management of epileptic seizures.

## References

- [1] "Non communicable diseases," World Health Organization, <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases> (accessed Mar. 8, 2024).
- [2] World Health Organization, "Epilepsy," WHO, Feb. 07, 2024. <https://www.who.int/news-room/fact-sheets/detail/epilepsy> (accessed Mar. 1, 2024).

- [3] Klein. E., "Epilepsy: Definition, symptoms, treatment, causes, and more," medicalnewstoday. Nov. 21. 2019  
<https://www.medicalnewstoday.com/articles/8947#symptoms> (accessed Mar. 1, 2024).
- [4] Sirven, J. and Schachter, S., "Electroencephalography (EEG)," Epilepsy Foundation. <https://www.epilepsy.com/diagnosis/eeg> (accessed Mar. 1, 2024).
- [5] An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023). "A Comprehensive Review on Machine Learning in Healthcare Industry...". Sensors (Basel, Switzerland), 23(9), 4178. <https://doi.org/10.3390/s23094178> .
- [6] "non-epileptic seizures and dissociative seizures | Epilepsy Society," epilepsysociety. (2023) <https://epilepsysociety.org.uk/about-epilepsy/what-epilepsy/non-epileptic-seizures> (accessed Mar. 1, 2024).
- [7] Lenio, Steven; Kerr, Wesley; Watson, Meagan; Baker, Sarah; Bush, Chad; Rajic, Alexander; Strom, Laura (2021), "Data & Code from Validation of a predictive calculator to distinguish between patients ...", Mendeley Data, V1, doi: 10.17632/cshccr8w3h.1 .
- [8] Chen, W., Wang, Y., Ren, Y., Jiang, H., Du, G., Zhang, J., & Li, J. (2023). An automated detection of epileptic seizures EEG using CNN classifier based on feature fusion with high accuracy. BMC Medical Informatics and Decision Making, 23(1). <https://doi.org/10.1186/s12911-023-02180-w>
- [9] M. Tawfik, E. Mahyoub, Z. A. Ahmed, N. M. Al-Zidi, and S. Nimbhore, "Classification of epileptic seizure using machine learning and deep learning based on electroencephalography (EEG)," Communication and Intelligent Systems, pp. 179–199, 2022. doi:10.1007/978-981-19-2130-8\_15
- [10] M. A. Jumaah, A.I.Shihab and A.A.Farhan, "Epileptic seizures detection using DCT-II and KNN classifier in long -Term EEG Signals"Researchgate, Feb. 2020, doi: [https://www.researchgate.net/publication/339302243\\_Epileptic\\_Seizures\\_Detection\\_Using\\_DCT-II\\_and\\_KNN\\_Classifier\\_in\\_Long-Term\\_EEG\\_Signals](https://www.researchgate.net/publication/339302243_Epileptic_Seizures_Detection_Using_DCT-II_and_KNN_Classifier_in_Long-Term_EEG_Signals)
- [11] S. Muhammad Usman, S. Khalid, and M. H. Aslam, "Epileptic seizures prediction using Deep Learning Techniques," IEEE Access, vol. 8, pp. 39998–40007, 2020. doi:10.1109/access.2020.2976866
- [12] Hairani, H., Anggrawan, A., & Priyanto, D. (n.d.). INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION journal homepage : [www.joiv.org/index.php/joiv](http://www.joiv.org/index.php/joiv) INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link. [www.joiv.org/index.php/joiv](http://www.joiv.org/index.php/joiv)
- [13] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189-1232.
- [14] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273-297.
- [15] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). John Wiley & Sons.
- [16] GfG (2023) Chi-square test in machine learning, GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/ml-chi-square-test-for-feature-selection/> (Accessed: 03 April 2024).

# **Comprehensible Machine Learning-Based Models for the Pre-emptive Diagnosis of Sickle Cell Anemia using Clinical Data**

**Sunday O. Olatunji<sup>1</sup>, Mohammad Aftab Alam Khan<sup>1</sup>, Fai Alanazi<sup>1</sup>, Rahaf yaan allah<sup>1</sup>, Shahad Alghamdi<sup>1</sup>, Razan Alshammari<sup>1\*</sup>, Fatimah Alkhathim<sup>1</sup>, Mehwash Farooqui<sup>1</sup>and Mohammed Imran Basheer Ahmed<sup>1</sup>**

<sup>1</sup> College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia.

Email: <sup>1</sup> [osunday@iau.edu.sa](mailto:osunday@iau.edu.sa) <sup>2</sup> [mkhan@iau.edu.sa](mailto:mkhan@iau.edu.sa) <sup>3</sup> [220000931@iau.edu.sa](mailto:220000931@iau.edu.sa) <sup>4</sup> [2200003935@iau.edu.sa](mailto:2200003935@iau.edu.sa) <sup>5</sup> [2200003434@iau.edu.sa](mailto:2200003434@iau.edu.sa)  
<sup>6</sup> [2200004035@iau.edu.sa](mailto:2200004035@iau.edu.sa) <sup>7</sup> [2200001977@iau.edu.sa](mailto:2200001977@iau.edu.sa) <sup>8</sup> [mfarooqui@iau.edu.sa](mailto:mfarooqui@iau.edu.sa) <sup>9</sup> [mbahmed@iau.edu.sa](mailto:mbahmed@iau.edu.sa)

\* Correspondence: [2200004035@iau.edu.sa](mailto:2200004035@iau.edu.sa)

## **Abstract**

Sickle cell anemia (SCA) is a chronic genetic condition that results in sickle-shaped red blood cells due to aberrant hemoglobin production. This condition results in impaired oxygen transport, vessel occlusion, and various complications, severely impacting patients' health and quality of life. In Saudi Arabia, SCA presents a significant public health challenge, particularly in the Eastern Province, where prevalence rates are notably high. Early diagnosis and intervention are crucial for effective management and improved outcomes. Leveraging machine learning (ML) algorithms alongside traditional diagnostic methods offers promising avenues for enhancing SCA diagnosis and prediction. By integrating clinical data from blood samples with ML techniques, this study seeks to develop accessible and interpretable predictive models for preemptive SCA diagnosis. Drawing on a dataset collected from a pediatric hospital-based study in Sudan, encompassing socio-demographic variables and clinical parameters, this research aims to overcome prior limitations in ML-based SCA diagnosis, particularly the scarcity of studies incorporating clinical data. Through the application of various ML algorithms, including AdaBoost, SVM, KNN, XGBoost, LogisticRegression, GradientBoosting, DecisionTree, RandomForest, GaussianNB, and SVM\_Gaussian, this study endeavors to facilitate targeted interventions and personalized treatment plans, ultimately mitigating the burden of SCA in affected populations. Results indicate significant improvements in diagnostic accuracy with Random Forest demonstrating the highest accuracy at 86.77% using only 5 features, alongside 89% precision, 87% recall, and 83% f1-score. To gain a deeper understanding of the top-performing AI model, the study also employed Explainable Artificial Intelligence (XAI) approaches such as Local Interpretable Model-Agnostic Explanations (LIME) and Feature importance.

**Keywords:** Pre-emptive Diagnosis; Chronic diseases; Sickle cell anemia; Explainable Artificial Intelligence; Machine Learning Algorithms; Early diagnosis.

## **1. Introduction**

Sickle cell anemia is one of the chronic genetic conditions that affects red blood cells, as their shape changes from spherical to a crescent or sickle shape, making them fragile, weak, and breaking easily. This change results in the inability of red cells to effectively transport oxygen to the body's tissues and causes blockage of small blood vessels as a result of red cells sticking together, leading to symptoms such as chronic pain attacks, shortness of breath, and pale skin [1]. Sickle cell anemia (SCA) is one of the most common

genetic diseases and one of the most difficult health challenges in the world. Estimates show that the number of new cases of this disease ranges from 30,000 to 40,000 newborns annually [2]. Sickle cell anemia, a recessive genetic disease, manifests when an individual carries both genes responsible for the condition. However, in cases where only one gene is present, the individual is identified as a carrier of the trait, known as the Sickle cell trait, without exhibiting symptoms. Most carriers of the sickle cell trait remain asymptomatic throughout their lives but possess the potential to pass the affected gene on to future generations, thereby perpetuating the inheritance of the disease within families and communities [3].

Sickle cell disease, a potentially life-threatening condition, arises from abnormalities in hemoglobin, the key molecule responsible for transporting oxygen in red blood cells. Normally, hemoglobin consists of four protein chains: two alpha globins ( $\alpha$ -globin) and two beta globins ( $\beta$ -globin), forming what's known as hemoglobin A (HbA). The beta globin gene, or HBB gene, encodes the beta subunit of hemoglobin. Various mutations in the HBB gene lead to this disorder. Each person inherits two copies of the HBB gene, and when both copies carry mutations, the production of normal beta globin is hindered, leading to the disease. Sometimes, each copy can undergo different mutations, resulting in distinct types of abnormal beta subunits within the same individual. Sickle cell anemia (HbSS), the most severe form of this illness, occurs when both copies carry the same mutation, resulting in the production of mutant hemoglobin S. Different combinations of mutations give rise to different forms of sickle cell disease, with each copy inherited from a parent. Hemoglobin S has a tendency to form polymers in low-oxygen conditions, a process called sickling or gelation, which thickens the cell's interior, giving it a crescent shape as polymer filaments entangle the cell membrane [4]. Figure 1 shows some combinations of HBB mutations.

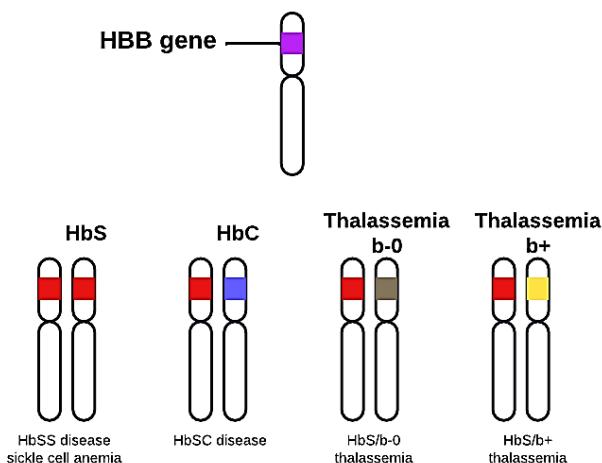


Figure 165 Combinations of HBB Mutations

The severe nature of sickle cell anemia and its associated complications pose significant risks to individuals' health and quality of life. Among these complications, stroke becomes a serious issue, wherein the abnormal shape of sickle cells can block blood vessels, depriving the brain of oxygen and causing long-term neurological damage. Acute chest syndrome, which is typified by fever, respiratory distress, and chest tightness, is another challenging issue. It is often brought on by sickle cell disease-related blockages in the pulmonary vasculature. Additionally, pulmonary hypertension, resulting from the constriction of lung blood vessels by sickle cells, places strain on the heart and may progress to heart failure if left untreated. Furthermore, Due to the shortened lifespan of sickle cells, the spleen compensates by working overtime to address this deficiency. But since the spleen is an essential organ for defending the body against infections, this increased strain on its function results in damage and makes patients more vulnerable to inflammatory illnesses. Moreover, the reduced flexibility of sickle cells in anemia leads to blockages in blood vessels, hindering the supply of blood to the body's tissues. These blockages cause severe pain and, in severe cases, may result in infarctions, leading to the loss of affected organs' functional abilities [5]. In order to maximize

the health outcomes for those with sickle cell anemia, comprehensive care techniques focused on avoiding, identifying, and managing these complex consequences are necessary.

In Saudi Arabia, Sickle Cell Anemia poses a significant burden on public health, with approximately 4.2% of the population carrying the sickle cell trait and around 0.26% of them affected by the disease [6]. Its effects are most noticeable in the Eastern Province, where the trait is most prevalent at 17% of the population; 1.2% of those afflicted with the illness. These statistics from the Ministry of Health highlight the urgency of addressing Sickle Cell Anemia within the country. Early diagnosis and intervention are paramount in mitigating the effects of the disease and improving patient outcomes. Timely identification enables healthcare providers to implement appropriate management strategies, including preventive measures and therapeutic interventions, thereby reducing the risk of complications associated with Sickle Cell Anemia. By prioritizing early diagnosis and access to comprehensive care, Saudi Arabia can significantly alleviate the burden of Sickle Cell Anemia on affected individuals and their families, enhancing both health outcomes and quality of life. Using machine learning algorithms and clinical data from blood samples, along with traditional diagnostic techniques, could improve the early diagnosis and prediction of sickle cell anemia (SCA) in Saudi Arabia. These parameters include hemoglobin (Hb), mean corpuscular hemoglobin concentration (MCHC), mean corpuscular hemoglobin (MCH), red blood cell count (RBCs), and total white blood cell count (TWBCs). By analyzing large datasets and finding patterns and connections indicative of SCA risk, these sophisticated computational approaches make it possible to discover sickle cell trait carriers or patients early. By utilizing the power of machine learning, healthcare providers can develop predictive models that help stratify patients according to their risk of getting SCA, allowing for targeted interventions and personalized treatment plans [7]. Consequently, several studies have been conducted to diagnose SCA using ML algorithms. Nevertheless, these studies have mostly concentrated on using imaging datasets for SCA diagnosis, which increases the difficulty of obtaining data and adds complexity to the use of the model, especially for stakeholders who are not technical.

Therefore, by creating easily understood machine learning-based models for the proactive diagnosis of sickle cell anemia (SCA) using clinical data, this work seeks to overcome the limitations of earlier studies. The dataset utilized in this study was collected through a descriptive cross-sectional hospital-based study conducted at Al Fashir Teaching Hospital in Sudan from December 2017 to August 2018 [8]. The study focused on children aged 0–18 years who were admitted to the hospital. A structured questionnaire capturing socio-demographic data such as age, gender, and tribe was administered to 400 patients at the pediatric ward during the study period. Various ML algorithms were utilized, including. Through the application of various ML algorithms, including AdaBoost, SVM, KNN, XGBoost, LogisticRegression, GradientBoosting, DecisionTree, RandomForest, GaussianNB, and SVM\_Gaussian, the study endeavors to facilitate targeted interventions and personalized treatment plans, ultimately mitigating the burden of SCA in affected populations. Results indicate significant improvements in diagnostic accuracy, particularly after employing SMOTETomek, with Random Forest demonstrating the highest accuracy at 86.77% using only 5 features, alongside 89% precision, 87% recall, and 83% f1-score.

## 2. Literature Review

The research paper that was produced by Darrin et al. [9] aimed to explore the potential of convolutional and recurrent neural networks in classifying red cell dynamics, specifically in the context of sickle cell disease. To achieve this, the researchers utilized a carefully collected dataset of red cell images, capturing different morphological variations associated with the disease. The methodology involves extracting features from sequential red cell images using CNNs and then leveraging RNNs to capture temporal dependencies within the data. The researchers trained their model on a dataset consisting of red cell images from healthy individuals and individuals with sickle cell disease. The results demonstrated the effectiveness of their approach, with the

model achieving high accuracy in classifying red blood cells between the two groups with an accuracy of 97% and an F1-score of 0.94.

A study by Ayoade et al. [10] aimed to enhance sickle cell disease (SCD) prediction accuracy by comparing individual Machine Learning (ML) algorithms and their ensemble models, targeting elongated erythrocyte shape identification. Moreover, three ML algorithms—MLR, XGBoost, and RF—were evaluated, and ensemble models were created. Performance metrics including accuracy, sensitivity (ROC-AUC), and F1 score were employed to assess the models' efficacy. Additionally, the analysis was conducted using Python programming, and medical datasets were employed. While individual algorithms achieved good accuracies (MLR = 87%, XGBoost = 90%, RF = 93%), the hybrid models RF-MLR and RF-XGBoost outperformed with accuracies of 92% and an impressive 99%, respectively.

The research study achieved by Nguyen et al. [11] was aimed at investigating the association of interleukin-6 (IL6) and interleukin-8 (IL8) with Sickle Cell Anemia (SCA) patients and exploring the possibility of predicting their presence using ANN based on hemoglobin alleles and other hematological variables. The dataset for this study was collected using a cross-sectional study that involved 74 healthy individuals and 60 sickle anemia patients. The researchers utilized a deep learning model to analyze the data, train the models using supervised learning techniques, and evaluate their performance using various evaluation metrics. They discovered a non-linear association between hemoglobin alleles and IL6 and IL8 production in SCA patients. The model achieved an impressive accuracy of 90.9% and an r-squared value of 0.88, demonstrating its potential to aid in the development of specific treatments and diagnostics for those suffering from Sickle Cell Anemia and associated immune complications.

A paper written by Saputra et al. [12], the authors propose the development of an automated prediction model to assist doctors in distinguishing between four types of anemia. This is essential because diagnosing anemia is challenging due to its wide range of symptoms and diverse forms. This study involved 165 females and 25 males aged 15 to 41 who had been diagnosed with various types of anemia. This study model was created by the ELM algorithm. Afterward, its performance was assessed using a confusion matrix with a dataset of 190 samples to represent the four types. The model achieved high results with an accuracy rate of 99.21%, sensitivity of 98.44%, precision of 99.30%, and an F1 score of 98.84%.

A paper by Vohra et al. [13], the authors considered the diagnosis problem as a classification task with three classes: mild, moderate, and severe, in contrast to previous studies, which represented it as a binary classification problem. Patient data from the Eureka diagnostic center in Lucknow, India, available on the Mendeley Data Repository, was used with 400 samples. They applied six ML algorithms for multi-class classification on the dataset, benchmarking their performance using both 10-fold cross-validation and hold-out methods. The results of the MLP network model showed the highest performance, achieving an accuracy of 99.35% on the SMOTE dataset.

Patgiri and Ganguly [14] conducted a comprehensive exploration of automatic disease diagnosis through image processing techniques, with a specific focus on sickle cell anemia (SCA) detection from microscopic blood images. In this study, the authors proposed an innovative approach for SCA detection, employing a segmentation method that combines local adaptive thresholding and an active contour-based algorithm. Furthermore, supervised classifiers, including SVM and ANN, were utilized to identify sickle cells based on the geometric features of RBCs. Their findings revealed that the SVM classifier outperformed the ANN, achieving an impressive accuracy rate of 99.2%, while the ANN, trained with resilient backpropagation and featuring ten hidden neurons, demonstrated promising accuracy at 99%. Additionally, the review also delved

into various segmentation methods used in medical image analysis, encompassing thresholding, edge-based, region-based, and clustering-based techniques.

A paper written by Shahzad et al. [15], the authors aimed to predict the severity of anemia by extracting essential morphological features to identify anemic pictures using a 3-tier deep convolutional fused network (3-TierDCFNet). The dataset includes 11,500 pictures with about 750,000 red blood cell elements. Out of these pictures, 5,750 are considered normal, and the other 5,750 show signs of anemia. The model consists of 2 modules: Module-I classifies pictures as Anemic or healthy, and Module-II detects Mild or Chronic anemia. Validation reduces inappropriate feature selection after each module's training. Evaluation metrics include specificity, recall, F1-Score, and accuracy. The Tier-III model achieved the highest result with 89.29% accuracy, 95.96% recall, 95.87% F1-Score, and 96.34% specificity.

This study was conducted by Srivastave et al. [16] and aimed to diagnose sickle cell anemia using the AutoML approach on UV-VIS absorbance spectroscopy data. The researchers utilized a dataset consisting of UV-VIS absorbance spectra collected from blood samples of individuals with and without sickle cell anemia. The dataset was collected from a diverse set of patients representing different age groups and ethnicities. The methodology involved the use of AutoML, a machine learning technique that automates the process of model selection and hyperparameter tuning. The dataset was divided into 70% for training and 30% for testing. The researchers trained various machine learning models on the dataset and evaluated their performance using cross-validation. This approach accurately identifies the presence of sickle hemoglobin with a sensitivity of 100% and a specificity of 93.84%. This study demonstrates the potential of AutoML in the field of medical diagnostics, offering an effective and efficient approach for diagnosing sickle cell anemia based on UV-VIS absorbance spectroscopy data.

A study was done by Yeruva et al. [17]. The paper introduced a deep neural network used to identify red blood cells and classify them into three categories: normal, sickle cell, and thalassemia. To enhance the accuracy of distinguishing sickle cells from other types of cells, the authors used an approach called Multi-Layer Perceptron (MLP) that provides more efficiency than the rest of the algorithms that have been used in the study. In addition to other machine learning approaches such as SVM, RF, KNN, Logistic Regression, and DT, the dataset used was received from the Thalassemia and Sickle Cell Society (TSCS), which has 1387 records. The results of the study confirm that using the MLP approach will enhance the accuracy of recognizing sickle cells. MLP has an accuracy of 99%, which is the best, followed by RF with 97%.

A paper written by Abdulkarim, H.A. et al. [18], the focus was to construct a deep-learning AlexNet model for red blood cell classification in sickle cell anemia (SCA) patients. The researchers used a dataset of over 9,000 single red blood cell (RBC) images from 130 SCA patients, with 750 cells in each class. Furthermore, the algorithm was developed in two stages: the automation of RBC extraction to identify the region of interest (ROI) in the blood smear image and the use of a deep-learning AlexNet model for classifying and predicting the presence of abnormalities in SCA patients. Finally, the results demonstrate that the proposed framework achieved a classification prediction accuracy of 95.92% for identifying abnormalities in the RBCs of SCA patients using a deep-learning AlexNet model.

Research conducted by Alzubaidi et al. [19] aims to develop a classification model for red blood cells to distinguish sickle cells from other kinds of cells. Besides, the authors wanted to highlight the lack of red blood cells in people infected with sickle anemia. The model used lightweight deep learning with transfer learning techniques to resolve the problem of lacking data. In addition to data augmentation to expand the quantity of data to be trained, the model was trained and tested with three different datasets and different scenarios. The

final results confirm that the model has an accuracy of 99.54%. Additionally, adding an SVM classifier to the model enhanced the performance slightly, achieving an accuracy of 99.98%.

A study by Mohammed et al. [20], the researchers investigated the potential of ML techniques in predicting the appearance of organ failure in adult SCD patients admitted to intensive care units (ICUs). A dataset comprising continuous physiological data was collected from 134 adult subjects at Methodist Le Bonheur Hospital in Memphis, Tennessee. Moreover, four machine learning algorithms were used to build classification models: multi-layer perceptron (MLP), RF, LoR, and SVM. Finally, the RF model accurately predicted organ failure up to six hours before its onset, with an accuracy of 94.57%, sensitivity of 90.24%, and specificity of 98.9%.

The paper by Alzubaidi et al. [21] discussed using a deep convolutional neural network to categorize red blood cells (RBC) into three categories: normal cells, sickle cells, and other cells that have another kind of disease. The proposed model overcomes the problem of classifying RBC using traditional methods. It needs more time and effort because of the complexity of RBC shapes. To train and test the model, 340 images of RBC were used as a dataset for this model, collected from Wadsworth Center data. With the help of error-correcting output codes (ECOC), the model's accuracy is 92.06%. The result shows the effectiveness of this model in classifying the RBC, so less time is needed to diagnose sickle anemia.

A study authored by Xu et al. [22], the authors present a comprehensive overview of their research on sickle cell disease (SCD) and the development of an automated RBC pattern classification framework. The authors' framework is divided into three phases: the automatic RBC extraction method, the RBC patch-size normalization method, and deep convolutional neural network (CNN) classification. Furthermore, the study includes experiments on microscope image datasets from eight SCD individuals from two different hospitals. Additionally, the authors highlight the importance of shape factor quantification and discuss the potential for clinical applications in SCD management.

The purpose of Alkrimi et al. [23] research is to develop an automated method for classifying red blood cells (RBCs) as normal or abnormal using Support Vector Machine (SVM) classification. The authors highlight the significance of medical imaging in the diagnosis of blood disorders as well as the function of RBC shape in clinical diagnosis. Furthermore, image processing techniques such as segmentation and mean filtering are used in this study to extract geometric, texture, and color properties from RBC images using photo imaging microscopy. To distinguish between normal and abnormal RBCs, the SVM is used as the classifier. Moreover, the dataset utilized comprises 1000 images of RBCs obtained from the Department of Hematology at Serdang Hospital in Malaysia. Finally, the experimental findings show that the proposed classifier algorithm achieves high accuracy rates, with an accuracy of 99.9%.

A survey of the literature on sickle cell anemia prediction found that half of the previous studies on the subject used imaging datasets of blood cells. Moreover, prior studies were constrained by small sample sizes and inadequate clinical data. Therefore, the purpose of this study is to close this gap by developing machine learning models that uses clinical data to predict sickle cell anemia more reliably and accurately. Furthermore, the study will employ explainable AI (XAI) approaches to ensure the prediction model's reliability and openness, so ensuring that healthcare professionals may successfully apply it in a variety of settings.

### 3. Materials and Methods

This study aimed to develop a preemptive model for diagnosing sickle cell anemia utilizing the Python programming language. The materials and methods employed in this research encompassed a series of procedural steps designed to effectively handle the dataset and implement machine learning algorithms. Initially, the dataset was imported into the analysis environment using the Pandas package. Preprocessing steps were then executed to ensure data quality and consistency. Duplicate entries were identified and removed, while missing values were handled through a combination of visualization techniques and imputation strategies. Visualization methods, including heatmaps, histograms, and count plots, were employed to assess the distribution of missing values, wrong values, and identify outliers. Categorical features with missing values were imputed using the most frequent strategy, and data cleaning procedures were conducted to standardize categorical values. Subsequently, categorical features were encoded into numerical representations using label encoding, facilitating compatibility with machine learning algorithms. The dataset was then partitioned into training (70%) and testing (30%) sets to facilitate model evaluation. Feature scaling was performed using standardization to ensure uniformity in feature magnitudes across the dataset. Ten machine learning algorithms, including RandomForest, SVM, KNN, GradientBoosting, XGBoost, LogisticRegression, DecisionTree, AdaBoost, GaussianNB, and SVM\_Gaussian, were implemented for predictive modeling. Hyperparameter tuning was conducted using GridSearchCV to optimize algorithm performance. Evaluation of model performance involved assessing accuracy scores, confusion matrices, and classification reports on both training and testing datasets. Once the top-performing model was identified, eXplainable Artificial Intelligence (XAI) techniques, such as Feature Importance and LIME, were employed. The complete procedural workflow is illustrated in Figure 2.

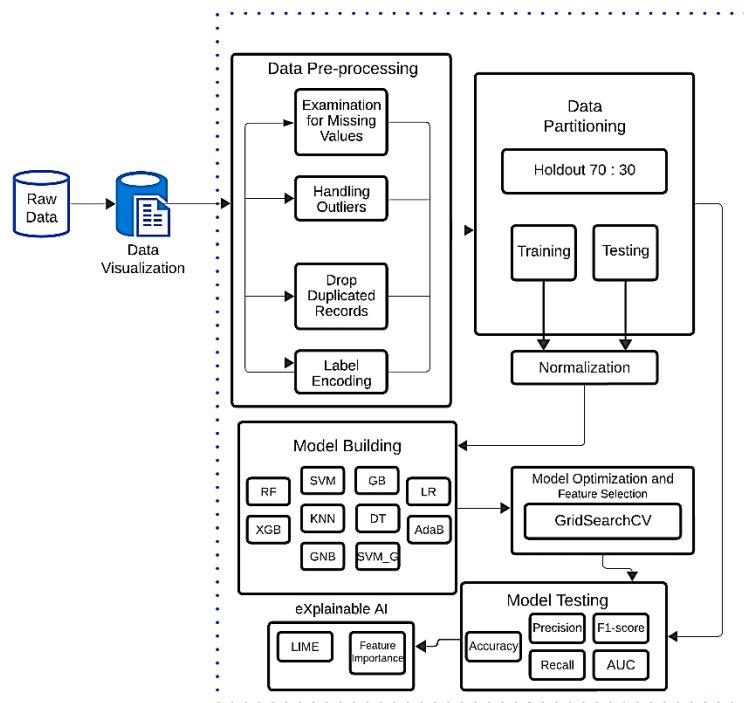


Figure 166 The proposed framework for the pre-emptive diagnosis of SCA

### 3.1. Data Description

The dataset for this study was derived from a descriptive cross-sectional analysis conducted at Al Fashir Teaching Hospital in Sudan, spanning from December 2017 to August 2018 [8]. This study included 400 pediatric patients who were admitted to the hospital during the specified period. These patients, aged between

0 and 18 years, were approached to participate in the study, with their parents providing informed consent. The selection process employed random sampling, ensuring equal opportunity for any child admitted to the pediatric ward to be included.

To gather sociodemographic information, including age, gender, and tribal affiliation, a structured questionnaire was used. 400 parents gave their consent, indicating a moderate percentage of refusal.

Blood samples, amounting to 5 ml each, were collected via peripheral venipuncture under strictly aseptic conditions. The samples were promptly transferred to the laboratory within 5 minutes of collection for further analysis .

Upon arrival at the laboratory, the blood samples underwent immediate testing using advanced diagnostic equipment. Complete blood counts were performed using an automated haematological analyzer, Sysmex Kx 21N, while hemoglobin electrophoresis was conducted utilizing the MINICAP HEMOGLOBIN capillary zone electrophoresis (CZE) system by Sebia, France.

For the haemoglobin electrophoresis procedure, blood samples were initially cooled to 2–8 °C to facilitate sediment formation over several hours. Following centrifugation at 5000 rpm for 5 minutes, the plasma was removed, and the sediment was vortexed briefly. Reagent cups were prepared for analysis, ensuring proper calibration of haemoglobin buffers and waste disposal. Each sample, along with a normal Hb A2 control, was labeled with specific barcodes and placed into haemolysing tubes, which were then positioned within the carousel for automated analysis. Table 1 shows each feature with its corresponding type.

Feature	Type
Hb Electro	object
Platelets	Integer
Total White Blood Cells	float64
Mean Cell Hemoglobin Concentration	float64
Mean Cell Hemoglobin	float64
Mean Cell Volume	float64
Red Blood Cells	float64
Packed Cell Volume	float64
Hemoglobin	float64
Tribe	Integer
Age/ Years	object
Sex	object
No	Integer

Table 181 Features' description

### 3.2. Statistical Analysis

Statistical analysis is an essential process to explore and understand the characteristics of both categorical and numerical data. In addition, determine any further needed preprocessing technique is needed. This will ensure that the model will perform smoothly with the highest possible accuracy.

Numerical attributes were explored through various statistical matrices to ensure the understanding of these attribute and its characteristics. Table 2 shows the statistical matrices for the numerical attribute.

Feature	Mean	Standard deviation	Min	25 <sup>th</sup> quartile	50 <sup>th</sup> quartile	75 <sup>th</sup> quartile	Max	Missing value counts
PLTs	284.59	134.52	21.00	195.25	272.00	351.00	780.00	0

<b>TWBCs</b>	7.95	6.27	1.00	4.60	6.65	9.40	61.70	0
<b>MCHC</b>	32.67	1.80	25.80	31.60	32.80	33.80	40.20	0
<b>MCH</b>	27.06	2.99	15.90	24.52	27.30	28.90	39.50	0
<b>MCV</b>	82.57	8.61	34.00	77.90	83.05	88.00	107.70	0
<b>PCV</b>	35.48	6.77	11.20	31.82	36.20	39.87	53.00	0
<b>RBCs</b>	4.47	1.53	1.14	4.07	4.49	4.83	22.10	0
<b>Hb</b>	11.61	2.23	3.50	10.50	11.90	13.00	17.70	0
<b>Tribe</b>	11.06	13.09	1.00	2.00	4.00	15.75	59.00	0

Table 2 Statistical analysis for numerical attribute

The figure below shows the distribution of the only categories attribute which is the gender. The dataset contains 226 female patient and 176 male patients.

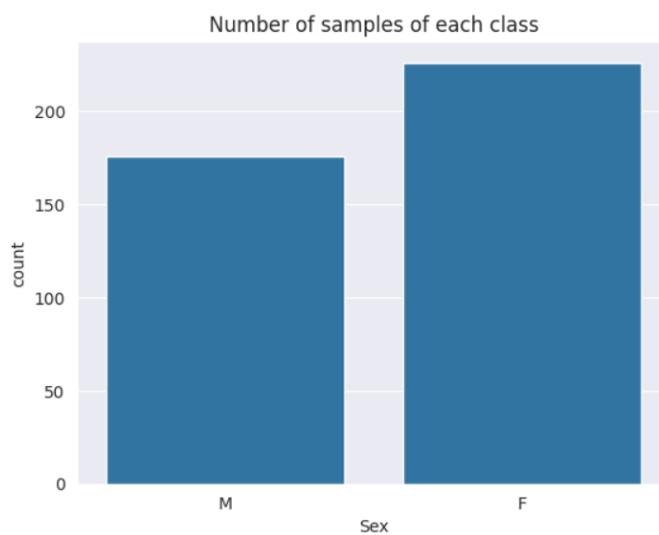


Figure 3 the distribution of gender attribute

### 3.3 Data Pre-processing

Raw data usually comes along with some issues that prevent the model from performing well. The data might contain noise, outliers or missing values. Before start constructing the model these issues must be deal with in order to get the optimal performance and highest possible accuracy. To get a clean data set, any noise was removed from the dataset, moreover the dataset in this study has no missing nor duplicate values. In the target attribute, the class S (sickle cell trait) and SS (sickle cell trait) were combined as one class. Deleting unnecessary columns such as patients' number and date is essential to enhance accuracy. At the beginning the dataset had 13 features and after deleting unwanted features it ended up with 11 including the target class. Standardizing the data is one of the common requirements for building a machine learning model as it enhances the performance and the coverage speed. To meet this requirement StandardScaler() function was utilized. The function works by eliminating the mean and scaling them to unit variance. To calculate the standard score for a sample data X the function needs the mean and standard deviation. This is shown in equation 1 [ 24].

$$Z = \frac{(X - u)}{s} \quad (1)$$

Lastly, to balance the number of samples in the target class, SMOTE-Tomek method was utilized. It is a method used for oversampling that combine two other methods called SMOTE and Tomeklink. Combining these methods helps to avoid the disadvantages of these techniques. The method works by generating samples

in the monitor class and then removing samples from the majority class. Balancing the data help in improving performance and accuracy of classification [25].

### 3.4 Description of Utilized Machine Learning Algorithms

#### 3.4.1 Support Vector Machine (SVM)

SVM is a powerful classifier that works by finding the hyperplane that best divides a dataset into classes. It is effective in high-dimensional spaces and versatile as different Kernel functions can be specified for the decision function. Robust against overfitting, especially in high-dimensional space, SVM is widely used for pattern recognition [26].

**3.4.2 AdaBoost (Adaptive Boosting):**  
AdaBoost is an ensemble technique that combines multiple weak classifiers to form a strong classifier. By reweighting the training samples, it focuses on the hard-to-classify instances in subsequent models, enhancing classification accuracy. AdaBoost is particularly effective for binary classification tasks and is less susceptible to overfitting compared to other algorithms [27].

**3.4.3 K-Nearest Neighbors (KNN):**  
KNN is a simple, instance-based learning algorithm where the class of a sample is determined by the majority of the classes of its nearest neighbors. It is highly adaptable and easy to implement, making it suitable for solving both classification and regression problems. However, it becomes significantly slower as the size of the data increases [28].

**3.4.4 XGBoost (Extreme Gradient Boosting):**  
XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. It is renowned for its performance and speed in classification tasks. XGBoost provides a scalable and efficient solution, often winning many Kaggle competitions [29].

**3.4.5 Logistic Regression:**  
Logistic Regression is used for binary classification problems, where predictions are mapped to probabilities via the logistic function. It is robust to noise and efficient for linearly separable classes. This model is often used as a baseline for binary classification problems [30].

**3.4.6 Gradient Boosting:**  
Gradient Boosting constructs an additive model in a forward stage-wise fashion, allowing optimization of arbitrary differentiable loss functions. Known for its effectiveness in handling heterogeneous features and variable interactions, it is widely used in both classification and regression tasks [31].

**3.4.7 Decision Tree:**  
Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision Trees are easy to interpret and visualize [32].

**3.4.8 Random Forest:**  
Random Forest is an ensemble of Decision Trees, typically trained via the bagging method. The individual trees operate as weak learners, while their collective decisions, taken by majority voting, provide robust

predictions against overfitting. Random Forest performs well on large datasets with high dimensionality [33].

<b>3.4.9</b>	<b>Gaussian</b>	<b>Naive</b>	<b>Bayes</b>	<b>(GaussianNB):</b>
				GaussianNB applies Bayes' theorem with the assumption of independence among predictors. GaussianNB is particularly suited when features are continuous and normally distributed. It is simple and effective, especially for large datasets [34].

<b>3.4.10</b>	<b>SVM</b>	<b>with</b>	<b>Gaussian</b>	<b>Kernel</b>	<b>(SVM_Gaussian):</b>
					SVM with a Gaussian kernel allows the model to create more complex boundaries around classes. The kernel transforms the data into a higher-dimensional space where a hyperplane can effectively separate classes that are non-linearly separable in the original space. This variant is useful for datasets with complex patterns [35].

### 3.5 performance measure

Evaluating the performance of the model is a significant step to ensure the model effectiveness in classifying between patients who suffer from sickle cell anemia and normal patients. Accuracy, Recall, Precision and F1-score were the measurements tools that used in this study along with confusion matrices that contains True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). These measures help to determine how well the model performs on unseen data to ensure optimal performance on real life scenarios.

- True Positive (TP): represents the number of patients who were correctly classified as diagnosed with SCA.
- False Positive (FP): represents the number of patients who were classified as diagnosed with SCA but actually are normal.
- True Negative (TN): represents the number of patients who were correctly classified as normal
- False Negative (FN): represents the number of patients who were classified as normal but actually are SCA patients

Accuracy is the ratio of accurately classified SCA and normal patients over the total number of patients. The equation below shows it mathematically.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Precision measures the proportion of true positive predictions among all positive predictions made. Equation 3 provides a mathematical representation of it.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

the proportion of true positive predictions among all actual positive instances is called recall. Equation 4 provides a mathematical representation of it.

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

F1-score can be calculated by using precision and recall which is shown in equation 5.

$$F1 - score = \frac{2(Precision \times Recall)}{Precision + Recall} \quad (5)$$

### 3.6 Optimization strategy

Hyperparameters are the parameters that need to be set before the training starts. The impact of selecting the optimal parameter is significant on the model's performance. GridSearchCV () was employed for the aim of finding the best hyperparameter for each model. Table 3 illustrate the hyperparameter that have been selecting by GridSearch CV () method with the original dataset.

<b>algorithm</b>	<b>Hyperparameter</b>	<b>Best Hyperparameter</b>
RF	Max_depth	None
	Min_samples_leaf	1
	Min_samples_split	2
	N_estimator	100
SVM	C	10
	Gamma	Scale
	kernel	Rbf
K-NN	algorithm	Auto
	N_neighbors	5
	weights	Distance
Gboost	Learning_rate	0.1
	Max_depth	5
	N_estimator	100
XGboost	Gamma	0.1
	Learning_rate	0.01
	Max_depth	3
Logistic Regression	N_estimator	150
	C	10
	penalty	12
Decision Tree	Max_depth	5
	Min_samplesleaf	1
	Min_samples_split	2
GaussianNB	-	-
AdaBoost	Learning rate	0.1
	N estimators	100
SVM_Gaussian	C	10
	gamma	scale

Table 3 Best hyperparameter with original dataset

The table below shows the best hyperparameter with original dataset and features selection.

<b>algorithm</b>	<b>Hyperparameter</b>	<b>Best Hyperparameter</b>
RF	Max_depth	None
	Min_samples_leaf	2
	Min_samples_split	5
	N_estimator	200
SVM	C	10
	Gamma	Scale
	kernel	Rbf
K-NN	algorithm	Auto
	N_neighbors	5
	weights	Distance
Gboost	Learning_rate	0.01
	Max_depth	3
	N_estimator	150
XGboost	Gamma	0

	Learning_rate	0.01
	Max_depth	5
	N_estimator	150
Logistic Regression	C	0.1
	Penalty	12
Decision Tree	Max_depth	5
	Min_samplesleaf	2
	Min_samples_split	2
GaussianNB	-	-
AdaBoost	Learning rate	0.1
	N estimators	100
SVM_Gaussian	C	10
	gamma	scale

Table 4 Best hyperparameters with original dataset and features selection

Table 5 represent the selected hyperparameters with oversampled dataset.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	None
	Min_samples_leaf	1
	Min_samples_split	2
	N_estimator	200
SVM	C	10
	Gamma	auto
	kernel	Rbf
K-NN	algorithm	Auto
	N_neighbors	3
	weights	Distance
Gboost	Learning_rate	0.1
	Max_depth	5
	N_estimator	100
XGboost	Gamma	0
	Learning_rate	0.01
	Max_depth	7
Logistic Regression	N_estimator	200
	C	10
	Penalty	12
Decision Tree	Max_depth	10
	Min_samplesleaf	1
	Min_samples_split	2
GaussianNB	-	-
AdaBoost	Learning rate	0.5
	N estimators	150
SVM_Gaussian	C	10
	gamma	auto

Table 5 Best hyperparameters with oversampled dataset

#### 4. Empirical Results

In this section, we report the results of our developed models after the application of GridSearchCV on the sampled data obtained from SMOTETomek, a technique that balances the distribution of classes in the dataset by undersampling the majority class and oversampling the minority class. [36] Following the process of identifying the optimal hyperparameters, training the model with all ten features through stratified 10-fold cross-validation, and displaying the results in Table 6, the next section offers a detailed analysis of the findings.

Classifier	Dataset	Testing Accuracy	Precision	Recall	F1-Score
Random Forest	Original	85.95%	86.00%	86.00%	83.00%
	Using SMOTETomek	75.15%	82.00%	75.00%	78.00%
	SMOTETomek				
SVM	Original	83.47%	80.00%	83.00%	79.00%
	Using SMOTETomek	61.49%	81.00%	61.00%	67.00%
	SMOTETomek				
KNN	Original	85.95%	86.00%	86.00%	83.00%
	Using SMOTETomek	62.11%	82.00%	62.00%	68.00%
	SMOTETomek				
Gradient Boosting	Original	84.29%	83.00%	84.00%	80.00%
	Using SMOTETomek	73.91%	83.00%	74.00%	77.00%
	SMOTETomek				
XGBoost	Original	84.29%	87.00%	84.00%	78.00%
	Using SMOTETomek	60.24%	84.00%	60.00%	66.00%
	SMOTETomek				
Logistic Regression	Original	80.99%	73.00%	81.00%	75.00%
	Using SMOTETomek	49.68%	75.00%	50.00%	57.00%
	SMOTETomek				
Decision Tree	Original	84.29%	82.00%	84.00%	80.00%
	Using SMOTETomek	70.80%	81.00%	71.00%	75.00%
	SMOTETomek				
AdaBoost	Original	81.81%	76.00%	82.00%	77.00%
	Using SMOTETomek	60.86%	83.00%	61.00%	67.00%
	SMOTETomek				
GaussianNB	Original	80.99%	78.00%	81.00%	79.00%
	Using SMOTETomek	67.70%	81.00%	68.00%	72.00%
	SMOTETomek				
SVM_Gaussian	Original	83.47%	80.00%	83.00%	79.00%
	Using SMOTETomek	61.49%	81.00%	61.00%	67.00%
	SMOTETomek				

Table 6 The results of the proposed models before and after sampling was applied.

The findings indicate that while the SMOTETomek oversampling strategy was implemented to reduce class imbalance, classifier performance was not consistently enhanced by this technique. All classifiers' testing accuracy was decreased by SMOTETomek in comparison to the original dataset. This result suggests that the original results will be considered as more reliable for this dataset. Random Forest proved to be the most effective method for modeling this dataset without the need for the SMOTETomek methodology, with the best testing accuracy of 85.95%. This investigation highlights how crucial it is to evaluate the SMOTETomek technique's impacts carefully and select the optimal classifier when dealing with unbalanced datasets. Thus, we conclude that using the original dataset without oversampling produced better results for our sickle cell anemia prediction task, with RandomForest appearing as the most successful classifier with an accuracy of 85.95%, a precision of 86.00%, recall of 86.00%, and F1-score of 83.00%. The number of samples for each class of the target feature, both before and after using SMOTETomek, is displayed in Figure 4. (Class 0 stands for N, and class 1 for SS.)

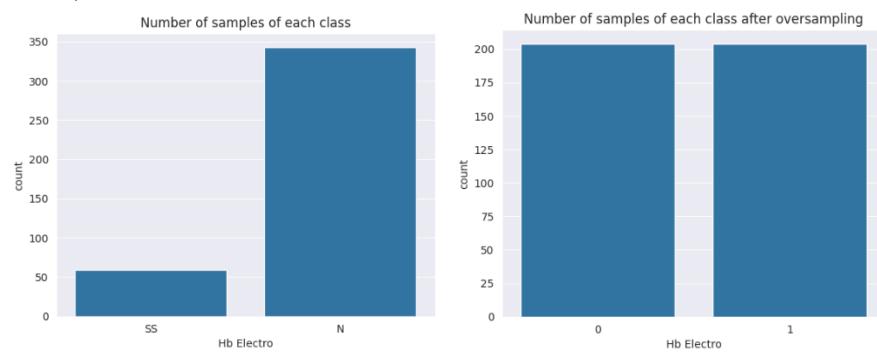


Figure 4 Number of samples of each class of the target feature before and after applying SMOTETomek.

#### 4.1 Results with Feature Selection

The feature selection utilized is the tree based embedded method to select the features with the highest importance to the chosen classifier [37]. In this research, a dataset of 400 patients at the pediatric ward were taken from Al Fashir Teaching Hospital, Sudan [8]. In this methodology the feature selection is embedded in the classifier learning process [37]. The ‘RandomForestClassifier’ is utilized in ‘SelectFromModel’ method for computing the feature importance scores during the training process. For each classifier the feature selection method chooses a subset of all the features and uses ‘RandomForestClassifier’ to calculate the feature importance score, then the feature subset with the highest score is chosen for the final modeling. Figure 5 demonstrate the process of feature selection.

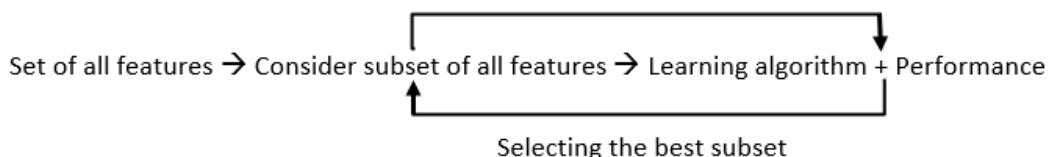


Figure 5 Feature selection process

The RF classifier with 5 features achieved the highest performance with an accuracy rate of 86.77%, a recall of 89%, a precision of 87% and 83% F1-score. Table 7 shows the result of the ML algorithms with feature selection with their best feature subsets, compared to the results of the classifiers with all features.

ML classifiers	All features (10)	With feature selection	Number of features	Feature subset
RF	85.95%	86.77%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
SVM	83.47%	82.64%	6	['PLTs', 'MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
KNN	85.95%	82.64%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
GB	84.29%	83.47%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
XGBoost	84.29%	85.12%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
LoR	80.99%	82.64%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
DT	84.29%	84.29%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
AdaBoost	81.81%	82.64%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
GaussianNB	80.99%	81.81%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
SVMGaussian	83.47%	82.64%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']

Table 7 Results with feature selection

## 4.2 Further Discussion of the Results

The confusion matrices provided offer insightful observations into the performance of various machine learning models in classifying instances into two categories: SS (sickle cell trait and sickle cell disease) and N (no sickle cell). Each model shows unique strengths and weaknesses in handling this binary classification problem.

Starting with **RandomForest** Table 8, we see a perfect classification for the SS category (100 correctly classified), but the model struggles slightly with the N category, misclassifying 5 cases as SS out of 21. This suggests a high sensitivity but potential overfitting to the SS category, which might lead to false positives in practical scenarios.

Transitioning to **SVM** Table 9, the model demonstrates robust performance with a high accuracy rate for both categories, accurately classifying 96 SS and 17 N cases. However, like RandomForest, it shows a tendency to misclassify 4 instances of N as SS. This indicates that SVM is generally effective but might also be slightly biased towards predicting SS.

**KNN** Table 10 exhibits a slightly lower performance compared to SVM, with more misclassifications: 93 SS correctly identified but 7 misclassified, and 14 N correctly identified with 7 misclassified. This could imply that KNN may require tuning of parameters such as the number of neighbors to improve its precision, especially in distinguishing SS from N effectively.

**GradientBoosting** Table 11 shows competitive accuracy, particularly in minimizing false negatives for the SS category with 99 correct predictions and 1 misclassification. However, it still misclassifies 2 out of 21 N instances as SS, similar to the previous models, suggesting a common challenge among these algorithms in avoiding type I errors in this dataset.

Similarly, **XGBoost** Table 12 aligns closely with GradientBoosting in performance, achieving high accuracy in the SS category with 100 correct predictions and no SS misclassifications but encountering some difficulty with 3 out of 21 N category misclassifications. Both models display a strong capability to identify SS correctly but need to improve in specificity to reduce the false positive rate.

**LogisticRegression** Table 13 provides a balanced approach with a strong classification record for both categories, accurately identifying 99 SS and 20 N cases but misclassifying 1 SS and 1 N. It presents a slightly higher rate of misclassifications compared to SVM and RandomForest, but its overall performance is solid, indicating good generalizability across different data distributions.

**DecisionTree** Table 14 shows variability in its performance, with a near-perfect score in classifying SS (99 correct, 1 misclassified) but a higher misclassification rate for N, correctly identifying 18 and misclassifying 3. This could reflect DecisionTree's sensitivity to the training dataset's nuances, potentially leading to overfitting and less robust predictions.

**AdaBoost** Table 15 offers balanced results, with decent accuracy in both categories: 98 SS correctly identified with 2 misclassified, and 19 N correctly identified with 2 misclassified. AdaBoost's performance suggests that while it is capable of handling diverse data, like many ensemble methods, it might also benefit from further tuning to enhance its discriminative ability.

**GaussianNB** Table 16 shows moderate sensitivity with 94 SS cases correctly identified and 6 misclassified. For the N category, it again misclassifies 5 out of 21 cases as SS, reinforcing the common challenge of managing false positives among these models.

Finally, in Table 17

**SVM with a Gaussian kernel** mirrors the performance of the standard SVM with 96 SS correctly classified and 4 misclassified. In the N category, it also misclassifies 4 cases as SS, indicating that the kernel choice does not significantly affect performance in this dataset but still highlights issues with potential overfitting to SS.

Overall, each model presents unique strengths and areas for improvement in classifying sickle cell anemia disease, highlighting the importance of model selection and parameter optimization in predictive healthcare analytics.

RandomForest		Predicted	
		SS	N
Actual	SS	100	0
	N	16	5

Table 8 confusion matrix of RF classifier

SVM		Predicted	
		SS	N
Actual	SS	96	4
	N	17	4

Table 9 confusion matrix of SVM classifier

KNN		Predicted	
		SS	N
Actual	SS	93	7
	N	14	7

Table 10 confusion matrix of KNN classifier

GradientBoosting		Predicted	
		SS	N
Actual	SS	99	1
	N	19	2

Table 11 confusion matrix of GBoost classifier

XGBoost		Predicted	
		SS	N
Actual	SS	100	0
	N	18	3

Table 12 confusion matrix of XGBoost classifier

LogisticRegression		Predicted	
		SS	N
Actual	SS	99	1
	N	20	1

Table 13 confusion matrix of Logistic Regression classifier

DecisionTree		Predicted	
		SS	N
Actual	SS	99	1
	N	18	3

Table 14 confusion matrix of Decision Tree classifier

AdaBoost		Predicted	
		SS	N
Actual	SS	98	2
	N	19	2

Table 15 confusion matrix of AdaBoost classifier

GaussianNB		Predicted	
		SS	N
Actual	SS	94	6
	N	16	5

Table 16 confusion matrix of GaussianNB classifier

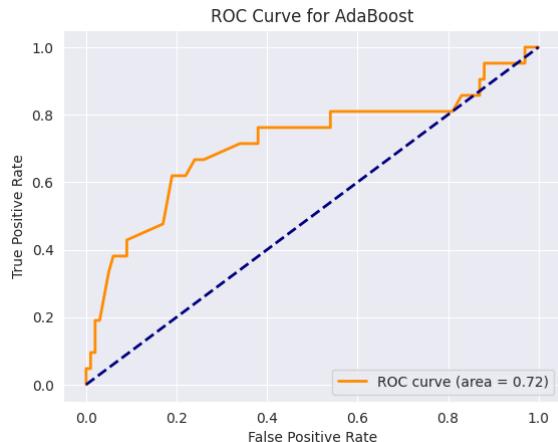
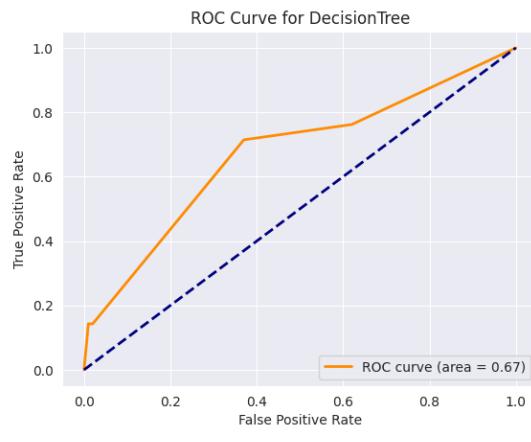
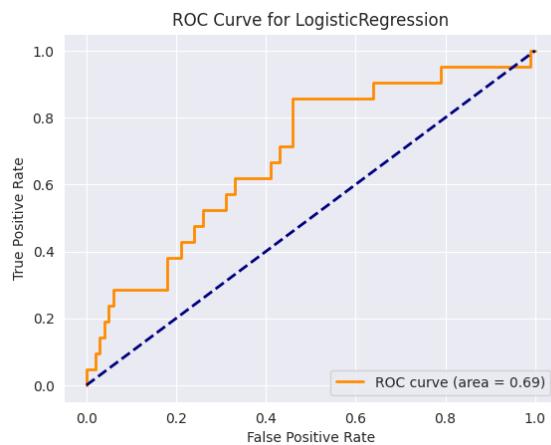
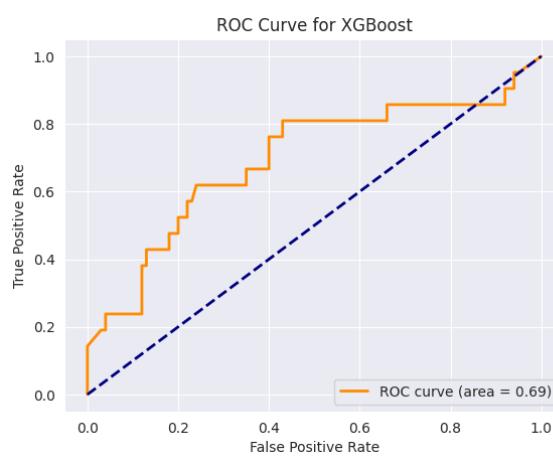
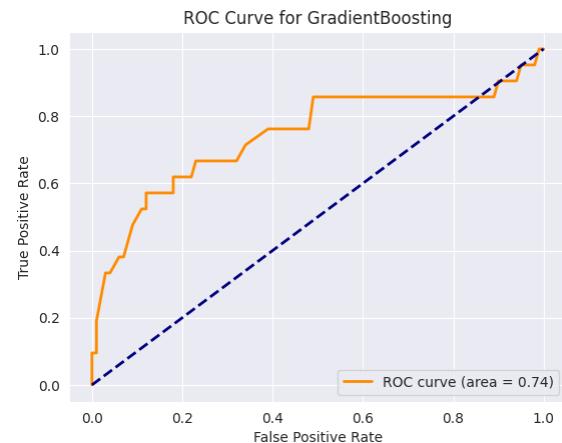
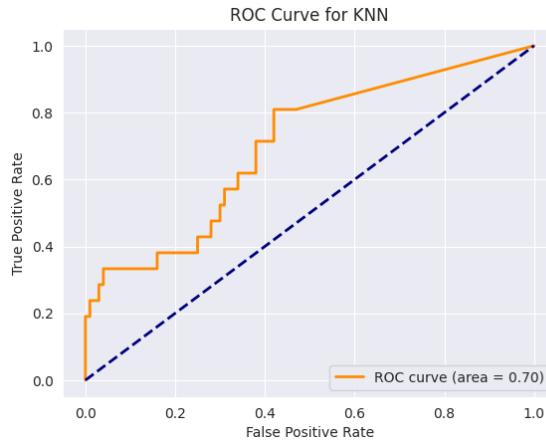
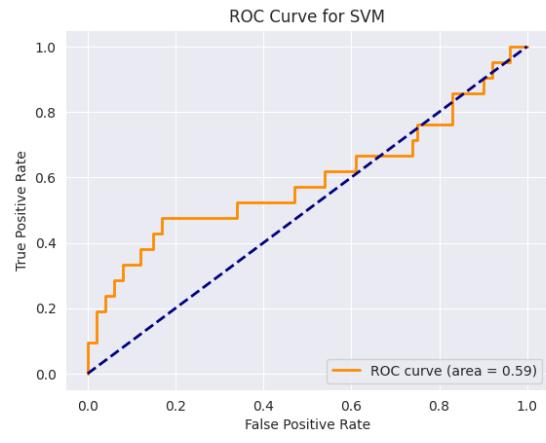
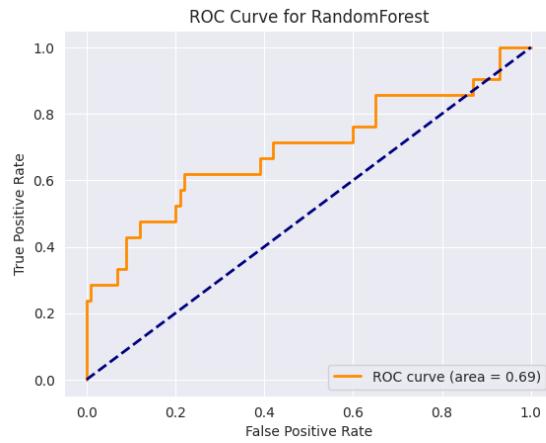
SVM_Gaussian		Predicted	
		SS	N
Actual	SS	96	4
	N	17	4

Table 17 confusion matrix of SVM Gaussian classifier

The ROC curves displayed across these figures measure the performance of various machine learning models utilized for predicting specific clinical outcomes. Each curve illustrates the true positive rate against the false positive rate for different thresholds, where a higher Area Under the Curve (AUC) indicates more effective model performance. The AUC serves as a summary metric, with a value of 1.0 depicting perfect prediction and 0.5 a random guess.

In this analysis, the Gradient Boosting model showcases the highest effectiveness with an AUC of 0.74, indicating a strong capability to distinguish between outcomes accurately. Following closely is the AdaBoost model, with an AUC of 0.72, demonstrating robust predictive power. The K-Nearest Neighbors (KNN) algorithm also performs well, achieving an AUC of 0.70, suggesting its usefulness in this specific predictive task. Both the Random Forest and XGBoost models report an AUC of 0.69, showing good, albeit not optimal, predictive accuracies.

On the other hand, Logistic Regression and the Decision Tree models display moderate effectiveness with AUCs of 0.69 and 0.67, respectively, indicating reasonable capabilities but with potential limitations in more complex scenarios. The Support Vector Machine (SVM) with a Gaussian kernel presents an AUC of 0.65, and the Gaussian Naive Bayes model has an AUC of 0.63, both reflecting fair performance. The standard SVM model trails with an AUC of 0.59, suggesting it might be less suited for this particular prediction task compared to the others. These insights underline the importance of model selection in clinical data analysis, where Gradient Boosting and AdaBoost emerge as particularly effective for high-stakes predictions. This comparative analysis provides a valuable benchmark, guiding researchers towards selecting models that not only maximize the true positive rate but also effectively minimize false positives in clinical predictions.



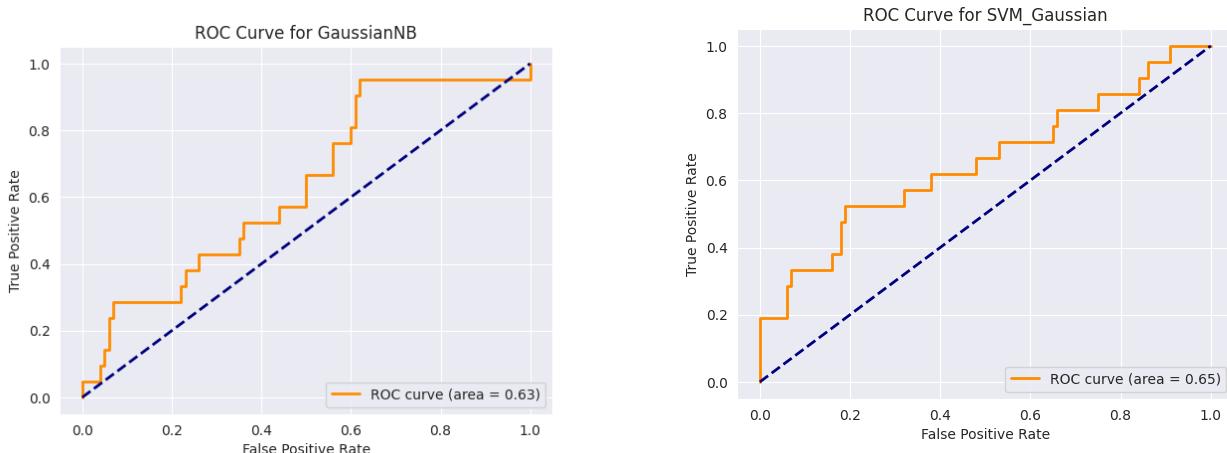


Figure 6 ROC curves of the ten models applied.

## 5 Interpretation of the Final Recommended Model

The field of artificial intelligence (AI) has witnessed a notable rise in interest recently due to the introduction of explainable artificial intelligence (XAI).[38] XAI algorithms provide comprehensible justifications for the predictions they make, therefore mitigating the intrinsic opacity of AI models. The nature of AI models can occasionally be obscure and hard to understand; these are what we call "black boxes," which means that the inner workings of the models are hidden or difficult to comprehend. [39] In this case, it becomes very important to interpret the final model that is suggested in machine learning sickle cell anemia prediction systems. Within the healthcare industry, where prompt and precise identification is essential, XAI plays a significant part in clarifying factors that impact forecasts, augmenting healthcare professionals' and patients' comprehension of risk factors along with important features that influence model results. In this work, two XAI techniques were used which are Feature Importance and LIME.

### 5.1. Feature Importance

One of the core concepts of eXplainable Artificial Intelligence (XAI) is feature importance, which seeks to explain how the algorithm chooses the most significant characteristics by assigning values to the input components based on how essential they were in making that decision [40]. By using this strategy, it is feasible to identify which features have a major impact on the model's predictions, providing valuable insights into the inner functioning of the model [39]. Feature importance analysis becomes useful in the context of sickle cell anemia prediction because it clarifies the essential clinical characteristics influencing disease susceptibility, facilitating reliable diagnosis and individualized treatment plans. The feature importance plot is displayed in Figure 7, with the length of the horizontal bar denoting the feature's importance. Hence, "PCV" is the most significant characteristic, whereas "RBCs" is the least significant.

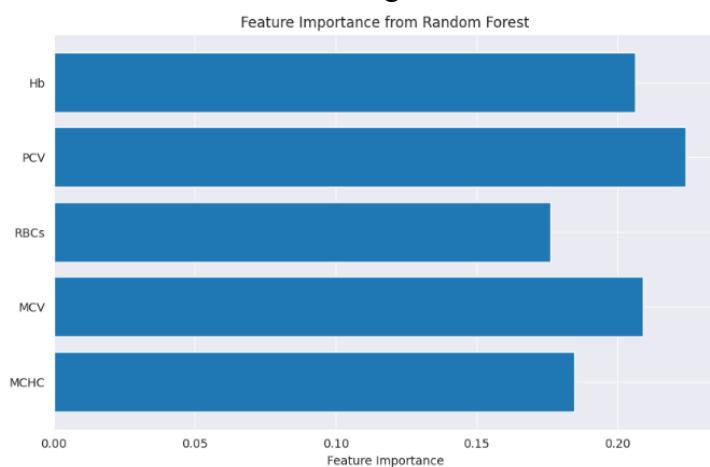


Figure 7 Feature importance of Random Forest.

## 5.2. Local Interpretable Model-Agnostic Explanations (LIME)

LIME is a method that generates local and understandable explanations for specific scenarios, providing a fresh answer to the problem of explaining predictions from complex black box models. [41] It sheds information on the model's decision-making process by varying the input data, creating a new model, and applying random perturbation to identify the relevance of features. LIME's model-agnostic nature allows it to be used with a variety of models following training, increasing the transparency and trustworthiness of AI applications. [42] Figure 8 illustrates the LIME map, which provides local prediction probabilities for normal and SS (sickle cell) patients using the Random Forest Classifier.

Figure 8a shows that the Random Forest model generated 96% of positive probability predictions. The image shows that features such as "MCHC", "MCV", Hb", and "PCV" helped to correctly classify the model for normal patients. In contrast, figure 8b explains the Random Forest model's negative prediction of 13%. The image shows that features such as "PCV", "MCHC", and "Hb" helped the model correctly classify Sickle Cell patients.

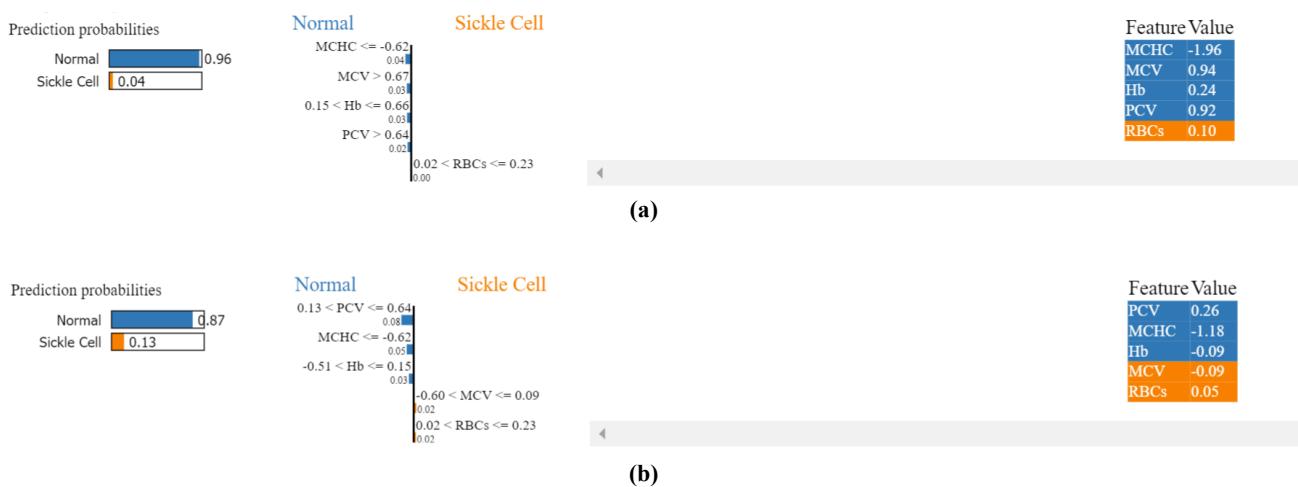


Figure 8 Lime prediction probability for the Logistic Regression model a) Normal class b) Sickle Cell class.

## 6. Discussion

The transformative impact of machine learning (ML) on the healthcare sector is multi-faceted, profoundly enhancing the analysis of vast amounts of data and refining patient care. ML algorithms excel in integrating diverse data types, including demographic information, lab results, imaging data, and free-text notes from healthcare professionals, into comprehensive predictive models for disease risk, diagnostics, and treatment planning. This capability facilitates personalized medicine and improves the efficiency of healthcare delivery by supporting decision-making processes and optimizing resource utilization. Notably, ML applications have progressed from experimental stages to becoming integral components of clinical settings, helping to predict patient outcomes and reduce the costs and time associated with medical testing, thereby boosting overall life expectancy [43], [44]. However, challenges such as data privacy, security, and the need for significant pre-processing of data to train ML models effectively must also be considered [45]. As the healthcare landscape continues to evolve, the integration of ML promises to enhance the standards of care delivery through data-driven insights and intelligent automation.

This work aims to improve sickle cell anemia early detection by the use of machine learning methods that only require clinical data, hence removing the need for imaging data. Ten machine learning algorithms are evaluated in this study using a dataset from a descriptive cross-sectional analysis that was carried out at Al Fashir Teaching Hospital in Sudan between December 2017 and August 2018. [8] The analysis included 400 patients. Using five characteristics, the Random Forest technique achieved 86.77% accuracy, 89% recall, 87% precision, and 83% F1 score, outperforming the other methods.

Sickle cell anemia can be caused by various features such as "PCV", "MCV", and "Hb". In our study "PCV" is the most important feature. [46] states that "PCV" is recognized as an important test in the assessment of many hematological disorders, such as sickle cell anemia. "PCV" calculates the percentage of red blood cells in a specific volume of blood, offering important information about the blood's ability to deliver oxygen. When it comes to sickle cell anemia, "PCV" is an important diagnostic metric that helps doctors determine the severity of the anemia and track the disease's evolution over time. As a diagnostic test for sickle cell anemia and similar hematological illnesses, "MCV" is very important too, according to [47], Red blood cell average volume or "MCV", provides important details on the size and morphology of the cells. Aberrations in "MCV" levels in the setting of sickle cell disease may point to differences in red blood cell morphology and size, which are distinguishing features of the illness.

A machine learning model's interpretability and explainability are crucial, particularly for medical applications like the detection of sickle cell anemia. The inner workings of predictive models must be understood by doctors in order to believe and accept suggestions made in the setting of healthcare, where decisions have a direct impact on the general health of patients [48]. Furthermore, interpretable models offer important insights into the underlying causes of sickle cell anemia diagnosis. Clinicians are better able to comprehend the biology of the disease and how it manifests in specific patients when the characteristics and patterns underlying the predictions are clarified. With this information, physicians may make better-informed decisions and customize treatment plans to meet the unique requirements and situations of each patient.

This paper's aim is to predict sickle cell anemia by employing clinical dataset. This has the main benefit of reducing interference and enhancing the detection process. Most of the research focused on using image-based datasets, however these medical checks are not always readily available and might be subject to influence.

## 7. Conclusion

Based on the extensive evaluation of machine learning algorithms for early diagnosis of Sickle Cell Anemia using clinical data, this research demonstrates the potential of these technologies to significantly enhance diagnostic precision and enable proactive treatment strategies. Employing a range of machine learning models, the study found that Random Forest algorithm, when integrated with clinical parameters, yields the highest diagnostic accuracy. These findings are pivotal, particularly in contexts where early diagnosis is crucial yet challenging due to resource constraints. The study further illustrates the effectiveness of machine learning models in transforming diagnostic processes, with Random Forest demonstrating the highest accuracy at 86.77% using a minimal set of five clinical features. This underscores the capacity of these models to function efficiently with streamlined data inputs, enhancing their applicability in diverse healthcare settings. Moreover, the application of Explainable Artificial Intelligence (XAI) techniques such as Feature Importance and Local Interpretable Model-Agnostic Explanations (LIME) provided valuable insights into the decision-making processes of the algorithms. This aspect of the research not only increases the transparency of the machine learning models but also assists healthcare providers in understanding the predictive factors influencing the algorithms' outcomes, thereby fostering trust and facilitating wider adoption in clinical practice.

Looking ahead, it would be beneficial to expand this research to include longitudinal studies that monitor patients over time, potentially incorporating real-time data to continuously refine the predictive models. Such dynamic approaches could further improve the accuracy and reliability of the predictions, ultimately enhancing patient outcomes in the management of Sickle Cell Anemia.

## References

- [1] Ministry of Health Portal Team, "Ministry Of Health Saudi Arabia," Ministry Of Health Saudi Arabia, 2019.<https://www.moh.gov.sa/en/HealthAwareness/EducationalContent/Diseases/Hematology/Pages/SickleCell-Anemia.aspx>.
- [2] WHO, "Sickle Cell Disease," WHO | Regional Office for Africa, 2017. <https://www.afro.who.int/health-topics/sickle-cell-disease>.
- [3] "Sickle Cell Disease - Sickle Cell Trait | NHLBI, NIH," www.nhlbi.nih.gov, Aug. 31, 2023. <https://www.nhlbi.nih.gov/health/sickle-cell-disease/sickle-cell-trait>
- [4] CDC, "What Is Sickle Cell Disease?," Centers for Disease Control and Prevention, Jul. 06, 2023. <https://www.cdc.gov/ncbddd/sicklecell/facts.html>
- [5] Elendu, C., Amaechi, D. C., Alakwe-Ojimba, C. E., Elendu, T. C., Elendu, R. C., Ayabazu, C. P., Aina, T. O., Aborisade, O., & Adenikinju, J. S. (2023). "Understanding Sickle cell disease: Causes, symptoms, and treatment options". *Medicine*, 102(38), e35237. <https://doi.org/10.1097/MD.00000000000035237>.
- [6] W. Jastaniah, "Epidemiology of sickle cell disease in Saudi Arabia," *Annals of Saudi Medicine*, vol. 31, no. 3, p. 289, 2011, doi: <https://doi.org/10.4103/0256-4947.81540>.
- [7] Arishi, W. A., Alhadrami, H. A., & Zourob, M. (2021). "Techniques for the Detection of Sickle Cell Disease: A Review". *Micromachines*, 12(5), 519. <https://doi.org/10.3390/mi12050519>.
- [8] Adam, M.A., Adam, N.K. & Mohamed, B.A. "Prevalence of sickle cell disease and sickle cell trait among children admitted to Al Fashir Teaching Hospital North Darfur State, Sudan". *BMC Res Notes* 12, 659 (2019). <https://doi.org/10.1186/s13104-019-4682-5>.
- [9] M. Darrin et al., "Classification of red cell dynamics with convolutional and recurrent neural networks: A sickle cell disease case study," *Nature News*, <https://www.nature.com/articles/s41598-023-27718-w> (accessed Oct. 8, 2023).
- [10] O. B. Ayoade, A. L. Imoize, J. A. Adeloye, and T. O. Oladele, "An Ensemble Models for the Prediction of Sickle Cell Disease from Erythrocytes Smears," ResearchGate, [https://www.researchgate.net/publication/374046799\\_An\\_Ensemble\\_Models\\_for\\_the\\_Prediction\\_of\\_Sickle\\_Cell\\_Disease\\_from\\_Erythrocytes\\_Smears](https://www.researchgate.net/publication/374046799_An_Ensemble_Models_for_the_Prediction_of_Sickle_Cell_Disease_from_Erythrocytes_Smears) (accessed Oct. 9, 2023).
- [11] D. Nguyen, L. Abraham, and A. Amatya, "Application of deep learning models into the prediction of interleukin ...," Application of Deep Learning Models into the Prediction of Interleukin-6 and -8 Cytokines in Sickle Cell Anemia Patients, [https://www.researchgate.net/publication/372051232\\_Application\\_of\\_Deep\\_Learning\\_Models\\_into\\_the\\_Prediction\\_of\\_Interleukin-6\\_and\\_-8\\_Cytokines\\_in\\_Sickle\\_Cell\\_Anemia\\_Patients](https://www.researchgate.net/publication/372051232_Application_of_Deep_Learning_Models_into_the_Prediction_of_Interleukin-6_and_-8_Cytokines_in_Sickle_Cell_Anemia_Patients) (accessed Oct. 8, 2023).
- [12] D. C. Saputra, K. Sunat, and T. Ratnaningsih, "A new artificial intelligence approach using extreme learning machine as the potentially effective model to predict and analyze the diagnosis of anemia," *Healthcare*, vol. 11, no. 5, p. 697, 2023. doi:10.3390/healthcare11050697
- [13] R. Vohra, A. Hussain, A. K. Dudyala, J. Pahareeya, and W. Khan, "Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting," *PLOS ONE*, vol. 17, no. 7, 2022. doi:10.1371/journal.pone.0269685.
- [14] C. Patgiri and A. Ganguly, "Machine Learning Techniques for Automatic Detection of Sickle Cell Anemia using Adaptive Thresholding and Contour-based Segmentation Method," ResearchGate, [https://www.researchgate.net/publication/369739403\\_Machine\\_Learning\\_Techniques\\_for\\_Automatic\\_Detection\\_of\\_Sickle\\_Cell\\_Anemia\\_using\\_Adaptive\\_Thresholding\\_and\\_Contour-based\\_Segmentation\\_Method](https://www.researchgate.net/publication/369739403_Machine_Learning_Techniques_for_Automatic_Detection_of_Sickle_Cell_Anemia_using_Adaptive_Thresholding_and_Contour-based_Segmentation_Method) (accessed Oct. 9, 2023).
- [15] M. Shahzad et al., "Identification of anemia and its severity level in a peripheral blood smear using 3-tier deep neural network," *Applied Sciences*, vol. 12, no. 10, p. 5030, 2022. doi:10.3390/app12105030
- [16] S. Srivastava, R. N. K, R. Srinivasan, N. K. Nambison, and S. S. Gorthi, "Diagnosis of sickle cell anemia using AutoML on UV-vis absorbance ...," Diagnosis of sickle cell anemia using AutoML on UV-Vis

- absorbance spectroscopy data,  
[https://www.researchgate.net/publication/356602448\\_Diagnosis\\_of\\_sickle\\_cell\\_anemia\\_using\\_AutoML\\_on\\_UV-Vis\\_absorbance\\_spectroscopy\\_data](https://www.researchgate.net/publication/356602448_Diagnosis_of_sickle_cell_anemia_using_AutoML_on_UV-Vis_absorbance_spectroscopy_data) (accessed Oct. 8, 2023).
- [17] S. Yeruva, M. S. Varalakshmi, B. P. Gowtham, Y. H. Chandana, and PESN. K. Prasad, "Identification of sickle cell anemia using deep neural networks," Emerging Science Journal, vol. 5, no. 2, pp. 200–210, 2021. doi:10.28991/esj-2021-01270
- [18] Abdulkarim, H.A. et al. (2020) 'A deep learning alexnet model for classification of red blood cells in sickle cell anemia', IAES International Journal of Artificial Intelligence (IJ-AI), 9(2), p. 221. doi:10.11591/ijai.v9.i2.pp221-228.
- [19] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, J. Zhang, and Y. Duan, "Deep learning models for classification of red blood cells in microscopy images to aid in sickle cell anemia diagnosis," Electronics, vol. 9, no. 3, p. 427, 2020. doi:10.3390/electronics9030427
- [20] Mohammed, A. et al. (2019) Machine learning predicts early-onset acute organ failure in critically ill patients with sickle cell disease [Preprint]. doi:10.1101/614941.
- [21] Alzubaidi, L. et al. (2019) 'Classification of red blood cells in sickle cell anemia using deep convolutional neural network', Advances in Intelligent Systems and Computing, pp. 550–559. doi:10.1007/978-3-030-16657-1\_51.
- [22] M. Xu, S. Abidi, M. Dao, and D. P. Papageorgiou, "A deep convolutional neural network for classification of red blood cells in sickle cell anemia," ResearcgGate, [https://www.researchgate.net/publication/320510877\\_A\\_deep\\_convolutional\\_neural\\_network\\_for\\_classification\\_of\\_red\\_blood\\_cells\\_in\\_sickle\\_cell\\_anemia](https://www.researchgate.net/publication/320510877_A_deep_convolutional_neural_network_for_classification_of_red_blood_cells_in_sickle_cell_anemia) (accessed Oct. 9, 2023).
- [23] Akrimi, J.A. et al. (2014) 'Classification red blood cells using support Vector Machine', Proceedings of the 6th International Conference on Information Technology and Multimedia [Preprint]. doi:10.1109/icimu.2014.7066642.
- [24] Scikit-Learn, "sklearn.preprocessing.StandardScaler — scikit-learn 0.21.2 documentation," Scikit-learn.org, 2019. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
- [25] Hairani, H., Anggrawan, A., & Priyanto, D. (n.d.). INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION journal homepage : www.jiov.org/index.php/jiov INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link. www.jiov.org/index.php/jiov
- [26] C. Cortes and V. Vapnik, "Support Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
- [27] J. Zhu et al., "Adaboost algorithm," 2009. [Online]. Available: <https://www.intlpress.com/site/pub/pages/journals/items/sii/content/vols/0002/0003/a008/>
- [28] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.
- [29] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
- [30] D. Hosmer Jr, S. Lemeshow, and R. Sturdivant, "Applied Logistic Regression," 3rd ed., Wiley, 2013.
- [31] J. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.
- [32] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and Regression Trees," Wadsworth, 1984.
- [33] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
- [34] I. Rish, "An empirical study of the naive Bayes classifier," in IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 2001.

- [35] B. Schölkopf et al., "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond," MIT Press, 2002.
- [36] R. A. A. Viadinugroho, "Imbalanced classification in Python: Smote-tomek links method," Medium, <https://towardsdatascience.com/imbalance-classification-in-python-smote-tomek-links-method-6e48dfe69bbc> (accessed Mar. 28, 2024).
- [37] GeeksforGeeks (2024) Feature selection techniques in machine learning, GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/> (Accessed: 04 May 2024).
- [38] Rachha, A., & Seyam, M. (2023). Explainable AI In Education : Current Trends, Challenges, And Opportunities. Conference Proceedings - IEEE SOUTHEASTCON, 2023-April, 232–239. <https://doi.org/10.1109/SoutheastCon51012.2023.10115140>
- [39] Alfeo, A. L., Zippo, A. G., Catrambone, V., Cimino, M. G. C. A., Toschi, N., & Valenza, G. (2023). From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks. Computer Methods and Programs in Biomedicine, 236. <https://doi.org/10.1016/j.cmpb.2023.107550>
- [40] Saarela, M., & Jauhainen, S. (2021). Comparison of feature importance measures as explanations for classification models. SN Applied Sciences, 3(2). <https://doi.org/10.1007/s42452-021-04148-9>
- [41] Gabbay, F., Bar-Lev, S., Montano, O., & Hadad, N. (2021). A lime-based explainable machine learning model for predicting the severity level of covid-19 diagnosed patients. Applied Sciences (Switzerland), 11(21). <https://doi.org/10.3390/app112110417>
- [42] Viswan, V., Shaffi, N., Mahmud, M., Subramanian, K., & Hajamohideen, F. (2023). Explainable Artificial Intelligence in Alzheimer's Disease Classification: A Systematic Review. In Cognitive Computation. Springer. <https://doi.org/10.1007/s12559-023-10192-x>
- [43] K. Ngiam and I. W. Khor, "Big data and machine learning algorithms for health-care delivery," *The Lancet. Oncology*, vol. 20, no. 5, pp. e262-e273, 2019. [Online]. Available: [https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045\(19\)30149-4/abstract](https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(19)30149-4/abstract).
- [44] C. Jutzeler and K. Borgwardt, "Machine Learning in Medicine," *Circulation*, vol. 132, no. 20, pp. 1920-1930, 2015. [Online]. Available: <https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.115.001593>.
- [45] A. Arora and N. Basu, "Machine Learning in Modern Healthcare," *International Journal of Advanced Medical Sciences and Technology*, 2023. [Online]. Available: <https://www.ijamst.latticescipub.com/portfolio-item/D3037063423/>.
- [46] "MCV (mean corpuscular volume): MedlinePlus Medical Test," MedlinePlus, <https://medlineplus.gov/lab-tests/mcv-mean-corpuscular-volume/> (accessed May 4, 2024).
- [47] S. A. A. Khaled et al., "Hematological, biochemical properties, and clinical correlates of hemoglobin S variant disorder: A new insight into sickle cell trait," *Journal of Hematology*, <https://thejh.org/index.php/jh/article/view/977/653> (accessed May 4, 2024).
- [48] R. Marcinkevičs and J. E. Vogt, "Interpretable and explainable machine learning: A methods-centric overview with concrete examples," *WIREs Data Mining and Knowledge Discovery*, Feb. 2023, doi: <https://doi.org/10.1002/widm.1493>.

## **Appendix F.1: Sickle Cell Anemia Conference Paper**

### **Machine Learning-Based Models for the Pre-emptive Diagnosis of Sickle Cell Anemia using Clinical Data**

Sunday O. Olatunji<sup>1\*</sup>, Mohammad Aftab Alam Khan<sup>1\*</sup>, Fai Alanazi<sup>1\*</sup>, Rahaf yaanallah<sup>1\*</sup>, Shahad alghamdi<sup>1\*</sup>, Razan Alshammari<sup>1\*</sup>, Fatimah Alkhatri<sup>1\*</sup>, Mehwash Farooqui<sup>1\*</sup>and Mohammed Imran Basheer Ahmed<sup>1\*</sup>

<sup>1</sup> College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia.

\* Correspondence: <sup>1</sup>[osunday@iau.edu.sa](mailto:sunday@iau.edu.sa) <sup>2</sup>[mkhan@iau.edu.sa](mailto:mkhan@iau.edu.sa)

<sup>3</sup>[220000931@iau.edu.sa](mailto:220000931@iau.edu.sa) <sup>4</sup>[2200003935@iau.edu.sa](mailto:2200003935@iau.edu.sa)

<sup>5</sup>[2200003434@iau.edu.sa](mailto:2200003434@iau.edu.sa) <sup>6</sup>[2200004035@iau.edu.sa](mailto:2200004035@iau.edu.sa)

<sup>7</sup>[2200001977@iau.edu.sa](mailto:2200001977@iau.edu.sa) <sup>8</sup>[mfarooqui@iau.edu.sa](mailto:mfarooqui@iau.edu.sa) <sup>9</sup>[mbahmed@iau.edu.sa](mailto:mbahmed@iau.edu.sa)

**Abstract.** Sickle cell anemia (SCA) is a chronic genetic condition that results in sickle-shaped red blood cells due to aberrant hemoglobin production. This condition results in impaired oxygen transport, vessel occlusion, and various complications, severely impacting patients' health and quality of life. In Saudi Arabia, SCA presents a significant public health challenge, particularly in the Eastern Province, where prevalence rates are notably high. Early diagnosis and intervention are crucial for effective management and improved outcomes. Leveraging machine learning (ML) algorithms alongside traditional diagnostic methods offers promising avenues for enhancing SCA diagnosis and prediction. By integrating clinical data from blood samples with ML techniques, this study seeks to develop accessible and interpretable predictive models for preemptive SCA diagnosis. Drawing on a dataset collected from a pediatric hospital-based study in Sudan, encompassing socio-demographic variables and clinical parameters, this research aims to overcome prior limitations in ML-based SCA diagnosis, particularly the scarcity of studies incorporating clinical data. Through the application of various ML algorithms, including KNN, XGBoost, and Random Forest, this study endeavors to facilitate targeted interventions and personalized treatment plans, ultimately mitigating the burden of SCA in affected populations. Results indicate significant improvements in diagnostic accuracy with Random Forest demonstrating the highest accuracy at 86.77% using only 5 features, alongside 89% precision, 87% recall, and 83% f1-score.

**Keywords:** Pre-emptive Diagnosis; Chronic diseases; Sickle cell anemia; Machine Learning Algorithms; Early diagnosis.

### **1. Introduction**

Sickle cell anemia is one of the chronic genetic conditions that affects red blood cells, as their shape changes from spherical to a crescent or sickle shape, making them fragile, weak, and breaking easily. This change results in the inability of red cells to effectively transport oxygen to the body's tissues and causes blockage of small blood vessels as a result of red cells sticking together, leading to symptoms such as chronic

pain attacks, shortness of breath, and pale skin [1]. Sickle cell anemia (SCA) is one of the most common genetic diseases and one of the most difficult health challenges in the world. Estimates show that the number of new cases of this disease ranges from 30,000 to 40,000 newborns annually [2].

Sickle cell disease, a potentially life-threatening condition, arises from abnormalities in hemoglobin, the key molecule responsible for transporting oxygen in red blood cells. Normally, hemoglobin consists of four protein chains: two alpha globins ( $\alpha$ -globin) and two beta globins ( $\beta$ -globin), forming what's known as hemoglobin A (HbA). The beta globin gene, or HBB gene, encodes the beta subunit of hemoglobin. Various mutations in the HBB gene lead to this disorder. Each person inherits two copies of the HBB gene, and when both copies carry mutations, the production of normal beta globin is hindered, leading to the disease. Sometimes, each copy can undergo different mutations, resulting in distinct types of abnormal beta subunits within the same individual. Sickle cell anemia (HbSS), the most severe form of this illness, occurs when both copies carry the same mutation, resulting in the production of mutant hemoglobin S [3].

The severe nature of sickle cell anemia and its associated complications pose significant risks to individuals' health and quality of life. Among these complications, stroke becomes a serious issue, wherein the abnormal shape of sickle cells can block blood vessels, depriving the brain of oxygen and causing long-term neurological damage. Acute chest syndrome, which is typified by fever, respiratory distress, and chest tightness, is another challenging issue [4]. In order to maximize the health outcomes for those with sickle cell anemia, comprehensive care techniques focused on avoiding, identifying, and managing these complex consequences are necessary.

In Saudi Arabia, Sickle Cell Anemia poses a significant burden on public health, with approximately 4.2% of the population carrying the sickle cell trait and around 0.26% of them affected by the disease [5]. Its effects are most noticeable in the Eastern Province, where the trait is most prevalent at 17% of the population; 1.2% of those afflicted with the illness. These statistics from the Ministry of Health highlight the urgency of addressing Sickle Cell Anemia within the country. Early diagnosis and intervention are paramount in mitigating the effects of the disease and improving patient outcomes. By prioritizing early diagnosis and access to comprehensive care, Saudi Arabia can significantly alleviate the burden of Sickle Cell Anemia on affected individuals and their families, enhancing both health outcomes and quality of life.

Using machine learning algorithms and clinical data from blood samples, along with traditional diagnostic techniques, could improve the early diagnosis and prediction of sickle cell anemia (SCA) in Saudi Arabia. By utilizing the power of machine learning, healthcare providers can develop predictive models that help stratify patients according to their risk of getting SCA, allowing for targeted interventions and personalized treatment plans [6]. Consequently, several studies have been conducted to diagnose SCA using ML algorithms. Nevertheless, these studies have mostly concentrated on using imaging datasets for SCA diagnosis, which increases the difficulty of obtaining data and adds complexity to the use of the model, especially for stakeholders who are not technical.

Therefore, by creating easily understood machine learning-based models for the proactive diagnosis of sickle cell anemia (SCA) using clinical data, this work seeks to

overcome the limitations of earlier studies. The dataset utilized in this study was collected through a descriptive cross-sectional hospital-based study conducted at Al Fashir Teaching Hospital in Sudan from December 2017 to August 2018 [7]. The study focused on children aged 0–18 years who were admitted to the hospital. A structured questionnaire capturing socio-demographic data such as age, gender, and tribe was administered to 400 patients at the pediatric ward during the study period. By using different machine learning methods, such as Random Forest, XGBoost, and KNN, the results indicate significant improvements in diagnostic accuracy with Random Forest demonstrating the highest accuracy at 86.77% using only 5 features, alongside 89% precision, 87% recall, and 83% f1-score.

## 2. Literature Review

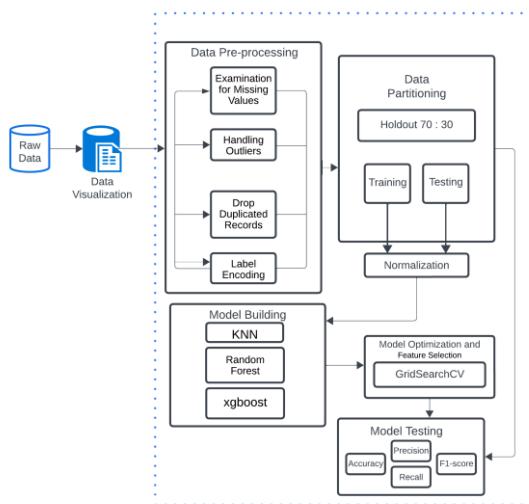
The research paper that was produced by Darrin et al. [8] explored the use of convolutional and recurrent neural networks in classifying red blood cells, specifically in sickle cell disease. They used a dataset of red cell images, extracting features using CNNs and leveraging RNNs to capture temporal dependencies. The model achieved high accuracy of 97% and an F1-score of 0.94. Another study by Ayoade et al. [9] compared three Machine Learning algorithms (MLR, XGBoost, RF) and their ensemble models for elongated erythrocyte shape identification in sickle cell disease prediction. Results showed individual algorithms achieved good accuracy, but hybrid models RF-MLR and RF-XGBoost outperformed with accuracies of 92% and an impressive 99%, respectively. Moreover, Shahzad et al. [10], used a 3-tier deep convolutional fused network to predict anemia severity using 11,500 pictures. The model, consisting of two modules, classifies healthy and mild/chronic anemia. The Tier-III model achieved the highest accuracy, recall, F1-Score, and specificity with 89.29%, 95.96%, 95.87%, and 96.34%, respectively. Abdulkarim et al. [11], developed a deep-learning AlexNet model for red blood cell classification in sickle cell anemia patients. They used 9,000 single RBC images from 130 patients and automated RBC extraction. The model achieved a classification prediction accuracy of 95.92% for identifying abnormalities in RBCs.

Previous studies primarily focused on imaging datasets, neglecting clinical data and limiting models' reliability due to limited sample numbers and insufficient clinical data. This study aims to improve sickle cell anemia prediction by developing machine learning models using clinical data, filling a gap in previous studies.

## 3. Materials and Methods

This study aimed to develop a preemptive model for diagnosing sickle cell anemia utilizing the Python programming language. The procedure included importing the dataset, applying analysis using Pandas, handling duplicates and missing values, and visualizing the distribution of data and identifying outliers. Categorical features with missing values were imputed using the most frequent strategy, and data cleaning

procedures were conducted to standardize categorical values. Subsequently, categorical features were encoded into numerical representations using label encoding. The dataset was then partitioned into training (70%) and testing (30%) sets to facilitate model evaluation. Feature scaling was performed using standardization to ensure uniformity in feature magnitudes across the dataset. Three machine learning algorithms, including Random Forest, KNN, and XGBoost were implemented for predictive modeling. Hyperparameter tuning was conducted using GridSearchCV to optimize algorithm performance. Evaluation of model performance involved assessing accuracy scores, confusion matrices, and classification reports on both training and testing datasets. The complete procedural workflow is illustrated in Figure 1.



**Fig 1.** The proposed framework for the pre-emptive diagnosis of Sickle Cell Anemia

### 3.1 Data Description

The dataset for this study was derived from a descriptive cross-sectional analysis conducted at Al Fashir Teaching Hospital in Sudan, spanning from December 2017 to August 2018 [7]. This study included 400 pediatric patients who were admitted to the hospital during the specified period. These patients, aged between 0 and 18 years, were approached to participate in the study, with their parents providing informed consent. The selection process employed random sampling, ensuring equal opportunity for any child admitted to the pediatric ward to be included.

To gather sociodemographic information, including age, gender, and tribal affiliation, a structured questionnaire was used. 400 parents gave their consent, indicating a moderate percentage of refusal. Table 1 shows each feature with its corresponding type.

**Table 182.** Features' description

Feature	Type
Hb Electro	object
Platelets	Integer

Total White Blood Cells	float64
Mean Cell Hemoglobin Concentration	float64
Mean Cell Hemoglobin	float64
Mean Cell Volume	float64
Red Blood Cells	float64
Packed Cell Volume	float64
Hemoglobin	float64
Tribe	Integer
Age/ Years	object
Sex	object
No	Integer

### 3.2 Statistical Analysis

Statistical analysis is an essential process to explore and understand the characteristics of both categorical and numerical data. In addition, determine any further needed preprocessing technique is needed. This will ensure that the model will perform smoothly with the highest possible accuracy.

Numerical attributes were explored through various statistical matrices to ensure the understanding of these attribute and its characteristics. The dataset contains 226 female patient and 176 male patients. Table 2 shows the statistical matrices for the numerical attribute.

**Table 183.** Statistical analysis of numerical features

Feature	Mean	Standard deviation	Min	25 <sup>th</sup> quartile	50 <sup>th</sup> quartile	75 <sup>th</sup> quartile	Max	Missing value counts
PLTs	284.59	134.52	21.00	195.25	272.00	351.00	780.00	0
TWBCs	7.95	6.27	1.00	4.60	6.65	9.40	61.70	0
MCHC	32.67	1.80	25.80	31.60	32.80	33.80	40.20	0
MCH	27.06	2.99	15.90	24.52	27.30	28.90	39.50	0
MCV	82.57	8.61	34.00	77.90	83.05	88.00	107.70	0
PCV	35.48	6.77	11.20	31.82	36.20	39.87	53.00	0
RBCs	4.47	1.53	1.14	4.07	4.49	4.83	22.10	0
Hb	11.61	2.23	3.50	10.50	11.90	13.00	17.70	0
Tribe	11.06	13.09	1.00	2.00	4.00	15.75	59.00	0

### 3.3 Data Pre-processing

Raw data usually comes along with some issues that prevent the model from performing well. The data might contain noise, outliers or missing values. Before starting to construct the model, these issues must be deal with in order to get the optimal performance and highest possible accuracy. To get a clean data set, any noise was removed from the dataset, moreover the dataset in this study has no missing nor duplicate values. In the target attribute, the class S (sickle cell trait) and SS (sickle cell trait) were combined as one class. Deleting unnecessary columns such as patients' number and date is essential to enhance accuracy. At the beginning the dataset had 13 features and after deleting unwanted features it ended up with 11 including the target class.

Standardizing the data is one of the common requirements for building a machine learning model as it enhances the performance and the coverage speed. To meet this requirement StandardScaler() function was utilized. The function works by eliminating the mean and scaling them to unit variance. To calculate the standard score for a sample data X the function needs the mean and standard deviation. This is shown in equation 1 [12].

$$Z = \frac{(X - u)}{s} \quad (1)$$

### 3.4 Description of Utilized Machine Learning Algorithms

**Random Forest:** Random Forest is an ensemble of Decision Trees, typically trained via the bagging method. The individual trees operate as weak learners, while their collective decisions, taken by majority voting, provide robust predictions against overfitting. Random Forest performs well on large datasets with high dimensionality [13].

**XGBoost (Extreme Gradient Boosting):** XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. It is renowned for its performance and speed in classification tasks. XGBoost provides a scalable and efficient solution, often winning many Kaggle competitions [14].

**K-Nearest Neighbors (KNN):** KNN is a simple, instance-based learning algorithm where the class of a sample is determined by the majority of the classes of its nearest neighbors. It is highly adaptable and easy to implement, making it suitable for solving both classification and regression problems. However, it becomes significantly slower as the size of the data increases [15].

### 3.5 Optimization strategy

Hyperparameters are the parameters that need to be set before the training starts. The impact of selecting the optimal parameter is significant on the model's performance. GridSearchCV () was employed for the aim of finding the best hyperparameter for each model. Table 3 and 4 illustrate the hyperparameter that have been selecting by GridSearchCV ()�.

**Table 3.** The best hyperparameter selected without feature selection.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	None
	Min_samples_leaf	1
	Min_samples_split	2
	N_estimator	100
XGboost	Gamma	0.1
	Learning_rate	0.01
	Max_depth	3
	N_estimator	150
KNN	algorithm	Auto
	N_neighbors	5
	weights	Distance

**Table 4.** The best hyperparameter selected with feature selection.

algorithm	Hyperparameter	Best Hyperparameter
RF	Max_depth	None
	Min_samples_leaf	2
	Min_samples_split	5
	N_estimators	200
XGboost	Max_depth	None
	Min_samples_leaf	2
	Min_samples_split	5
	N_estimators	200
KNN	algorithm	Auto
	N_neighbors	5
	weights	Distance

#### 4. Empirical Results

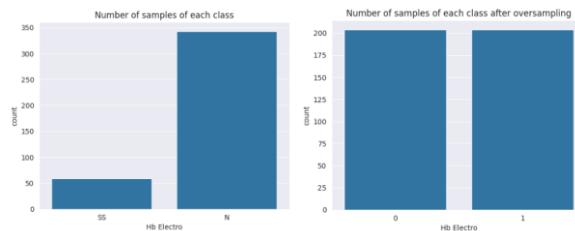
In this section, we report the results of our developed models after the application of GridSearchCV on the sampled data obtained from SMOTETomek, a technique that balances the distribution of classes in the dataset by undersampling the majority class and oversampling the minority class. [16] Following the process of identifying the optimal hyperparameters, training the model with all ten features through stratified 10-fold cross-validation, and displaying the results in Table 5, the next section offers a detailed analysis of the findings.

**Table 5.** The results of the proposed models.

Classifier	Dataset	Testing Accuracy	Precision	Recall	F1-Score
Random Forest	Original	85.95%	86.00%	86.00	83.00%
	Using SMOTETomek	75.15%	82.00%	75.00%	78.00%
KNN	Original	85.95%	86.00%	86.00%	83.00%
	Using SMOTETomek	62.11%	82.00%	62.00%	68.00%
XGBoost	Original	84.29%	87.00%	84.00%	78.00%
	Using SMOTETomek	60.24%	84.00%	60.00%	66.00%

The findings indicate that while the SMOTETomek oversampling strategy was implemented to reduce class imbalance, classifier performance was not consistently enhanced by this technique. All classifiers' testing accuracy was decreased by SMOTETomek in comparison to the original dataset. This result suggests that the original results will be considered as more reliable for this dataset. Random Forest proved to be the most effective method for modeling this dataset without the need for

the SMOTETomek methodology, with the best testing accuracy of 85.95%. This investigation highlights how crucial it is to evaluate the SMOTETomek technique's impacts carefully and select the optimal classifier when dealing with unbalanced datasets. Thus, we conclude that using the original dataset without oversampling produced better results for our sickle cell anemia prediction task, with RandomForest appearing as the most successful classifier with an accuracy of 85.95%, a precision of 86.00%, recall of 86.00%, and F1-score of 83.00%. The number of samples for each class of the target feature, both before and after using SMOTETomek, is displayed in Figure 2. (Class 0 stands for N, and class 1 for SS.)



**Fig 2.** Number of samples of each class before and after applying SMOTETomek.

#### 4.1 Results with Feature Selection

The feature selection utilized is the tree based embedded method to select the features with the highest importance to the chosen classifier [17]. In this research, a dataset of 400 patients at the pediatric ward were taken from Al Fashir Teaching Hospital, Sudan [7]. In this methodology the feature selection is embedded in the classifier learning process [17]. The ‘RandomForestClassifier’ is utilized in ‘SelectFromModel’ method for computing the feature importance scores during the training process. For each classifier the feature selection method chooses a subset of all the features and uses ‘RandomForestClassifier’ to calculate the feature importance score, then the feature subset with the highest score is chosen for the final modeling.

The RF classifier with 5 features achieved the highest performance with an accuracy rate of 86.77%, a recall of 89%, a precision of 87% and 83% F1-score. Table 6 shows the result of the ML algorithms with feature selection with their best feature subsets, compared to the results of the classifiers with all features.

**Table 6.** Results with feature selection

ML classifiers	All features (10)	With feature selection	Number of features	Feature subset
RF	85.95%	86.77%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
KNN	85.95%	82.64%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']
XGBoost	84.29%	85.12%	5	['MCHC', 'MCV', 'RBCs', 'PCV', 'Hb']

The confusion matrices provided offer insightful observations into the performance of various machine learning models in classifying instances into two categories: SS (sickle cell trait and sickle cell disease) and N (no sickle cell). Each model shows unique strengths and weaknesses in handling this binary classification problem. Table 7, we see a perfect classification for the SS category (100 correctly classified), but the model struggles slightly with the N category, misclassifying 5 cases as SS out of 21.

**Table 7.** confusion matrix of Random Forest algorithm.

		Predicted	
		SS	N
Actual	SS	100	0
	N	16	5

## 5. Conclusion

Based on the extensive evaluation of machine learning algorithms for early diagnosis of Sickle Cell Anemia using clinical data, this research demonstrates the potential of these technologies to significantly enhance diagnostic precision and enable proactive treatment strategies. Employing a range of machine learning models, the study found that Random Forest algorithm, when integrated with clinical parameters, yields the highest diagnostic accuracy. These findings are pivotal, particularly in contexts where early diagnosis is crucial yet challenging due to resource constraints. The study further illustrates the effectiveness of machine learning models in transforming diagnostic processes, with Random Forest demonstrating the highest accuracy at 86.77% using a minimal set of five clinical features. This underscores the capacity of these models to function efficiently with streamlined data inputs, enhancing their applicability in diverse healthcare settings.

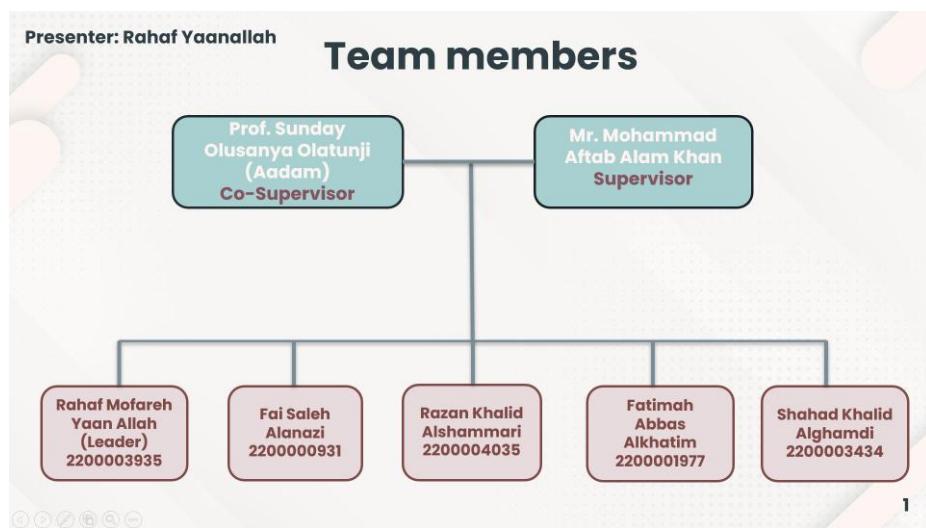
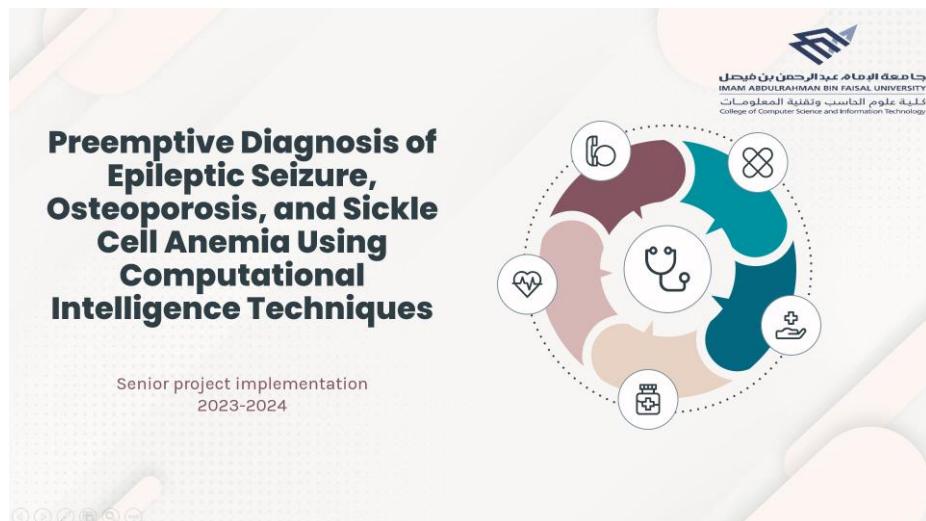
Looking ahead, it would be beneficial to expand this research to include longitudinal studies that monitor patients over time, potentially incorporating real-time data to continuously refine the predictive models. Such dynamic approaches could further improve the accuracy and reliability of the predictions, ultimately enhancing patient outcomes in the management of Sickle Cell Anemia.

## References

- Ministry of Health Portal Team, "Ministry Of Health Saudi Arabia," Ministry Of Health Saudi Arabia, 2019.<https://www.moh.gov.sa/en/HealthAwareness/EducationalContent/Diseases/Hematology/Pages/SickleCell-Anemia.aspx>.
- WHO, "Sickle Cell Disease," WHO | Regional Office for Africa, 2017. <https://www.afro.who.int/health-topics/sickle-cell-disease>.
- CDC, "What Is Sickle Cell Disease?," Centers for Disease Control and Prevention, Jul. 06, 2023. <https://www.cdc.gov/ncbddd/sicklecell/facts.html>
- Elendu, C., Amaechi, D. C., Alakwe-Ojimba, C. E., Elendu, T. C., Elendu, R. C., Ayabazu, C. P., Aina, T. O., Aborisade, O., & Adenikinju, J. S. (2023)."Understanding Sickle cell

- disease: Causes, symptoms, and treatment options". Medicine, 102(38), e35237. <https://doi.org/10.1097/MD.00000000000035237>.
5. W. Jastaniah, "Epidemiology of sickle cell disease in Saudi Arabia," Annals of Saudi Medicine, vol. 31, no. 3, p. 289, 2011, doi: <https://doi.org/10.4103/0256-4947.81540>.
  6. Arishi, W. A., Alhadrami, H. A., & Zourob, M. (2021). "Techniques for the Detection of Sickle Cell Disease: A Review". Micromachines, 12(5), 519. <https://doi.org/10.3390/mi12050519>.
  7. Adam, M.A., Adam, N.K. & Mohamed, B.A. "Prevalence of sickle cell disease and sickle cell trait among children admitted to Al Fashir Teaching Hospital North Darfur State, Sudan". BMC Res Notes 12, 659 (2019). <https://doi.org/10.1186/s13104-019-4682-5>.
  8. M. Darrin et al., "Classification of red cell dynamics with convolutional and recurrent neural networks: A sickle cell disease case study," Nature News, <https://www.nature.com/articles/s41598-023-27718-w> (accessed Oct. 8, 2023).
  9. O. B. Ayoade, A. L. Imoize, J. A. Adeloye, and T. O. Oladele, "An Ensemble Models for the Prediction of Sickle Cell Disease from Erythrocytes Smears," ResearchGate, [https://www.researchgate.net/publication/374046799\\_An\\_Ensemble\\_Models\\_for\\_the\\_Prediction\\_of\\_Sickle\\_Cell\\_Disease\\_from\\_Erythrocytes\\_Smears](https://www.researchgate.net/publication/374046799_An_Ensemble_Models_for_the_Prediction_of_Sickle_Cell_Disease_from_Erythrocytes_Smears) (accessed Oct. 9, 2023).
  10. M. Shahzad et al., "Identification of anemia and its severity level in a peripheral blood smear using 3-tier deep neural network," Applied Sciences, vol. 12, no. 10, p. 5030, 2022. doi:10.3390/app12105030
  11. Abdulkarim, H.A. et al. (2020) 'A deep learning alexnet model for classification of red blood cells in sickle cell anemia', IAES International Journal of Artificial Intelligence (IJ-AI), 9(2), p. 221. doi:10.11591/ijai.v9.i2.pp221-228.
  12. Scikit-Learn, "sklearn.preprocessing.StandardScaler — scikit-learn 0.21.2 documentation," Scikit-learn.org, 2019. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
  13. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
  14. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016
  15. T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.
  16. R. A. A. Viadinugroho, "Imbalanced classification in Python: Smote-tomek links method," Medium, <https://towardsdatascience.com/imbalanced-classification-in-python-smote-tomek-links-method-6e48dfe69bbc> (accessed Mar. 28, 2024).
  17. GeeksforGeeks (2024) Feature selection techniques in machine learning, GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/> (Accessed: 04 May 2024).

## Appendix G: Final Presentation



Presenter: Rahaf Yaanallah

## AGENDA

**01.**  
INTRODUCTION

**02.**  
IMPLEMENTATION

**03.**  
WEBSITE  
INTERFACES

**04.**  
SOFTWARE  
TESTING

**05.**  
CONCLUSION

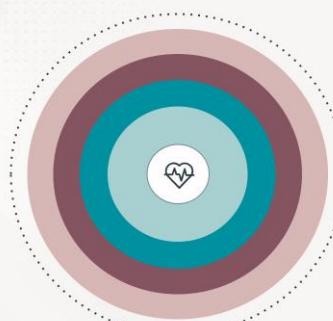
**06.**  
DEMO



2

Presenter: Rahaf Yaanallah

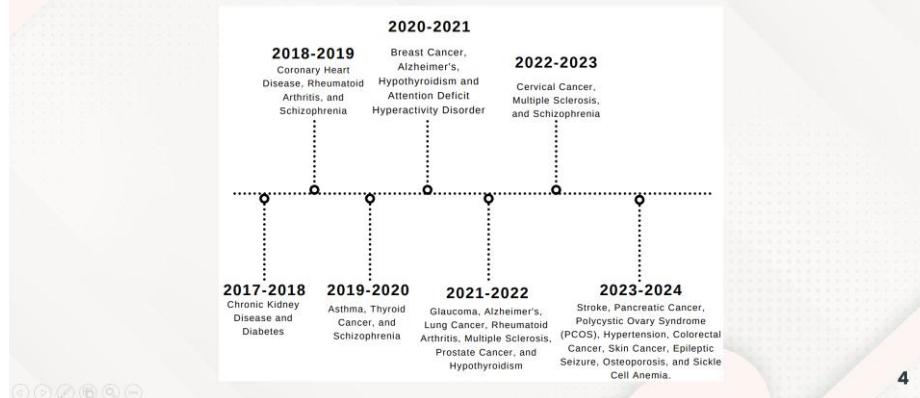
## INTRODUCTION



3

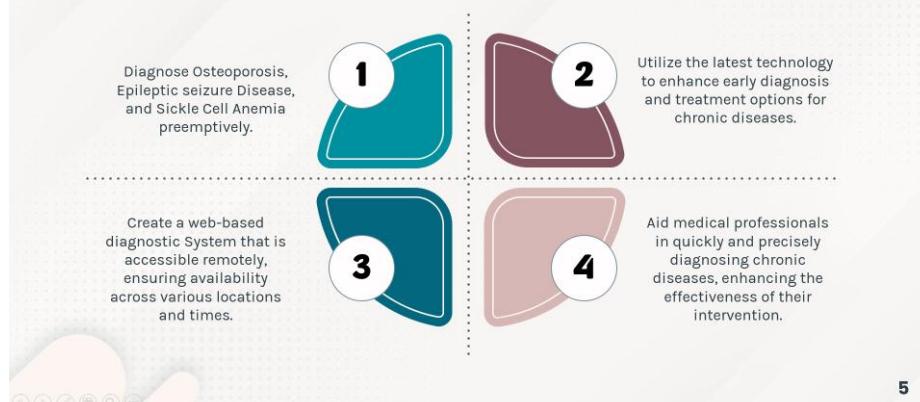
Presenter: Rahaf Yaanallah

## Project Phases



Presenter: Rahaf Yaanallah

## Overview



Presenter: Rahaf Yaanallah

## Changes from the proposal



**Dataset Update:** The datasets for epileptic seizures and sickle cell anemia were changed to enhance diagnostic precision for these specific diseases.



**Algorithm Expansion:** Expanded our computational methods by incorporating five new algorithms: Logistic Regression, SVM with Gaussian kernel, Gaussian Naive Bayes, Decision Tree, and AdaBoost.

6



Presenter: Fatimah Alkhatim

## Implementation



7

## Methodology

### ● LoR

A statistical method used for binary classification problems by modeling the probability that a given input belongs to a particular class. It uses the logistic function to constrain the output between 0 and 1.

### ● KNN

A non-parametric algorithm that classifies instances based on the majority class among its  $k$  nearest neighbors in the feature space. It relies on distance metrics, such as Euclidean distance, to determine the proximity of instances.

### ● SVM

Supervised learning algorithm that finds the optimal hyperplane to separate different classes in the feature space. It maximizes the margin between the closest points of the classes, known as support vectors, to improve generalization.

8

### ● RF

An ensemble learning method that constructs multiple decision trees during training and merges their results to improve accuracy and control overfitting. Each tree is built from a random subset of the training data and features, ensuring diversity among the trees.

### ● Methodology

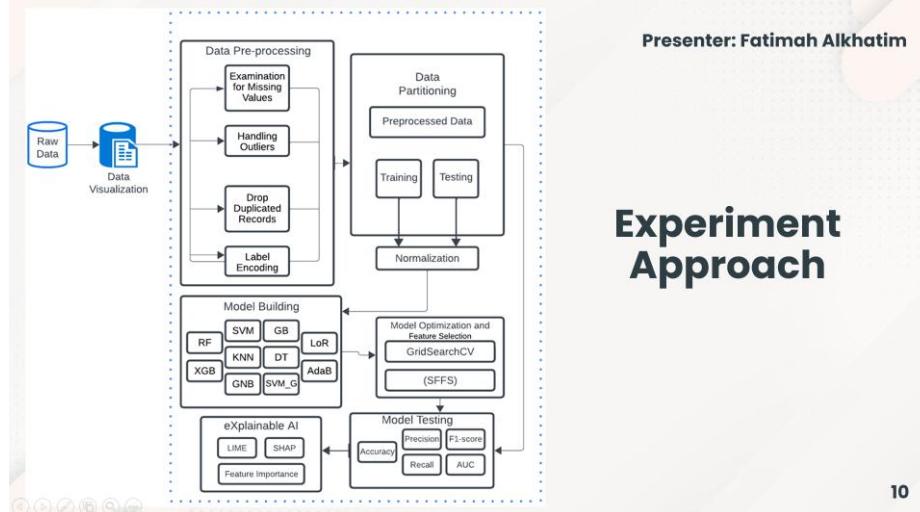
### ● GBoost

An ensemble technique that builds models sequentially, where each new model attempts to correct the errors made by the previous ones. It optimizes a loss function by adding new models that minimize the residuals of the previous models.

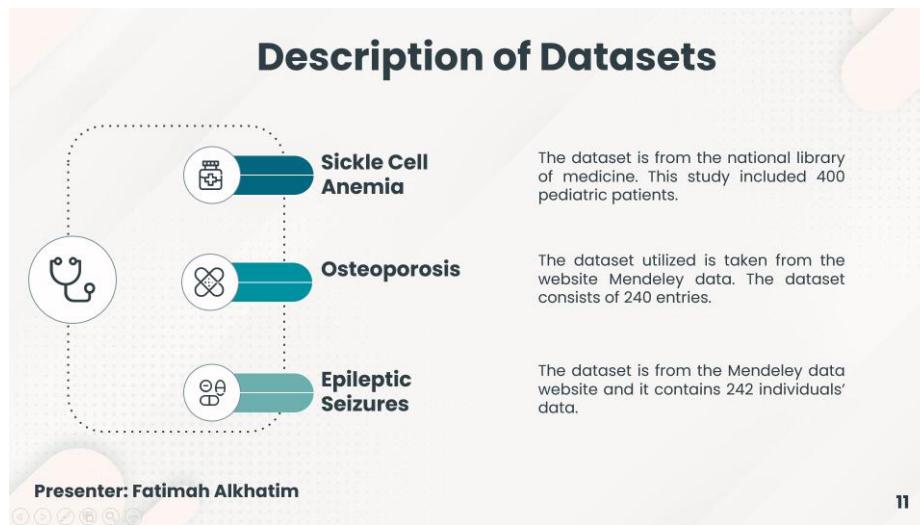
### ● XGBoost

Advanced implementation of gradient boosting that is designed for efficiency and performance. It includes regularization terms to prevent overfitting and supports parallel and distributed computing.

9



## Experiment Approach



## Results of Osteoporosis disease

Technique	Accuracy
RF	91.11 %
SVM	87.77 %
XGBoost	87.77 %

Random forest algorithm achieved the highest accuracy with 91.11 % with all 20 feature and SMOTETomek oversampling technique

Presenter: Shahad alghamdi

12

## Results of Epileptic Seizure disease

Technique	Accuracy
LoR	87.67 %
SVM	86.30 %
GBoost	82.19 %

Logistic Regression algorithm achieved the highest accuracy with 87.67 % with all 36 feature

Presenter: Shahad alghamdi

13

## Results of Sickle Cell Anemia disease

Technique	accuracy
RF	86.77 %
KNN	85.95 %
XGBoost	85.12 %

Random forest algorithm achieved the highest accuracy with 86.77% with 5 feature

Presenter: Shahad alghamdi

14

## Website Interfaces



Presenter: Razan Alshammari

15

## Website Developing Process

The website was divided into two separate parts:

### The Client side



- o HTML (HyperText Markup Language)
- o CSS (Cascading Style Sheets)
- o JavaScripts
- o JQuery

### The Server side



- o Flask used as Python's web framework.

Presenter: Razan Alshammary

16

## Users

### Admin

- Manage users' accounts.
- Generate, update, and view models.
- Manage the system's database.

### Registered User

- Create and delete Accounts
- Manage Personal Information.
- Diagnose themselves, print diagnosis results.

### Guest User

- diagnose themselves, print diagnosis results

### Laboratory Specialist

- Update profile information.
- Enter the patients' test data for disease diagnosis.

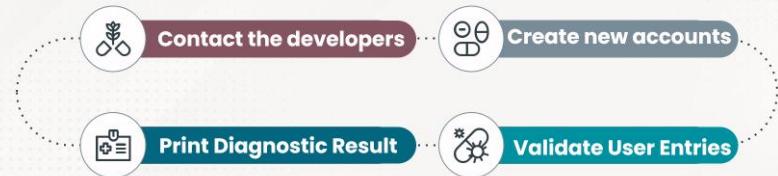
### Medical Specialist

- Add new patients.
- Diagnose patients

Presenter: Razan Alshammary

17

## Additional Website Functionalities



Presenter: Razan Alshammari

18

## Diagnosis Interface

Enables medical specialists, registered users, and guests to perform diagnosis by entering the required information.  
The result will indicate whether a patient will possibly have the disease or not.

The Diagnosis Interface screenshot shows a navigation bar with "Diagnose" and "Home - Diagnose". Below is a list of diseases: Chronic Kidney Disease, Diabetes, Rheumatoid Arthritis Disease, Asthma Disease, Thyroid Cancer, Schizophrenia, Hypothyroidism, Prostate Cancer, Alzheimer's Disease, Lung Cancer, Glaucoma, Liver Cirrhosis, Parkinson's Disease, Multiple Sclerosis, Cervical Cancer, Hepatitis C, Coronary Heart Disease, Depression, Chronic Obstructive Pulmonary Disease, Sickle Cell Anemia, and Osteoporosis. A sidebar lists "Epileptic Seizure". The main form is titled "Sickle Cell Anemia" and includes three checked input fields: "Hemoglobin", "Packed Cell Volume (PCV)", and "Red Blood Cells (RBC)".

Presenter: Razan Alshammari

19

## Medical Diagnosis Interface

The screenshot shows a web-based medical diagnosis system. On the left, a sidebar has 'Dashboard' and 'Diagnose history' options. The main area is titled 'Patient Information' with a search bar containing 'W0nK33'. Below it is a list of diseases: Chronic Kidney Disease, Diabetes, Coronary Heart Disease, Hypertension, Mammalian Adhesive Disease, Alzheimer Disease, Thyroid Cancer, Schizophrenia, Hypothyroidism, Prostate Cancer, Alzheimer's Disease, Lung Cancer, Glaucoma, Liver Cirrhosis, Parkinson Disease, Multiple Sclerosis, Gastroesophageal Reflux Disease, Chronic Obstructive Pulmonary Disease, Sickle Cell Anemia, Osteoporosis, and Epileptic Seizure. At the bottom, 'Chronic Kidney Disease' is highlighted.

Presenter: Razan Alshammari

20

By using a MySQL query, the system will display the registered user by their national ID. In case of the unregistered users, the medical specialist can fill their information and register them in the system.



## Rebuild Model Interface

The screenshot shows a 'Rebuild Model' interface. On the left, a sidebar lists diseases: Asthma, Glaucoma, Cervical Cancer, and Hypertension. The main area has a 'UPLOAD DATA FILE' button, a dropdown for 'Disease' set to 'Asthma', a 'Training Percentage %' input field, and a 'GENERATE MODEL' button. To the right, a 'MODEL SUMMARY' section displays: Model Name: 2024-05-05\_00:35:09.sav, Disease Type: Asthma, Total Number of Instances: 1000, Correctly Classified Instances: 980, Incorrectly Classified Instances: 20, and Training Percentage: 100%. A 'Update Diagnostic Model' button is at the bottom.

Only accessed by the admin.

1- upload the dataset file.

2- select the disease and insert the training percentage.

Presenter: Razan Alshammari

21

# Software Testing

Presenter: Shahad alghamdi

22

## Software Testing

- Test Item**  
Testing the functionality and the implementation of the website
- Approach**  
Different testing type were employed
- Features to be tested**  
All features to be tested were listed on the requirements
- Pass/Fail test**  
All functionality must work as expected
- Testing process**  
Including deliverables, responsibilities, resources, and timetable

Presenter: Shahad alghamdi

23



# CONCLUSION

Presenter :Fai Alanazi

24

## Conclusion

### Aim

Using ML approaches  
for detecting chronic  
diseases using clinical  
data.

### contribution

Improving Saudi Arabia's  
public health care.

### Fill the gap

employing Saudi  
clinical datasets that  
are available with  
minimum resources.

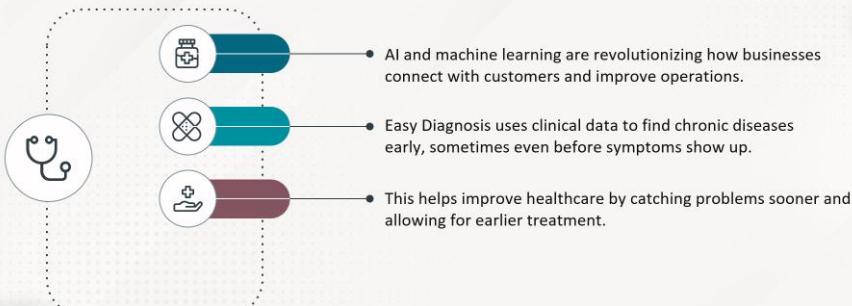
### System Employing

Each user has certain  
requirements, constraints.

Presenter : Fai Alanazi

25

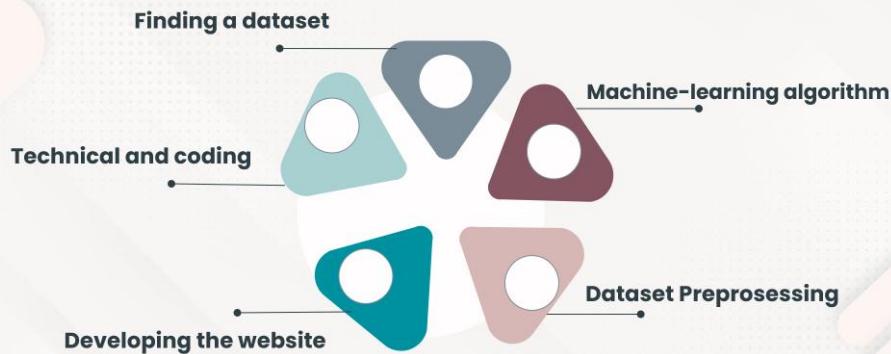
## Entrepreneurship Impact



Presenter : Fai Alanazi

26

## Issues Faced



Presenter : Fai Alanazi

27

# Demo



Presenter: Razan Alshammari

28

# Demo

Welcome to Easy Diagnosis

Watch the Video

29

Presenter: Rahaf Yaanallah

## Achievements

**Waiting for publishing**

**Under processing**

- 1. Osteoporosis Disease Journal Paper
- 2. Epileptic Seizure Disease Journal Paper
- 3. Sickle Cell Anemia Disease Journal Paper

**Accepted**

Of International Conference of Business and Innovative Technology (ICBIT) 2024 Team

- 1. Epileptic Seizure Disease Conference Paper
- 2. Sickle Cell Anemia Disease Conference Paper

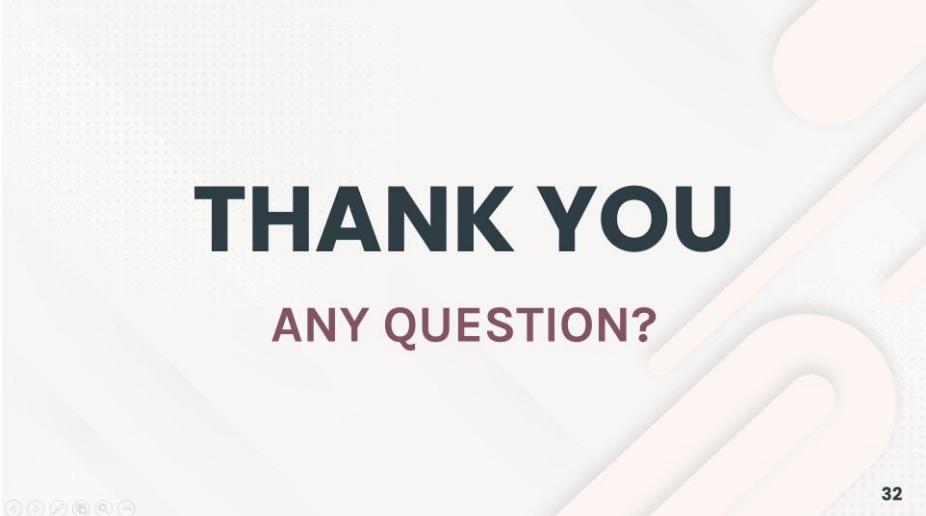
30

## Acknowledgment

First and foremost, we express our deepest gratitude to Allah, who most deserves our thanks and praises for His countless blessings. We pray that Allah blesses our efforts in this project and makes it beneficial for the community. We extend our heartfelt thanks to our supervisor, Mr. Aftab Khan, for his unwavering support and guidance throughout this endeavor. We are also immensely grateful to our Co-supervisor, Dr. Sunday Olatunji, for his encouragement and steadfast belief in our capabilities. Finally, we would like to express our appreciation for the continuous guidance, feedback, and advice provided by our evaluators, Dr. Mohammed Imran and Mrs. Mehwash Farooqui.

Presenter: Rahaf Yaanallah

31



# **THANK YOU**

**ANY QUESTION?**

© © © © © ©

32