DESIGN OF PREDICTIVE MODEL USING VARIOUS MACHINE LEARNING ALGORITHMS FOR IMPROVING DIAGNOSIS OF TYPE 2 DIABETES

Submitted in partial fulfillment of the requirements

of the degree of

Doctor of Philosophy (Technology)

In

Information Technology

By

Mrs. Rohini Anant Patil

(Registration No. 12/18-12-2019)

Under the Supervision of

Dr. Kamal Shah

Department of Information Technology

Thakur College of Engineering and Technology

Thakur Village, Kandivali (East),

Mumbai-400101

University of Mumbai

August 2022

STATEMENT BY THE CANDIDATE

As required by the University Ordinance 770, I wish to state that the work embodied in this thesis titled **"Design of predictive model using various machine learning algorithms for improving diagnosis of type 2 diabetes""** forms my contribution to the research work carried out under the guidance of Dr. Kamal Shah at the Thakur College of Engineering and Technology. This work has not been submitted to any other university. Whenever references have been made to previous works of others, it has been indicated as such and included in the Bibliography.

Mrs. Rohini Anant Patil
(Registration No. 12/18-12-2019)

Certified by

Dr. Kamal Shah
(Supervisor/Guide)

Date: 01/08/22

CERTIFICATE

This is to certify that the thesis entitled "**Design of Predictive Model Using Various Machine Learning Algorithms for Improving Diagnosis of Type 2 Diabetes"** is a bonafide work of **"Mrs. Rohini Anant Patil" (Ph.D. Registration No. 12/18-12-2019)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of "**Ph.D. (Technology)**" in **"Information Technology".**

Dr. Kamal Shah
Supervisor/Guide

Dr. R. R. Sedamkar
Head, Research Centre,
HOD, Ph.D. Program

Dr. B. K. Mishra
Principal

PRINCIPAL
Thakur College of Engg. & Tech.
A-Block, Thakur Educational Campus
Shyamnarayan Marg, Thakur Village
Kandivali (E), Mumbai-400 101

Thesis Approval for Ph.D.

This thesis entitled **Design of Predictive Model Using Various Machine Learning Algorithms for Improving Diagnosis of Type 2 Diabetes** by **Mrs. Rohini Anant Patil (Ph.D. Registration No. 12/18-12-2019)** is approved for the degree of **Ph.D. (Technology) in Information Technology.**

Examiners

1. ----------

2. ----------

Date:

Place:

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Mrs. Rohini Anant Patil
(Registration No. 12/18-12-2019)

Date: 01/08/2022

ABSTRACT

Diabetic mellitus (DM) is most the most prevalent non-communicable disease. The burden of diabetes is growing due to its increasing prevalence day by day. Preventive care in an early stage is essential to avoid complications. Early diagnosis and better care at less cost are the key aspects of improving patient satisfaction. Enormous data and increased complexities have led to rising interest in using machine learning (ML) in healthcare. ML uses several approaches to learn and improvise the prediction without explicit programming automatically. It develops on existing statistical methods and finds patterns in the data. Predictive modelling can make predictions based on specific properties from a training dataset. As per the International Diabetes Federation, type 2 diabetes prevalence is projected to reach 101 million by 2030. To reduce the adverse impact of rising rates of diabetes, a risk predictive model can be valuable for public health.

This research aimed to design a predictive model for estimating diabetes in healthy population with diverse ages based on different lifestyle factors, i.e. stress, food habit, smoking, profession and exercise.

The overall process of research encompassed the following four objectives; evaluation of the effect of stress on the prediction of type 2 diabetes, estimating the effect of stress related to COVID-19 on blood glucose level in type 2 diabetes, evaluation of the effect of lifestyle parameters on the prediction of type 2 diabetes using six machine learning models and designing a novel- Improved Ensemble Learning with Dimensionality Reduction Model (IELDR).

In order to study the results of the first objective, a dataset of 374 records was collected via a self-designed web-based questionnaire validated by a consulting physician. The collected data were subjected to feature selection methods, namely the Feature importance, Chi-square select-k best, Correlation, Recursive feature elimination (RFE) and Information gain. Then, using these six classifiers, Gradient boosting classifier (GBC), Random forest (RF), Logistic regression (LR), Support Vector Machine (SVM), Decision tree (DT), k-nearest neighbor (KNN). K fold cross validation accuracy was calculated. Hyperparameter tuning using the grid search method was used before estimating the model's accuracy. Comparative assessments based on Receiver operating

characteristics (ROC), accuracy, precision, sensitivity and f1-measure were performed among six machine learning models. SVM with the select k-best method showed an accuracy of 82.14%. Accuracy, precision and recall values of GBC and SVM were 86.67%, 87% and 87%, respectively.

As a part of the second objective, the impact of COVID-19 related stress on blood glucose levels in type-2 diabetes patients was examined using statistical methods. A dataset of 199 diabetes patients was collected from MGM Hospital, Aurangabad. To estimate the severity of stress, Depression Anxiety Stress Scale (DASS-12) stress sub-scale was used. Fasting blood glucose (FBG) and post-prandial blood glucose (PPBG) levels of these patients were measured before and during discharge and compared. Observation showed a rise in glucose levels with without and moderate stress discharged after COVID-19 treatment, implying an impact of disease related stress. Also, there was a rise in the blood glucose level in elderly patients.

As a part of the third research objective, using six machine learning models, the effect of lifestyle parameters on the prediction of type 2 diabetes was analyzed. The proposed model was implemented by a grouping of five feature selection techniques and six different classification methods. The highest feature selection accuracy was observed in the GBC with the RFE method of 82.92%. Based on the comparative analysis, LR, RF, and KNN showed the highest accuracy value of 85.33%.

Finally, we designed a novel- Improved Ensemble Learning with Dimensionality Reduction Model (IELDR). Auto encoder-based feature extraction method along with ensemble learning algorithm was used. The experiments were carried out using the LS_diabetes dataset. The IELDR model showed results in terms accuracy, precision, specificity, sensitivity, f1 score, roc and Mathew correlation coefficient (MCC) of 98.67%, 95.24%, 100%, 98.18%, 97.56%, 99.09% and 0.97 respectively. IELDR model applied on Diabetes_2019 dataset and PIMA diabetes dataset. Comparative analysis was done with the LS_diabetes dataset. This proposed IELDR system predicts the risk of type 2 diabetes in a healthy person based on the person's current lifestyle pattern; the system calculates the risk score and offers a recommendation.

Keywords: Machine learning, Improved Ensemble Learning with Dimensionality Reduction Model, diabetes, risk, lifestyle, stress, prediction

Table of Content

Abstract ……v

List of Tables xiii

List of Figures xv

Abbreviations and Symbols xviii

Chapter 1. Introduction 1

1.1 Background 2

1.1.1 Importance of Health 2

1.1.2 Lifestyle Disease 2

1.2 Diabetes Mellitus 3

1.2.1 Types of Diabetes Mellitus 3

1.2.2 Diagnosis of Type-2 Diabetes 4

1.2.3 Management of Type-2 Diabetes Mellitus 4

1.3 Machine Learning: An Overview 5

1.3.1 Requirements of Machine Learning 5

1.3.2 Types of Machine Learning 6

1.3.3 Strength and Challenges of Machine Learning 8

1.3.4 Importance of Machine Learning in Healthcare 9

1.3.5 Predictive Modelling Approaches 10

1.4 Motivation 11

1.5 Problem Statement 12

1.6 Objectives of the Research 13

1.7 Scope of the Research 14

1.8 Research Contributions 14

1.9 Organization of the Report 16

1.10 Summary 17

Chapter 2. Theoretical Background 18

2.1 Introduction 19

2.2 Feature Selection Approaches.....19

2.2.1 Filter Method.....20

2.2.2 Wrapper Method 21

2.2.3 Embedded Method.....21

2.2.4 Autoencoder based Method.....22

2.3 Conventional Machine Learning Classifiers.....24

2.4 Model Selection using hyperparameter tunning 30

2.5 Model Validation 32

2.6 Summary.....33

Chapter 3. Literature Review.....34

3.1 Introduction.....35

3.2 Survey on Machine Learning Algorithm for Type-2 Diabetes.....35

3.3 Survey on Feature Selection Techniques.....40

3.4 Survey on Type-2 Diabetes Prediction Using Machine Learning 45

3.4.1 Type-2 Diabetes Prediction for Stress Parameter.....45

3.4.2 Type-2 Diabetes Prediction for Lifestyle Parameter.....50

3.5 Survey on Impact of COVID-19 Related Stress on Blood Glucose Level for Type-2 Diabetes patients 52

3.6 Research Gap and Findings 55

3.7 Summary.....56

Chapter 4. Proposed Framework for Analyzing Impact of Various Parameters on Type 2 Diabetes Prediction 57

4.1 Introduction.....58

4.2 Data Collection Process 59

4.2.1 Development of Machine Learning Model Using LS_Diabetes Dataset 59

4.2.2 Development of Model Using COVI_Stress Dataset 61

4.2.3 Development of Machine Learning Model to Test Using PIMA Indian Diabetes Dataset. 63

4.2.4 Development of Machine Learning Model to Test Using Diabetes_2019 Dataset……. 64

4.3 Proposed Research Framework.....66

4.4 Performance Metrics.....70

4.5 Impact of Stress Parameter on Type-2 Diabetes Prediction 72

4.5.1 Description of Stress Data.....72

4.5.2 Data Pre-processing 73

4.5.3 Optimal Feature Selection Using Preferred Techniques 73

4.5.4 Diabetes Prediction for Stress Parameter Using Preferred Classifier 74

4.6 Impact of COVID-19 Related Stress on Blood Glucose Level for Type-2 Diabetes Patients 75

4.6.1 Proposed COVI_Stress Methodology 76

4.6.2 Data Cleaning.....77

4.6.3 Preferred Statistical Methods 77

4.7 Impact of Lifestyle Parameters on Type-2 Diabetes Prediction 78

4.7.1 Dataset Description 78

4.7.2 Data Pre-processing 79

4.7.3 Optimal Feature Subset Selection Using Preferred Techniques 80

4.7.4 Diabetes Prediction Using Preferred Classifier.....80

4.8 Algorithm for D_Stress and Diabe_lifestyle methodology 82

4.9 Mathematical Model for D_Stress and Diabe_lifestyle 83

4.10 Summary 84

Chapter 5. Design of Novel Model for Type 2 Diabetes Prediction- "Improved Ensemble Learning with Dimensionality Reduction" (IELDR) 85

5.1 Introduction 86

5.2 Type-2 Diabetes Prediction System Using Proposed IELDR.....86

5.2.1 Dataset Description 86

5.2.2 Data Pre-Processing 88

5.2.3 Autoencoder based Optimal Feature Extraction 88

5.2.4 Diabetes Prediction Using IELDR 89

5.3 Mathematical Model of IELDR 90

5.4 Summary 92

Chapter 6. Result and Performance Analysis 94

6.1 Introduction 95

6.2 Experimental Setup 95

6.3 Performance Analysis of Type-2 Diabetes Prediction for Stress Parameters .96

6.3.1 Analysis of Feature Selection Methods 98

6.3.2 Experimentation Results.....101

6.4 Statistical Analysis for Impact of COVID-19 Related Stress on Blood Glucose Level for Type-2 Diabetes Patients 106

6.4.1 Data Analysis 106

6.4.2 Intra and Inter Group Analysis 108

6.4.3 Groupwise Analysis 111

6.4.4 Age wise Analysis 112

6.5 Performance Analysis of Type-2 Diabetes Prediction for Lifestyle Related Parameters for LS_diabetes dataset 112

6.5.1 Analysis of Feature Selection Methods 114

6.5.2 Experimentation Results.....117

6.6 Comparative Analysis of Type-2 Diabetes Prediction for LS_diabetes Dataset and Diabetes_2019 Dataset for Lifestyle parameters 122

6.6.1 Experimental Validation.....122

6.7 Performance Analysis of Type-2 Diabetes Prediction Using IELDR Model for LS_Diabetes Dataset 124

6.7.1 Experimentation Result 124

6.7.2 Recommendations 127

6.8 Comparative Analysis of Type-2 Diabetes Prediction Using IELDR Model with PIMA Diabetes Dataset 127

6.8.1 Experimentatal Validation.....127

6.8.2 Comparative Analysis 129

6.9 Comparative Analysis of Type-2 Diabetes Prediction Using IELDR Model with Diabetes_2019 Dataset 130

6.9.1 Experimentatal Validation.....130

6.9.2 Comparative Analysis 132

6.10 Summary 133

Chapter 7. Conclusion and Future Scope.....134

7.1 Conclusion 135

7.2 Future Scope 137

References……. 138

List of Publications and Patent 149

Acknowledgement.....152

Appendix I Dataset Validation Letter 154

Appendix II Ethical Approval 156

Appendix III Dataset Screenshots 158

Appendix IV Plagarism Report.....164

Appendix V Synopsis Report and Compliances 166

List of Tables

Table No	Table Content	Page no
Table 1.1	Ranges of laboratory parameters for different levels of diagnosis	4
Table 4.1	LS_diabetes dataset characteristics	60
Table 4.2	Class wise LS_diabetes dataset distribution	60
Table 4.3	Characteristics of COVI_Stress dataset	62
Table 4.4	Attribute description of COVI_Stress dataset	62
Table 4.5	PIMA Indian diabetes dataset characteristics	64
Table 4.6	Attribute description of PIMA Indian diabetes dataset	64
Table 4.7	Characteristics of Diabetes_2019 dataset	65
Table 4.8	Attribute specification of Diabetes_2019 dataset	65
Table 4.9	Confusion matrix	70
Table 4.10	D_Stress methodology dataset description	73
Table 4.11	Depression, anxiety & stress scale (DASS)	77
Table 4.12	Diabe_lifestyle methodology dataset description	79
Table 4.13	Procedure for D_Stress and Diabe_lifestyle methodology	83
Table 5.1	LS_Diabetes dataset attribute description	87
Table 6.1	Training and testing accuracy of six classifiers	96
Table 6.2	Analysis of feature selection methods based on number of features	98
Table 6.3	Comparative analysis of different feature selection methods	100
Table 6.4	Hyperparameter tuning classification accuracy	101
Table 6.5	Selected hyperparameter based classifier performance	103
Table 6.6	Comparative analysis of different classifiers on various measures	105
Table 6.7	Experimental validation	106
Table 6.8	Demographics characteristics of participants	107
Table 6.9	Blood glucose parameters comparison	108
Table 6.10	Intragroup comparison of glycaemic parameters based on the severity of stress	109
Table 6.11	P values of intergroup comparison	111
Table 6.12	Group wise comparison	111

Table 6.13	Age group wise p values comparison	112
Table 6.14	Comparison of training and testing accuracy	113
Table 6.15	Classification accuracy based on number of selected features	115
Table 6.16	Classification accuracy of hyperparameter tuning	117
Table 6.17	Classifier based best hyperparameters	120
Table 6.18	Performance metric of classifiers	121
Table 6.19	Comparison of existing methodology vs proposed methodology	123
Table 6.20	Accuracy and error rate comparison vs. model	124
Table 6.21	Comparison of IELDR model with different classifier models on various measure	125
Table 6.22	Comparative analysis of PIMA dataset vs. LS_diabetes dataset	128
Table 6.23	Proposed IELDR model vs. best model on PIMA dataset	129
Table 6.24	Comparative analysis of LS_diabetes dataset vs. Diabetes_2019 dataset	130
Table 6.25	Proposed IELDR methodology vs. methodology used on Diabetes_2019 dataset	132

List of Figures

Figure No.	Figure Content	Page no
Figure 1.1	Flow of machine learning process	5
Figure 1.2	Types of machine learning algorithm	7
Figure 1.3	Global projection	11
Figure 1.4	Rising number of diabetes cases in India	11
Figure 1.5	Type-2 diabetes risk element	12
Figure 2.1	Types of feature selection techniques	20
Figure 2.2	Process of filter method	20
Figure 2.3	Process of wrapper method	21
Figure 2.4	Process of the embedded method	22
Figure 2.5	Hidden layer based autoencoder structure	22
Figure 2.6	Autoencoder structure	23
Figure 2.7	Graphical representation of logistic regression	25
Figure 2.8	Decision tree with n=2 nodes	26
Figure 2.9	Random forest model	27
Figure 2.10	Working principle of support vector machine	27
Figure 2.11	Basic neural network with two hidden layers	29
Figure 2.12	Process of hyperparameter tuning	31
Figure 2.13	Hyperparameter using grid search and random search	32
Figure 2.14	Detailed representation of data using k-fold validation	33
Figure 4.1	LS_diabetes dataset parameters	61
Figure 4.2	Flow of model development	66
Figure 4.3	Research framework	67
Figure 4.4	Proposed research methodologies	69
Figure 4.5	ROC curve	72
Figure 4.6	A proposed methodology for D_Stress	74
Figure 4.7	COVI- Stress methodology	76
Figure 4.8	Proposed Diabe_lifestyle framework	81

Figure 5.1	Proposed Improved Ensemble Learning with Dimensionality Reduction Model	90
Figure 6.1	Comparison of train_test classification accuracy	97
Figure 6.2	Comparative analysis of k-fold accuracy	97
Figure 6.3	Analysis with feature selection and without feature selection method	100
Figure 6.4	Accuracy based comparative analysis of feature selection and hyperparameter tuning	101
Figure 6.5	Accuracy comparison of classification algorithm	102
Figure 6.6	ROC curve analysis	104
Figure 6.7	Performance evaluation with D_Stress	105
Figure 6.8	Distribution based on severity of stress as assessed by DASS-12 score	108
Figure 6.9	Intra-group analysis	110
Figure 6.10	Training and testing classification accuracy	113
Figure 6.11	Analysis of k-fold accuracy	114
Figure 6.12	Comparative analysis of different feature selection methods	116
Figure 6.13	Comparison with and without feature selection accuracy	117
Figure 6.14	Comparison of feature selection and hyperparameter tuning accuracy	118
Figure 6.15	Comparison of K-fold, feature selection and hyperparameter tuning	119
Figure 6.16	Comparison of ROC curve analysis	120
Figure 6.17	Graphical analysis of various measures	122
Figure 6.18	Accuracy comparison of proposed Diabe_lifestyle method with existing method	123
Figure 6.19	Accuracy comparison of top five model	124
Figure 6.20	Comparison of precision vs. recall	126
Figure 6.21	Comparative analysis of f1-score for various classifiers	126
Figure 6.22	Accuracy based comparison with PIMA dataset	128
Figure 6.23	Proposed IELDR model vs best model for case-I	129
Figure 6.24	Accuracy based comparison with Diabes_2019 dataset	131

Figure 6.25 Performance comparison of IELDR method on LS_diabetes dataset vs.Diabetes_2019 dataset 131
Figure 6.26 Proposed IELDR model vs best model for case-II 132

Abbreviations and Symbols

BMI	Body Mass Index
DM	Diabetes Mellitus
FBG	Fasting Blood Gglucose
PPG	Postprandial Blood Glucose
HbA1c	Hemoglobin A1c
T2D	Type 2 Diabetes
IDF	International Diabetes Federation
WHO	World Health Organization
ML	Machine Learning
AI	Artificial Intelligence
NPV	Negative Predictive Value
PPV	Positive Predictive Value
ROC	Receiver Operating Characteristics Curve
AUC	Area Under Curve
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
SP	Specificity
SVM	Support Vector Machine
RF	Random Forest

GBC	Gradient Boosting Classifier
LR	Logistic Regression
KNN	K-Nearest Neighbor
DT	Decision Tree
MLP	Multilayer Perceptron
NB	Naïve Bayes
LDA	Linear Discriminant Analysis
SGD	Stochastic Gradient Descent
SARS-COV-2	Severe Acute Respiratory Syndrome Ccoronavirus 2
DASS	Depression Anxiety Stress Scale
COVID-19	Corona Virus Disease
RFE	Recursive Feature Elimination
UCI	University of California at Irvine
PIDD	Pima Indian Diabetic Database

Symbols

x*	Set of Hyperparameters
Σ	Summation
w'	weight
b'	bias
X_i	Input Variable
Y_i	Output Variable
$\mathcal{F}$	Feature

CHAPTER-1

INTRODUCTION

CHAPTER-1 INTRODUCTION

1.1 Background

1.1.1. Importance of Health

Health is more than just the absence of illness, injury, or pain. It is a person's overall state of wellness on all levels. As per World Health Organization, health is "a state of complete physical, mental, and social well-being and not merely the absence of disease" [1]. Health is derived from the words "hoelth" (sound) and "hale" (strength). A person is considered healthy if he/she can function effectively, adapt to environmental changes and feel well. As per the modern approach to science, all illnesses are potentially affected by different factors in the biological ecosystem, psychological status and social surroundings.

Physical health refers to physical fitness, which is the outcome of exercise, proper diet and nutrition, cleaning and adequate rest. Psychological health is "a state of well-being in which the individual realizes his or her own potential, can cope with the normal stresses of life and can work productively and fruitfully" [1]. The four determinants of health are environment, human biology, healthcare services and lifestyle. Health is maintained and improved not only through advancements in health care but also through societal and individual lifestyle choices.

1.1.2. Lifestyle Diseases

Lifestyle diseases are driven by several factors, including the globalization of unhealthy lifestyles, rapid urbanization and population aging. Such disorders are preventable, and their occurrence can be reduced through dietary and lifestyle changes. Examples include type 2 diabetes mellitus, depression, asthma, stroke, acne, metabolic syndrome, heart disease, osteoporosis, and obesity. Lifestyle diseases are brought on by improper interaction among people and their surroundings. It could take years for them to mature and once encountered, they are difficult to cure [1, 2]. Due to modern methods of living, poor eating habits, sleep deprivation, stress and a sedentary lifestyle are significantly contributing to the progression of lifestyle diseases, contributing a significant burden on healthcare.

1.2 Diabetes Mellitus

Diabetes is the most challenging disease to manage from a psychosocial and behavioural standpoint. A published WHO report in 2019 states that diabetes is the 9th cause of death [1]. Diabetes mellitus occurs due to an increased level of blood glucose as the body cannot use insulin effectively or produce an adequate amount of insulin. The cell's inability to respond to insulin or the lack of insulin leads to increased blood glucose levels, called hyperglycemia, which is the benchmark of diabetes. Over a long period if hyperglycemia, left unchecked, can source damage to the body organs, leading to health complications, i.e., problems related to the brain, eye, heart and kidney [2]. Diabetes mellitus is associated with mortality and morbidity due to short and long-term complications such as stroke, kidney failure, vision loss, heart attack and damage to nerves, physical and cognitive disability and depression.

1.2.1 Types of Diabetes Mellitus

Diabetes mellitus is classified as; type 1 diabetes, type 2 diabetes and gestational diabetes.

Type 1: In this case, the body does not properly metabolize food due to insufficient or lack of insulin resulting in a high level of blood glucose. This happens because the pancreas cannot produce sufficient insulin. Therefore insulin injections are required for such patients. [1, 2].

Type 2: Due to a lack of effectiveness of insulin or insufficient insulin, type 2 diabetes occurs [2]. By adhering to a healthy lifestyle, it can be prevented.

Gestational: Diabetes occuring during pregnancy is gestational diabetes [2]. During pregnancy, gestational diabetes is diagnosed due to an increased level of blood glucose. After pregnancy, it becomes normal, but there are chances of developing type 2 diabetes.

Pre-diabetes is a state of rise in blood glucose [2]. It raises the risk of developing diabetes in the future. At this earlier stage, preventive measures are necessary to prolong the occurrence of diabetes mellitus.

1.2.2 Diagnosis of Type 2 Diabetes

Early diagnosis and interventions are required for living well with diabetes [2]. Different laboratory investigations in diabetes include glycosylated haemoglobin (HbA1C), FBG and PPG. Ranges of these parameters for diabetes, pre-diabetes and normal people are shown in (Table 1.1).

Table1.1: Ranges of laboratory parameters for different levels of diagnosis

Status	Fasting blood glucose level(mg/dl)	Blood glucose level after oral glucose tolerance test (mg/dl)	HbA1c
Diabetes	≥126	≥200	≥6.5%
Pre-diabetes	100 to125	140 to 199	5.7 to 6.4%
Normal	≤99	≤139	~5%

1.2.3 Management of Type-2 Diabetes Mellitus

Reducing mortality rates and improving quality of life are the primary goals of managing chronic diseases. Effective diabetes management includes early detection, short- and long-term morbidity prevention, and self-care practices promotion [3].

In order to reduce the incidence of disease, preventive measures such as avoiding smoking and being overweight, regular physical activity, consumption of healthy types of fat, plenty of fruits and vegetables, replacing refined grains with whole grains, avoiding excessive calories, etc., are necessary. Physical activity or exercise increases metabolic control, insulin sensitivity, and cardiorespiratory fitness and helps to maintain body weight.

1.3 Machine Learning: An Overview

Machine learning (ML) concepts originated from pattern recognition have the ability to discover without programming to perform specific tasks. Samuel invented "ML" in the early1950s [4]. In ML, "a computer program is said to learn from experience "E" with respect to some class of tasks "T" and performance measure "P", if its performance at tasks in "T", as measured by "P", improves with experience "E"" [5]. Large, complex data sets can be analysed using ML to find patterns and create hypotheses by learning from previous examples [6].

1.3.1 Requirements of Machine Learning

ML uses several approaches to automatically learn and improvise the prediction without explicit programming. Through ML, computers learn to act without specific programming and models help computers to sort out required analysis from the extensive data. Algorithms are used to discover particular patterns to get the necessary insights from the data. These patterns help in generating hypotheses [7]. The flow of the ML process is depicted in figure (1.1).

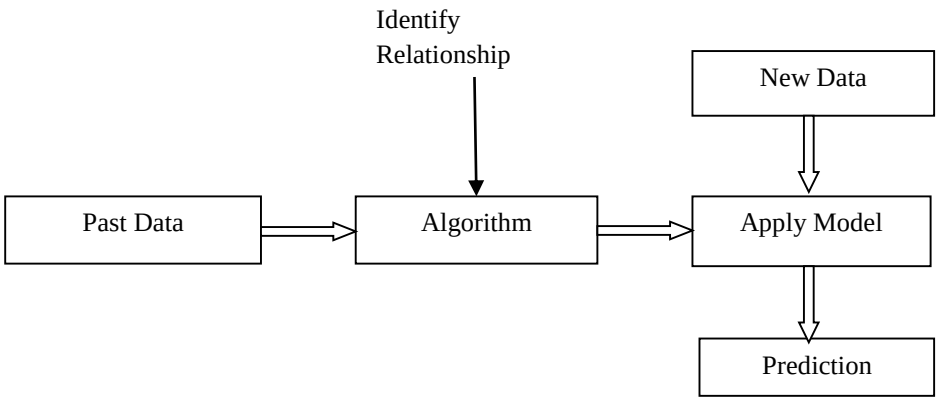


Figure 1.1: Flow of machine learning process

The intention of using AI and ML in healthcare is to increase the diagnostic accuracy and effectiveness of therapy and help clinicians in their practice of patient management with improved outcomes. The decision to use AI in the diagnostic field is quickly gaining momentum due to improvement in accuracy and huge amount of data availability [8].

Early diagnosis and better care at a reasonable cost are key aspects of improving patient satisfaction. Enormous data and increased complexities have led to rising

interest in using ML in healthcare. Disease prediction using ML is gaining significant attention in data science.

In ML, data plays an indispensable role. Data collection, a critical step, is associated with several challenges, including continuous growth in the amount of data, costs related to data storage cost, different types of data, complex data, heterogeneity of data, and diverse sources for data collection. Even though the ML domain has performed massive tasks, manpower can-not be entirely replaced by machines.

Datasets can be categorized as labeled and unlabeled datasets.

Labeled Dataset D_m: $X = \{ x^{(k)} \in R^d \}_{k=1}^{N}$

$Y = \{ y^{(k)} \in R \}$

Unlabeled Dataset D_m: $X = \{ x^{(k)} \in R^d \}_{k=1}^{N}$

Where X denotes the feature vector with N samples.

The feature vector with p dimensions are, $x(k) = [x_1(k), x_2(k), \ldots \ldots xp(k)]^T$ where each dimension of vector represents an attribute or feature. Y represents the dependent variable for the label set.

1.3.2 Types of Machine Learning

ML is used to optimize system parameters and get the expected outcome using training examples. Statistical techniques are often used to make computers able to "learn." Once computer learning is done, the resulting system is used for different applications, including automatic prediction of the type of data, feature extraction, or generating artificial examples. Supervised, unsupervised, reinforcement learning and semi-supervised are the different types of ML algorithms. These algorithms work with specific hypotheses and mathematical models for data. figure (1.2) shows types of ML.

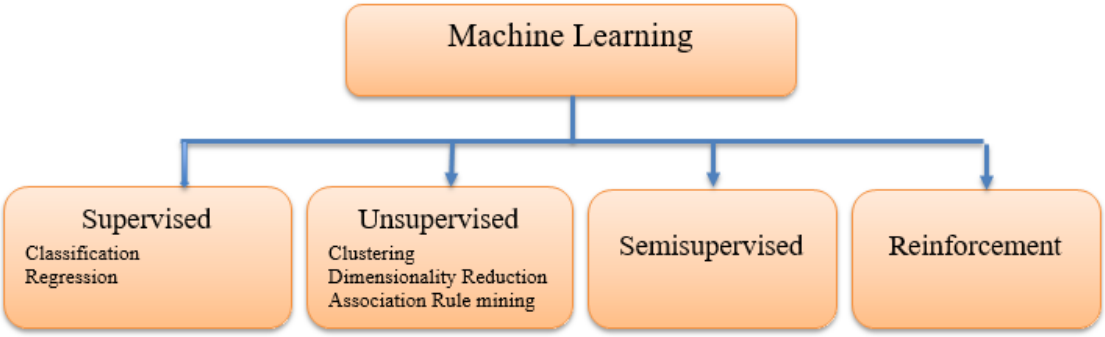


Figure 1.2: Types of machine learning algorithm

Supervised learning:

In this learning, the training consists of input vectors and target vectors. For each instance target vector is represented as y .The model is in the form [(x, y) € n] where n = 1.. N. The multiple feature vector is x1, x2,…xk, where x is input. The model is represented as f(x) = cx + d, where f(x) is the predictor variable, c and d are constant variables, and the noise in the data. For each input vector, loss functions are the differences between the output and target vector. This procedure is repeated until the system reaches to optimized value for c and d. The machine learns through experience and becomes ready for use [9-11]. Primary categories under supervised learning consist of classification and regression.

Classification:

If the dependent variable is categorical, the problems are called classification. Classification algorithms use features based on predefined criteria to classify data in classes, such as "diabetes" or "no diabetes." From classification algorithms, the machine learns to classify data items. Classification problems require labeled data for the identification of patterns.

Example: Independent variable consists of many features. For instance, if the input feature is "Fasting Blood Glucose (FBS)," which may read as:

. {x_training, output}→{"FBS, 140", "Diabetes"}

Regression:

If the dependent variable is continuous, the problems are known as regression. In the regression problem based on the pattern in the information, the machine predicts continuous variables. For example, if the continuous value is HbA1c =7.1%. Three groups can be defined based on the level of HbA1c.

- Diabetes (Group 1)→ HbA1c ≥ 6.5%
- Pre-diabetes (Group 2)→HbA1c = 5.7% to 6.4%
- Normal (Group 3)→HbA1c < 5.7%.

Unsupervised ML algorithms:

A learning model is trained on unlabeled data to predict outcomes. It is pattern based learning. This algorithm is used for instances where experts are unable to analyze data. This technique is used in descriptive modeling and pattern detection.

Semi-supervised ML algorithms:

Supervised and unsupervised learning lead to semi-supervised learning. As expert's knowledge is required for labeled data; increases cost. Therefore few models have been developed with label data and others work on unlabeled data. This method is used when there is only a limited dataset.

Reinforcement ML algorithms:

Reinforcement learning is learning behavior in a specific condition for a certain reward/ penalty signal. A current state and environment are defined in this type of learning. In order to determine the appropriate reward or punishment action, a criterion function evaluates the current state using a set of policies. It is self-sustainable and upgrades on its own. Reward can be positive and negative. The Monte Carlo method and Q-learning are popular reinforcement learning algorithms.

1.3.3 Strength and Challenges of Machine Learning

The strength and weakness of ML for employing in various real-time applications are given below [12].

- Simple and relatively fast algorithm doesn't require a lot of memory.
- To discover implicit relationships in data-sets.
- Capabilities to handle multi-dimensional and multi-variety data in uncertain environments.

- Capability to handle complex optimization problems

Machine Learning Challenges:

Although ML has been used widely, it has some challenges, as discussed below [8,13].

- There is no uniformity in the data collection process and data collection tools used by healthcare professionals. This results in difficulty in streamlining data and infrastructures.
- Limited availability of sufficient resources, including knowledgeable and trained manpower, technology, and financial aid.
- Data standardization and optimization for effective utilization in the improvement of existing knowledge.
- Regular update of the datasets.
- Data quality
- Issues related to data privacy and data integrity.
- Differences in results of prediction by different algorithms result in a biased diagnosis.
- High dimensional data increases model complexity and learning rate, which leads to overfitting.
- Although it can significantly help in automation, it cannot completely replace all tasks.

1.3.4 Importance of Machine Learning in Healthcare

With a rise in awareness and understanding of diseases, an increase in people's earnings, improved access to healthcare, and access to technology such as the internet and mobile phones, there has been significant growth in the disease-associated information in the medicinal field. Data collection from the patient, physician, hospital records, and insurance agencies has become more accessible today because of technological advances. The growth of healthcare coupled with technology has fueled the expanded need for ML in healthcare. As expected, ML helps in removing important information from enormous available data. ML is helping healthcare by providing innovative and relevant information, which is practically difficult to analyze manually given the constraints of time, manpower, and other resources [14].

Accuracy of diagnostic rates and quick results are the well-known benefits of AI and ML. Today, increasing costs is one of the crucial challenges in healthcare. The government and others involved in healthcare always seek cost-effective options. The solutions provided for healthcare should be cost-effective without compromising patient access and healthcare care delivery, resulting in better health outcomes. In this regard, AI and ML have all the required potential to fulfill such demands. The broad and diverse utility of AI in the medical field underlines its long presence over four decades and applications and research is expected to increase further over a while. Considering accuracy and cost efficiency, AI offers the advantage of replication.

The use of ML can help in healthcare in the following ways:

1] Prediction of onset and progression of chronic diseases

2] Reduction in mortality

3] To reduce readmissions in hospitals

4] For prevention of hospital acquired infections

5] Prediction of patient propensity to pay

1.3.5 Predictive Modelling Approaches

Today, valuable data extraction for diagnosing and treating diseases is increasingly becoming necessary. ML is used for building analytical models. Using ML, computers can find insights and patterns that would be missed by the human eye without being programmed to do so. In predictive analysis, certain variables are used to measure future behavior. This idea can be used in medicine to help diagnose and predict disease outcomes. Due to outstanding performance on large and noisy data, ML is popular for predictive analytics. Predictive analysis can be applied with several techniques, including data mining, statistics, modelling, ML, and AI. Due to exceptional performance on large datasets with their uniform characteristics and noisy data, ML techniques have steadily gained popularity in predictive analyti cs.

1.4 Motivation

Diabetes is affecting the health of several people across the world, turns to cause of death and disability. Moreover, the rising prevalence and the financial, emotional, and physical burden add to the difficulties for the individuals living with diabetes. Diabetes is one of the largest healthcare worldwide problems.

In 2019, it was anticipated that there will be 9.3 % of diabetes worldwide, by 2045 increasing to 10.2% to 10.9% [2]. Global Report published in 2017[2] of International Diabetes Federation claims that there are 82 million adults who have diabetes. South East Asia is the second highest region among all other areas. Overall global projections of diabetes mellitus are shown in figure (1.3).

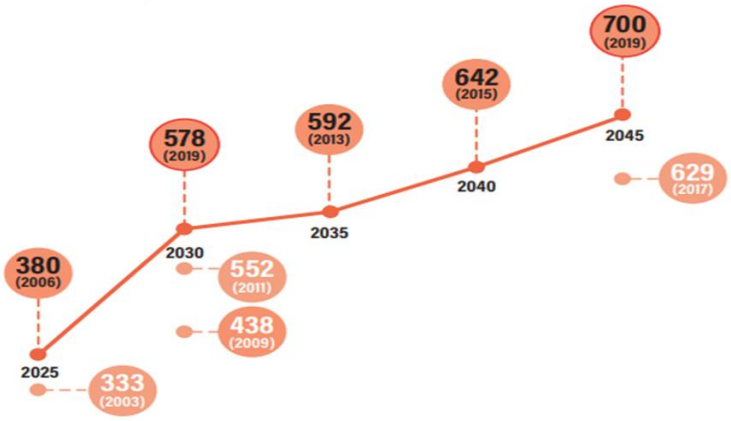


Figure 1.3: Global projection [2]

India contributes about 49% of the world's burden. In the Southeast Asia region, out of 88 million people with diabetes, India contributes 77 million people, which is expected to increase to 134.2 million in 2045, as shown in figure (1.4)[2].

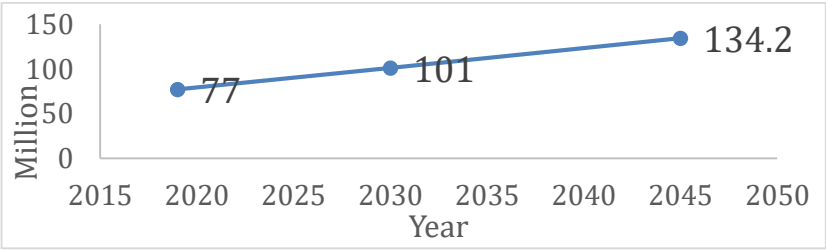


Figure 1.4: Rising number of diabetes cases in India [2]

In India, Diabetes prevalence increased upto 65 million in 2016. [15].In Maharashtra, overall reported diabetes prevalence in cities and rural area is 10.9% and 6.5%, respectively [16-17].

1.5 Problem Statement

Diabetes is not just high in numbers but also the leading cause of death worldwide. The disease can adversely affect the eye, nerves, arteries, kidneys and heart [2]. Every person is concerned about their health. Unlike other chronic diseases, self-care plays a significant part in managing diabetes. The risk elements include age, unhealthy eating habits, family history, obesity, exercise, smoking and stress related parameters figure (1.5).

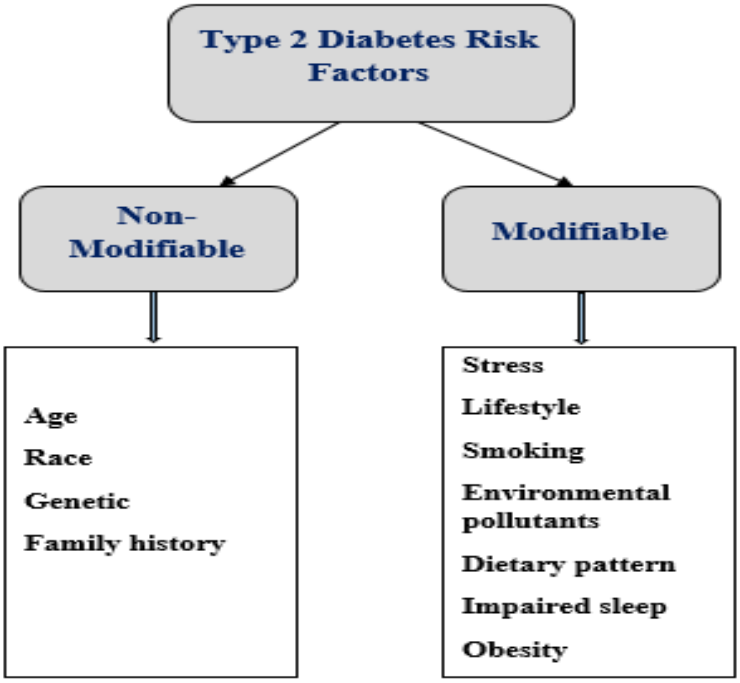


Figure 1.5: Type 2 diabetes risk element

The high prevalence and associated complications of diabetes underline the importance of early detection of the disease and the need for prompt measures to control it. Early detection is possible by regular screening of glucose level and glycosylated haemoglobin levels. However, it requires high motivation of the person for screening at regular intervals. Considering this background, there is an unmet need for a cost-efficient, convenient and accurate ML algorithm that can assist in the early prediction.

The problem statement of this research was “To design stress and lifestyle based type 2 diabetes prediction model using machine learning algorithms for improving diagnostic accuracy”.

Definitions of Key Terms

Stress:

Stress is an emotional strain caused by fatigue and emotional tension. Too many pathological conditions and hormonal imbalances are exacerbated by stress. Stress can also have a negative impact on glycaemic control. Stress has significant effects on metabolic activity. Stress is placing a substantial burden on healthcare due to modern lifestyles. Lifestyle diseases have increased due to inadequate sleep, unhealthy eating, and sedentary lifestyles. [18].

Lifestyle:

The term "lifestyle measure" refers to the individual or group's habits, attitudes, profession, income level, etc., collectively make up their way of life.

Predictive model:

A predictive model refers to predicting future outcomes by analyzing historical and current data using statistical and machine learning algorithms.

1.6 Objectives of the Research

Lifestyle changes are a crucial component of treatment for postponing or avoiding diabetes related complications. Hence, the research aims to design a predictive model for early stage diagnosis of type 2 diabetes based on different risk factors, i.e., stress, unhealthy food, smoking, profession and exercise. The research work is designed to analyze the influence of stress and lifestyle and examine the utility of ML models for the risk of development of diabetes based on different stress and lifestyle related features.

To summarize, the research objectives were

- To analyze the effect of stress related parameters on the prediction of type 2 diabetes.
- To analyze the impact of COVID-19 related stress on blood glucose in patients with diabetes of 3 different age groups.

- To analyze the effect of lifestyle factors on the prediction of type 2 diabetes.
- To design a novel predictive algorithm that can predict the risk of type-2 diabetes mellitus with up to 80% accuracy and provide a recommendation based on prediction.

1.7 Scope of the Research

Considering the high burden of diabetes and the risk of associated complications, there is a need to develop a diabetes prediction model in India. Using a prediction model may help in early diagnosis and intervention, reducing the healthcare and economic burden associated with the disease. The literature has used a standardized PIMA Indian diabetes dataset, which is female-centric [19]. There is a need for a study using a comprehensive dataset consisting of parameters from both genders. The limited literature from India shows the effect of stress using machine learning. The available literature does not show lifestyle parameters such as food habits at ground level.

1.8 Research Contributions

This research determined the effects of various factors on likelihood of developing type 2 diabetes. The contribution made by the work is described as given,

Contribution to Research Objective I

1. The proposed D_stress system worked on the LS_diabetes dataset with various stress parameters and demographical features.
2. D_stress system is based on five feature selection methods with six different classifiers.
3. The experimental result shows that stress contributes in the developing of type -2 diabetes with an accuracy 86.67% for Gradient Boosting Classifier and Support Vector Machine.

Contribution to Research Objective II

1. The proposed system worked on the COVI-Stress dataset. The study explored the effect of perceived stress on blood glucose levels in diabetes patients discharged from the hospital after their COVID-19 treatment.

2. The stress scale is used for the research.
3. A statistical analysis revealed an increase in FBG and PPBG levels in type 2 diabetes patients discharged following COVID-19 treatment. The increase was significant in no stress and moderate stress.
4. Compared age wise among three groups, gender wise as well as residence wise. The statistical analysis results revealed a significant difference in the FBG and PPBG based on their gender, age and place of residence (urban versus rural population). Among three different age groups in the old age group population found stress.

Contribution to Research Objective III

1. Evaluation of traditional machine learning models for predicting diabetes risk by utilizing a comprehensive LS_diabetes dataset with various lifestyle related parameters such as food habits and demographical features.
2. The dominant features that could cause complications with diabetes were identified with the help of five methods for selecting features.
3. The performance analysis of six classifiers: Decision Tree, K-Nearest Neighbor, Random Forest, Support Vector Machine, Logistic Regression, and Gradient Boosting Classifier.
4. Improved accuracy with minimum validation loss was achieved using tuning of hyperparameters using grid search.
5. Experimental result prediction model shows that a person's lifestyle is a risk element for developing type 2 diabetes with roc-auc value 92.8% for the Random Forest Classifier.
6. For validation Proposed Diabe_lifestyle model has been compared to cutting-edge techniques.

Contribution to Research Objective IV

1. Implementation and assessment of the proposed Improved Ensemble Learning with Dimensionality Reduction Algorithm (IELDR) predict the possibility of developing type 2 diabetes by utilizing a comprehensive India based LS_diabetes dataset.

2. Analyzing and forecasting type 2 diabetes using autoencoder-based feature extraction with two level stacking model as IELDR.
3. The IELDR model predicts the risk of development of diabetes by achieving an accuracy of 98.67% and provides recommendations.
4. The effectiveness of the developed models based on prediction accuracy has been validated through a thorough comparison against cutting-edge methods.

1.9 Organization of the Report

The research work is divided into seven chapters which are outlined as follows:

Chapter 1 deals with the importance of health, lifestyle diseases, diabetes mellitus causes, types, and effective management of diabetes. A brief overview of machine learning, its process, its types, strength and challenges, and the importance of machine learning in healthcare are discussed. Motivation, problem statement, objectives of this research, and scope and contribution work are presented.

Chapter 2 emphasizes the theoretical concept of conventional feature selection methods, auto encoder-based feature selection, and various machine learning classifiers. This chapter also describes hyperparameters based model selection followed by the model validation method, which is helpful for generating better predictions in terms of accuracy.

Chapter 3 studies the literature review associated with the diabetes mellitus prediction using ML classifiers include, k-nearest neighbour, random forest, logistic regression, support vector machine, decision tree and gradient boosting classifier. A review was done on various feature selection techniques, diabetes prediction using stress and lifestyle as risk factors and the impact of COVID-19 related stress to identify the research gaps for this research.

Chapter 4 describes the various datasets used for this research and the tool used to obtain the results. Provide insights about the proposed D_Stress model and Diabe_lifestyle model for type-2 diabetes prediction. Both models are implemented by using preferred classifiers and optimal feature selection techniques. Hyperparameter optimization using grid search shows the best output solution with higher accuracy for both models. This chapter also uses a statistical method to

describe the impact of COVID-19 related stress on blood glucose levels for type-2 diabetes patients.

Chapter 5 describes the novel hybrid model, “Improved Ensemble Learning with Dimensionality Reduction (IELDR).” The proposed model is implemented by combining autoencoder based feature extraction with an ensemble model.

Chapter 6 presents the type 2 diabetes analysis and prediction using D_stress and Diabe_lifestyle methodologies and a comparison of the suggested Diabe_lifestyle method with contemporary researchers. The proposed IELDR model has been evaluated and its performance is compared with the PIMA dataset and the diabetes_2019 dataset. Comparative analysis with contemporary researchers is also done. Based on the proposed IELDR model, provide a recommendation.

Chapter 7 reveals the conclusion of the thesis and summarizes the contributions made by the research work and future work in the area has been recommended.

1.10 Summary

This chapter described the fundamental notions about the importance of health, lifestyle diseases, diabetes mellitus causes, types, diagnosis and management. The requirements of machine learning, types of ML, strength and challenges of ML and importance of ML in healthcare. Motivation for the research, problem statement, objectives, scope and research contributions are presented and the thesis's organization is summarized. These are instrumental in emphasizing the remarkable potential of this research domain. Theoretical foundations for the research are discussed in the next chapter.

CHAPTER-2

THEROTICAL BACKGROUND

CHAPTER-2 **THEORETICAL BACKGROUND**

2.1 Introduction

This chapter emphasizes the theoretical groundwork for conventional feature selection approaches and autoencoder based automated feature selection. Various conventional classifiers of ML and ideas pertaining to the model selection followed by validation of the ML model are also discussed.

2.2 Feature Selection Approaches

Increased number of features, increases computation cost and affects the performance [19]. Hence, the primary goal is to create an optimal feature subset S = {f1,f2, f3,…, fi}containing useful information ensuring Minimum redundancy and Max-relevance properties. Global relevance is more crucial in feature selection. [20]. Feature Selection or feature extraction was the most major step. The process of choosing pertinent features, or a candidate subset of features, without updating them and validating them in light of the analysis objective is known as feature selection. [21]. Features can be classified as strongly relevant, weakly relevant, irrelevant, and redundant.

If data also have high dimensions, using the feature selection method, the dimension can be reduced without losing any information. The feature selection method aids in reducing dimension without losing information. It reduces overfitting, increases performance by selecting the right subset of features and minimizes execution time. The process of feature selection is automated or semi-automated. The feature selection is discussed as follows,

Let the input vector, X = {x1, x2, . . . , xp} with class label is,

Y = {0,1} with k features such as,

H = { f1, f2, . . . , fk}is given as,

$$F(X,Y) = \{(\mathrm{Xj}, \mathrm{Yj}) |\ \mathrm{Xj} \in \mathrm{R\ p}\ , \mathrm{Yj} \in \{0,1\}\}\} \tag{2.1}$$

Where Xj = {x_1, x_2, . . . , x_p} are independent variables and Yj ∈ {0,1} is class labels in the dataset.

The evaluation criteria are used for getting an optimal feature subset that can be categorized in a filter, wrapper and embedded techniques. Below figure (2.1) shows feature selection techniques.

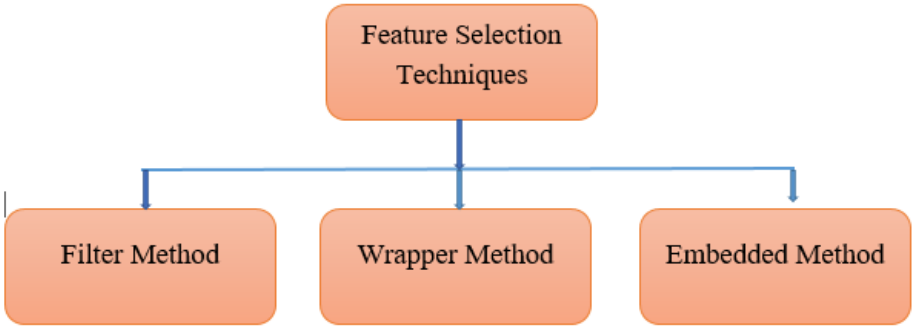


Figure 2.1: Types of feature selection techniques

2.2.1 Filter Method

Filter method is a ranking based techniques. Ranking criterion is used for variable selection. These methods selects features depending on a score irrespective of the classifier used. Below, figure (2.2) depicts the process of the filter method.

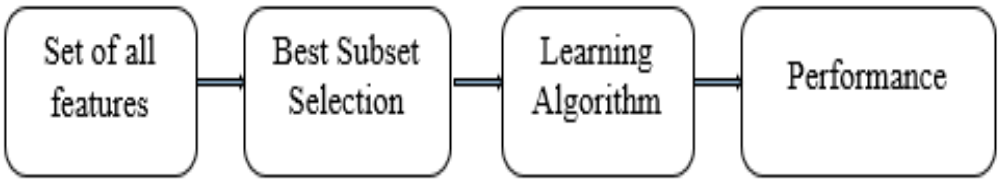


Figure 2.2: Process of filter method

By correlating the feature relevance with the target variable, the filter-based method quantifies feature relevance. Feature relevance provides the usefulness of the features in discriminating the different classes [22]. Measures such as, information, distance, consistency, similarity, and statistical measures all play a role in feature filtering. Information gain computes a reduction in entropy while transforming the dataset. The Chi-square test is used to deal with the categorical feature. Mutual information determines the statistical dependence between two features. Fisher score

returns the ranks of the variables in descending order. The linear relationship between two or more features is measured by correlation. Positive or negative correlation is possible.

2.2.2 Wrapper Method

The wrapper feature selection algorithm assesses the utility of a subset of features by training the classifier on them.The assessment is iterated for 2^N subsets and using a heuristic approach, the best subset of the features gets selected. Subset creation depends on the search mechanism. Below, figure (2.3) demonstrates how the wrapper method works. For getting the best subsets, the wrapper method depends on the classifier's resource requirements. To find a subset of variables various search algorithms can be used. Backward feature elimination, forward feature selection, recursive feature elimination and exhaustive feature selection are different types of wrapper techniques.

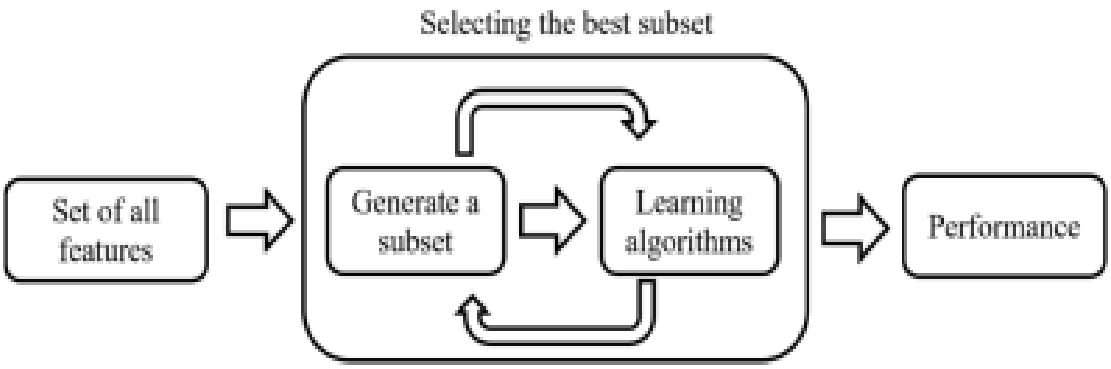


Figure 2.3: Process of a wrapper method

2.2.3 Embedded Method

The embedded technique is a built-in method for choosing variables. Automatic selection or rejection of the variables occurs. Regularization models with objective functions, which lower the fitting errors, are necessary for feature weighting. LASSO regression regularizes the estimates of a variable's coefficient to zero. Figure (2.4) below illustrates how the embedded method works.

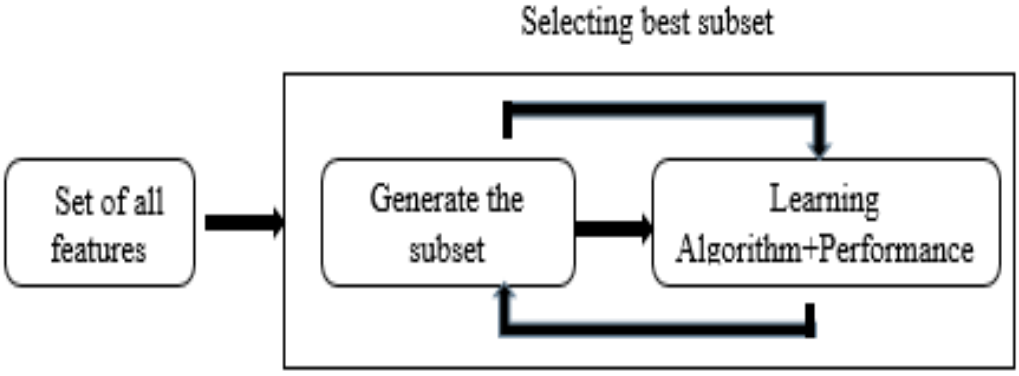


Figure 2.4: Process of the embedded method

2.2.4 Autoencoder based method

An autoencoder is a feedforward neural network with a hidden layer, an output layer, and an input layer are shown in figure (2.5). An autoencoder training is based on increasing /decreasing the hidden layer neurons [23].

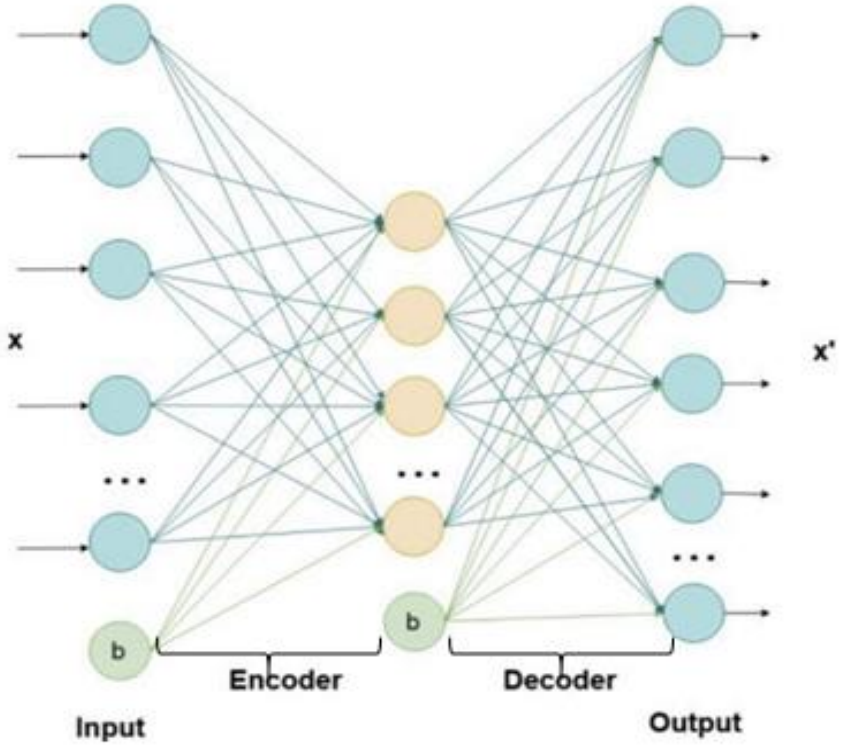


Figure 2.5: Hidden layer based autoencoder structure

Autoencoder consists of encoder and decoder. The input layer defines the internal representation that the encoder uses to interpret the input and compress it. The hidden layer encodes the input vector.The decoder uses the encoder's output to try

to reconstruct the input. Since objective is to reproduce the input data on the output layer, both x and x' have the same dimension. The basic structure is shown in figure (2.6).

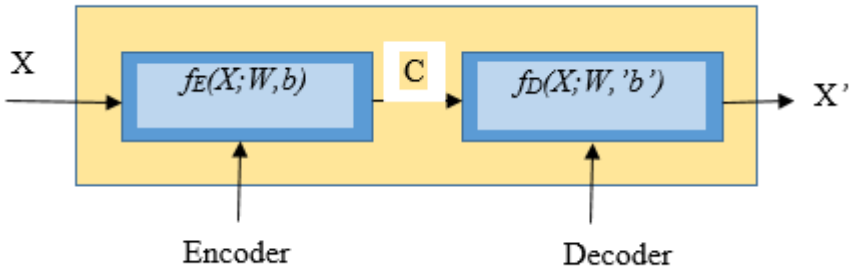


Figure 2.6: Autoencoder structure

Different activation functions used in autoencoder are linear, binary, rectified linear units (relu), sigmoid, hyperbolic tangent and Scaled Exponential Linear Units (Selu) functions [23]. The equation of each of these is mentioned below.

$$S_{linear}(x) = x \tag{2.2}$$

$$S_{binary}(x) = [x > 0] \tag{2.3}$$

$$S_{relu}(x) = x\,[x > 0] \tag{2.4}$$

$$S_{selu}(x) = \lambda \begin{cases} \alpha e^{x} - \alpha & x \leq 0 \\ x & x > 0 \end{cases} ; \lambda > 1 \tag{2.5}$$

$$S_{sigm}(x) = \sigma(x) = \frac{1}{1+e^{x}} \tag{2.6}$$

$$S_{tanh}(x) = \frac{e^{x} - e^{x}}{e^{x} + e^{x}} \tag{2.7}$$

2.3 Conventional Machine Learning Classifiers

Important algorithms of supervised learning are explained below.

Logistic Regression (LR):

It is one of the method of classification. that classifies the data in the form of binary variables means if the person has diabetes or not. The objective includes identifying independent variables that impact the dependent variable. Logistic regression (LR) is based on the Maximum Likelihood estimation technique [5,6]. For analyzing a dataset where >1 independent variables which find an outcome using logit function where, logit (z) = log (Probability of event / probability of no event) =log(z/1-z) =log(odds) . The importance of input features is determined by the coefficients of the model. LR requires large sample sizes. For prediction and classifications of medical problems, such as diagnosis of disease LR algorithm is used. The logit model is depicted in the equation below.

$$z = \frac{e^{y}}{1+e^{y}} \tag{2.8}$$

$$log\left[\frac{z}{1-z}\right] = y \tag{2.9}$$

$$log\left[\frac{z}{1-z}\right] = logit(z) = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots .. + \beta nxn \tag{2.10}$$

Where β0→ Intercept and β1, β2… βn are the coefficient.

Using intercept and coefficient for the features, the probability is calculated. The advantages of these are that it is a probabilistic model and this model can be updated easily. Limitations include not showing a linear relationship, a complex relationship between variables does not give good accuracy of the model and due to sampling bias, may overstate prediction accuracy [10-11]. Figure (2.7) demonstrates the data into two classes.

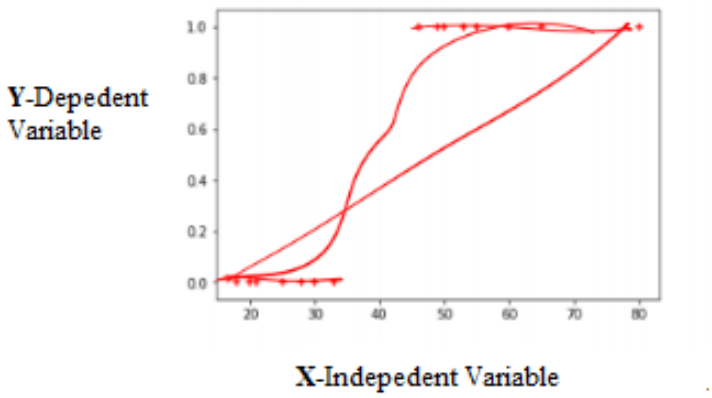


Figure 2.7: Graphical representation of logistic regression

Decision Trees (DT):

Decision tree (DT) are prepared using an algorithm that identifies a mechanism to divide features based on conditions. A DT makes a hierarchical decision about the sequential outcomes variable based on the predictor data. Figure (2.8) shows DT. In DT, each node, transition path, and leaf represents a feature, decision, and outcome, respectively and decisions are provided in the form of if-then-else statements. Through Entropy and Gini Index, information gain is calculated, which is used for the splitting of a node.

$$Entropy(S) = \sum_i -p_i log_2 p_i \tag{2.11}$$

$$Gain\ (S, A) = E(S) - \sum_t P_i\, E(t) \tag{2.12}$$

A predictive function, h(x): X → Y, predicts the outcome variable associated with sampling x from the root to a leaf node. It can work with Small data, both continuous & categorical. It can easily handle outliers. The pruning method reduces the size of the DT. DT can generate a robust classifier validated through statistical tests. Limitations include if a non-leaf node variable value is missing, it cannot branch and will work on the order of variables [5, 8, 10-11].

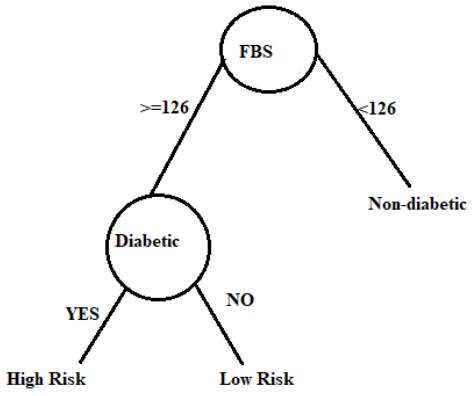


Figure 2.8: Decision tree with n=2 nodes

Random Forest (RF):

It is an ensemble classifier formed by training multiple DT simultaneously, called "forest". DT causes overfitting of the training data; Random forest (RF) prevents this. Accuracy is proportional to the number of DT. For decision, each RF takes a dataset sample and a subset of features. The prediction function f(x) is defined as follows:

$$f(x) = \sum_{m=1}^{M} Cm \prod(x, Rm) \tag{2.13}$$

Where Cm is a constant and Rm is region to m:

$$\prod(x, Rm) = \begin{cases} 1, & x \in Rm \\ 0, & otherwise \end{cases} \tag{2.14}$$

Each tree performs classification; the forest then selects the classification with more 'votes' or the average votes [11]. It reduces the variance resulting from a single DT. It also works for the large dataset and can estimate important variables in classification. Limitations include computationally complex and expensive; it favors only higher values of variables. Below, figure (2.9) shows a random forest model.

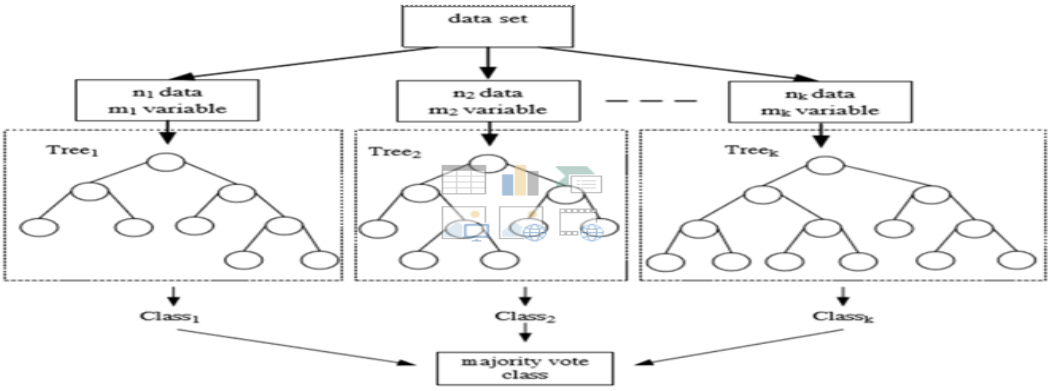


Figure 2.9: Random forest model [82]

Support Vector Machine (SVM):

This model applicable to classification and regression problems of linear and nonlinear data. Input is given as, {(xi, yi), . . ., (xn, yn) } while output data format is, $Y \in \{+1, -1\}$ here n is number of samples. The following is a linear classifier function:

$$g(x) = sign(h(x)) \tag{2.15}$$

Where $h(x) = w^T x + b_0$, with $\{w, x \in R_n\}$ and $\{b_0 \in R\}$.

The goal is to reduce classification errors and locate the highest-margin separating hyperplane. The hyperplane should be far away from the data points for better generalization. The points that lie nearest to the hyperplane are support vectors [10-11]. Working of SVM is shown in figure (2.10).

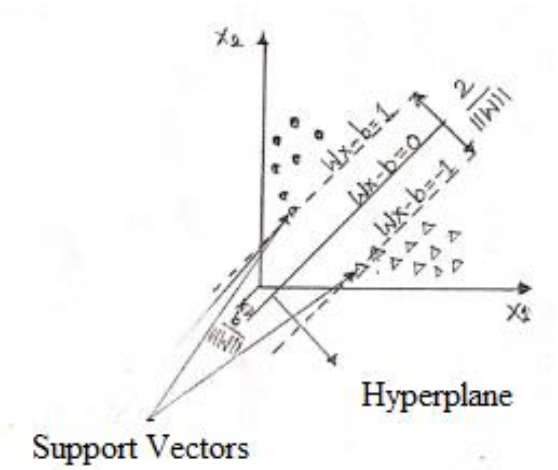


Figure 2.10: Working principle of support vector machine

In Linear SVM, a minimum number of features is used and the Kernel trick is applied for more features. Original training data is transferred into a higher dimensional space that can be done using a Kernel approach. SVMs work well on small datasets. There is less risk of overfitting. It can work for semi-structured or unstructured data. Limitations are less capable of noisy data, computationally expensive for complex data, generic SVM extension is required and cannot classify more than two classes.

Gradient Boosting Classifier:

Gradient boosting classifier (GBC) is an iterative algorithm in which different weak classifiers are trained and then combined to form a robust classifier. To optimize the results, each iteration defines a new weak classifier to overcome the existing deficiencies of the weak classifier. It is used to decrease the bias error. This algorithm works on the same training set. It is less prone to overfitting.

K-Nearest Neighborhood (KNN):

This algorithm is applicable to regression and classification. This algorithm determines the unclassified points by measuring the majority class vote from its k-nearest neighbor. KNN is considered lazy learners as there is no learning of the model. During the training phase, it just stores and during testing, after getting new data. It calculates the distance using from the query point to the points in the training phase. For distance calculation, Hamming distance, Euclidean's distance, Manhattan distance, or Minkowski distance method were used. After distance calculation, '*k*' nearest neighbors are selected and the resultant class is calculated based on major votes.

If a variable is a categorical variable, it uses hamming distance. If it is a continuous variable, the distance measure is Minkowski distance. Euclidean distance performs well for categorical and numerical datasets. Euclidean distance is the most frequently used and given as,

$$d(x,y) = \sqrt[2]{\left(\sum_{i=1}^{n}(xi - yi)^2\right)} \tag{2.16}$$

Advantages include can classify instances quickly and used for both classification and regression. Limitations include that the distances between vectors cannot be

calculated on missing data, expensive to increasing attributes and weightage is given to all attributes, which may lead to poor classification performance.

Artificial Neural Networks (ANN):

ANN is based on the idea of the function of nervous systems in the brain. Like brain neurons, the ANN algorithm is an interconnected group of nodes that allows the system to learn and streamline by analyzing new data. The components of ANN are the input, hidden, and output layers Figure 2.11. The network's total number of variables is the same as its total number of neurons. Input layer has input = n, output = (n/2). The variables in the hidden layer are (n+1), where n is the number of features, and are equal to the sum of the input and output variables. Activation function plays an important role in neural network e.g. sigmoidal function given by, $f(x) = 1/1+e^{x}$. This algorithm can be used to evaluate the function of a large number of input variables. A cost function is used for optimization. The parameter is learned using a back propagation neural network, and the gradient descent algorithm can be used to find the best parameters. The advantage of this algorithm is that it is suitable for both categorical and numeric variables. It needs lesser statistical training. Moreover, it can identify nonlinear relationships between outcome and independent variables. The limitation includes that complex classification problems, computationally expensive and predictor variables require pre-processing steps [9-11].

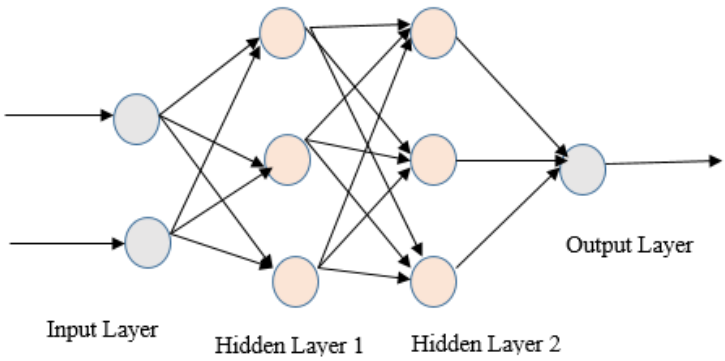


Figure 2.11: Basic neural network with two hidden layers

Naïve Bayes:

This classification algorithm is based on probability theory, considering that all variables are independent. Ability to handle continuous and discrete data and make

probabilistic predictions. The algorithm works on the simple Naïve Bayes formula, which is represented as,

$$P(C|B) = ((P(B|C)XP(C))/P(B) \tag{2.17}$$

Where,

P (C|B) denotes the target class's posterior probability.

P (B|C) stands for the class's likelihood probability.

P (C) is the predictor's prior probability.

P (B) is the prior probability of class

From the training data, the prior distribution and likelihood probability can be calculated. This classifier is used for high-dimensional data. Dependency between variables affects classification performance. By default, this algorithm assumes the normal distribution for the numerical data [11].

2.4 Model Selection using hyperparameter tuning

In ML techniques to evaluate models, hyperparameter tuning is important. It is the process of choosing the ideal set of hyperparameters. A hyperparameter is a parameter that regulates the learning process. Before starting task, the value of the hyperparameter model is fixed. These values are derived by training the data and maximizing the performance on a validation set. As data is divided into training, testing and validation phase, selecting appropriate hyperparameters is critical. Below, figure (2.12) depict the process of hyperparameter tuning.

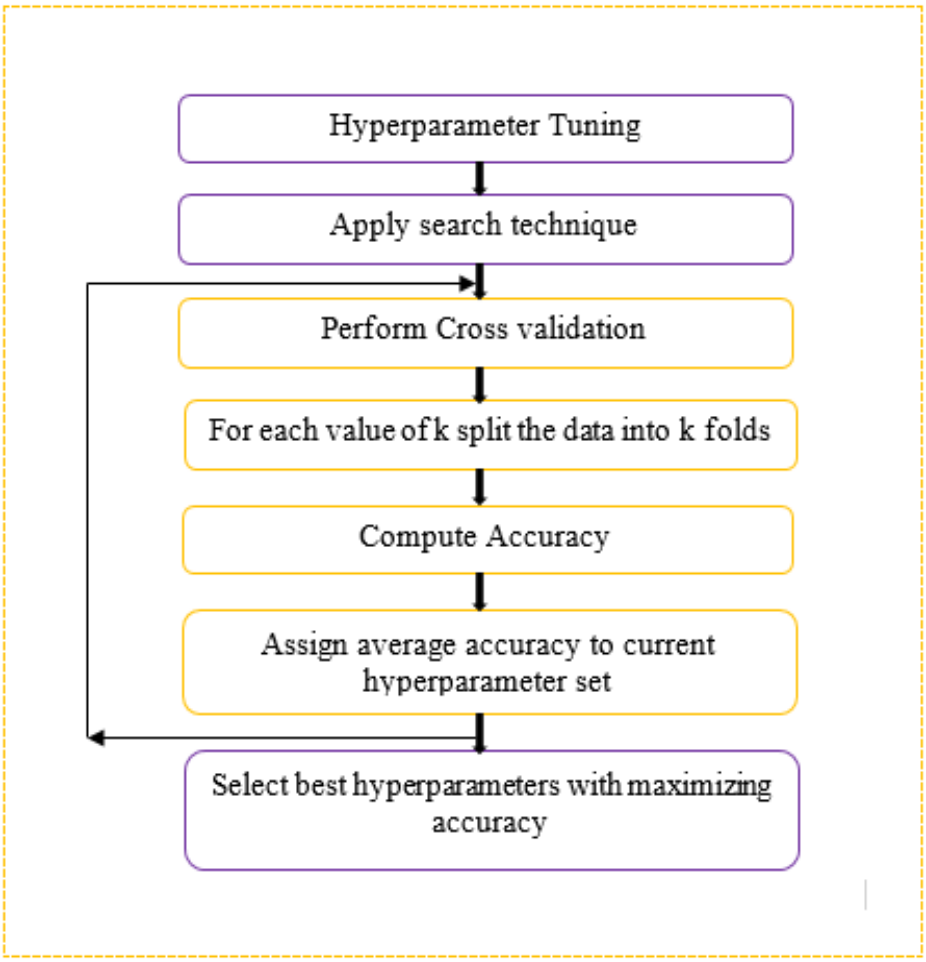


Figure 2.12: Process of hyperparameter tuning

Traditionally, one can assign values manually by trial and error with all possible combinations, but the best settings of values on one dataset may not work on another one. One classic and favored way is by using the automated grid search technique, create a grid of hyperparameter values, train a model, and to determine scores for the testing data for each combination. Random search algorithms select random combinations of the hyperparameters. Grid search and random search are depicted in figure (2.13). For example, RF has the hyperparameters as 'criterion,' 'max_depth,' 'max_features,' 'n_estimators,' etc. Penalty in LR, gamma, C and kernel in SVM and k in KNN are the hyperparameters for probabilistic models.

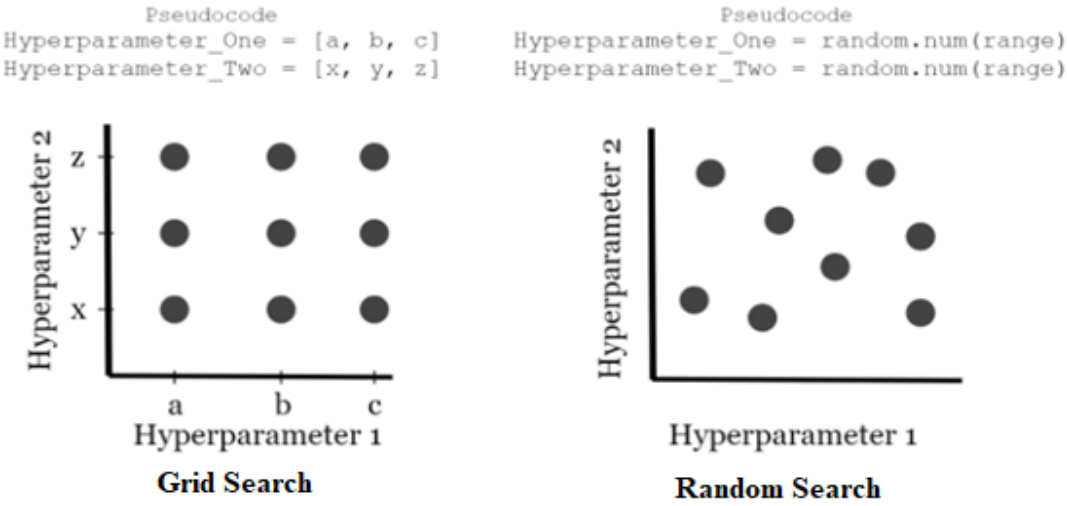


Figure 2.13: Hyperparameter using grid search and the random search

(Source: Medium.com)

2.5 Model Validation

K-fold cross-validation (KCV) is a technique for estimating the accuracy of a model. Once the model has been trained, we cannot be certain that it will perform well on new data; thus, assurance about the model's performance is required. Testing is to be done on some unseen data. We can assess the model's fit to unknown data and determine whether it is underfitting or overfitting or balanced. In underfitting, training and testing error rates are high, while in overfitting, the training rate is low and the testing rate is high and in balanced conditions, both trainings and testing rates are low. An efficient learning model should be balanced. Cross-validation is a beneficial technique to test on limited data. Figure (2.14) demonstrates a detailed data presentation using K-fold validation.

In this technique, the dataset has been folded K times. The first k-1 folds are used for training, and the last one is used for testing. This process is repeated k times, with each step employing a new fold for testing. In each step, evaluate the model's performance. Finally, the model's overall performance is equal to the average performance. (P). Let P be the average model accuracy, $P_i \in R$, i = 1, 2, . . . , K fold shown in equation (2.18).

$$P = \frac{1}{k} * \sum_{i=1}^{k} P_i \tag{2.18}$$

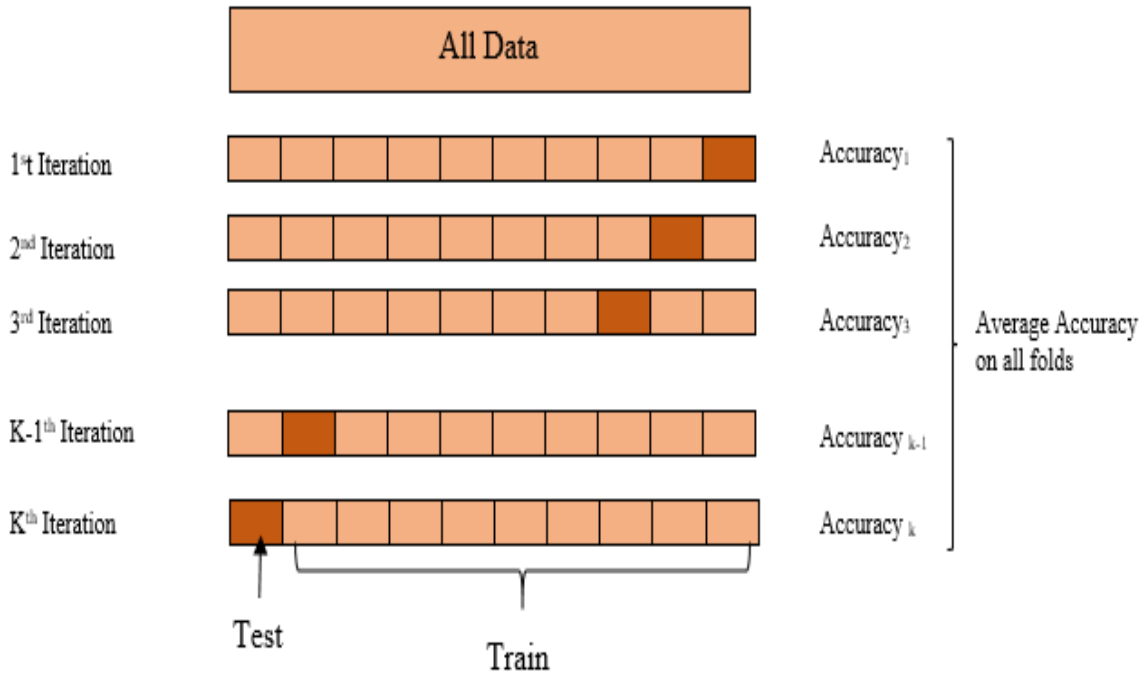


Figure 2.14: Detailed representation of data using k-fold validation

2.6 Summary

This chapter laid down the theoretical foundation for our research work. Chapter mentions feature selection approaches, both traditional and automated, based on feature selection. Conventional machine learning classifiers with their pros and cons. This chapter also explored model selection using the hyperparameter tuning method. Finally, model validation using K-fold cross-validation is also discussed.

CHAPTER-3

LITERATURE REVIEW

CHAPTER-3 LITERATURE REVIEW

3.1 Introduction

Diabetes mellitus is one of the serious issues. In recent years, ML has been the evolving, reliable and supporting tool in healthcare for solving many real-world problems. In recent days, early diagnosis of diabetes has been predicted using machine learning and has gained more attraction. This research uses ML classifiers to predict diabetes based on personal, stress and lifestyle-based information. It is helpful to identify the requirement for diabetes mellitus diagnosis using ML classifiers. This chapter presents works that have a close relationship with significant impacts on the early diagnosis of diabetes using a predictive model. The survey is organized into six sections. ML based survey for early diagnosis of type-2 diabetes is proposed by various authors listed in section3.2. Section 3.3 discussed various feature selection approaches used in the medical domain. Section 3.4.1 impact of stress on type 2 diabetes related works are surveyed. Section 3.4.2 mentioned lifestyle parameter related works are explored. Section 3.5 describes the research findings and gaps. Section 3.6 concludes the chapter.

3.2 Survey on Machine Learning Algorithm for Type-2 Diabetes

Contreas I et al.;[24] study focuses on all AI techniques for diabetic prediction and management. Study shows an AI is an effective tool applicable for prediction and prevention of complications due to diabetes. AI techniques are being progressively utilized in the area of medicine containing complex sets of diagnostic and clinical information. This tool aids to recover the patient's health through a predictive approach and helps for improving health outcomes.

Rigla and García-Sáez [25] examined the tools and AI method used in diabetes detection. Many different applications make use of AI techniques. It can be used to predict the likelihood of developing diabetes based on the types of abilities, such as reasoning and learning. Healthcare providers can benefit from a variety of AI techniques, including Case-based analysis, data mining, fuzzification,

defuzzification, SVM, evolutionary strategy, hybrid systems, naive Bayes, decision tree supervised learning, and unsupervised learning are helpful for healthcare providers.

Uddin et al.;[10]designed a study to provide an overview of various supervised machine learning algorithms for predicting disease. Due to the differences in research scope and clinical data across studies, comparisons could only be made once a uniform benchmark for dataset and coverage was established; hence, the author included studies on the same data. The Support Vector Machine (SVM) algorithm was found to be the most frequently used, followed by the Nave Bayes algorithm. Observation shows that out of 17 studies, best 9 show RF gives the performance, followed by SVM.

Perveen et al.; [26] study used diabetes is detected using the AdaBoost and bagging ensemble techniques. Along with these techniques, a base learner J48 (c4.5) decision tree is used using risk factors. Three diverse adult age groups are selected for classification. The Chi-square test is used to investigate the relationship between age groups and diabetes. The experimental results show that the AdaBoost ensemble method, outperforms among all three groups than bagging and J48 decision tree.

Jayanthi N. et al.; [27] discussed various predictive models, steps to develop, usage, assessment,and applications in diabetes. The survey summarizes developed predictive models, diverse datasets, techniques and tools used. The review shows the scope of the clinical prediction models, particularly for diabetes disease in India. Study also reveals a variety of prediction models used for diabetes. Observation shows hybrid model becomes more accurate than the traditional model. Based on the study author also mentioned some gaps.

Sisodia D. et al.;[28] designed a system to predict the possibility of diabetes by achieving higher accuracy. PIDD dataset from the UCI repository used for study development. Experimentation was done using the weka tool by applying NB, DT and SVM classifiers for early detection of diabetes. Model performance was also done. The performance results for NB were the best, with a maximum accuracy of 76.3% and the highest ROC value of 81.9.

Mujumdar and Vaidehi[29] designed a pipelined-based approach using machine learning. Proposed system involves some risk factors such as body mass index

(BMI), glucose, age, insulin, skin thickness, pregnancies and job type. The dataset contains 800 records and ten attributes. A total of 10 different classifiers involved in study. A pipeline model was created for improving classification accuracy. Comparative study of all models done on the PIMA dataset and the dataset used so far found that LR gives maximum accuracy of 96%. The application of the pipeline shows that the AdaBoost classifier gives an accuracy of 98.8%.

Zou et al.; [30] published a paper on the prediction of diabetes with ML techniques. The authors used classifiers such as -RF, DT, and ANN, and dimensionality reduction was done using PCA and minimum redundancy maximum revalence (mRMR). The study involves Luzhou China-82694-based dataset containing 14 features and a Pima dataset. The author used ML approach. The result reveals that for all features RF method shows accuracy of 80.84%. The experimental evaluation suggests that PCA is not suitable for this data. For the PIMA dataset, using mRMR –RF gives a better result of 77.21%. Study reveals that blood glucose is an important risk factor, but only using glucose cannot achieve the best result.

Lai H et al.;[31] developed a model to aid doctors in identifying patients who will develop diabetes in the future using LR, GBM, RF, and DT on different features age, sex, FBG, BMI, high-density lipoprotein(HDL), triglycerides, low-density lipoprotein(LDL) and blood pressure(BP)shows GBM and LR model performs better i.e.AROC value is 84.07% and 84% with a sensitivity of 71.6% and 73.4%.The dataset used for validation is Canadian-centric, containing demographic and laboratory results. FBG, HDL, BMI and triglycerides were the significant predictors.

Patil et al.;[32] using the simple k-means clustering algorithm, a hybrid prediction model (HPM) was implemented. The model is developed and validated using Pima Indian Diabetes Dataset. The final classifier model is built using the C4.5 algorithm and utilizes a k-fold cross-validation technique. Performance metrics such as specificity and sensitivity are used for evaluating the model.

Mahabub et al.;[33] designed a system to predict diabetes by improving accuracy. Used 11 classifiers such as NB, KNN, SVM, DT, RF, ANN, LR, GB, AdaBoosting, etc., on the PIMA dataset. Evaluations of all models are examined on the accuracy, precision, F-measure and recall. Ensemble voting classifier developed using the

three best classifiers as SVM, MLP and KNN by applying hyperparameter tuning and cross-validation. The proposed ensemble framework gives an accuracy of almost 86%.

Kumari et al.;[34] designed an ensemble soft voting approach. PIMA dataset and breast cancer datasets used for evaluation purposes. In this study, the author used an ensemble of three algorithms RF, LR and Naïve Byes. Compared performance with an existing system and found that the system outperforms with 79.08% accuracy and an F1-score value of 80.6% on the PIMA dataset. Comparative analysis of the proposed method on the breast cancer dataset shows an accuracy and an f1-score value of 97.02% and 96.32%.

Ahmed et al.; [35] DT, KNN, NB, RF, GB, LR, and SVM were used to create a system for predicting DM. To improve accuracy, label-encoding and data normalisation are used. PIMA and Tigga and Garg are two different datasets are used. The PIMA dataset achieves the highest accuracy for SVM and RF with 80.26 %; the other dataset achieves the highest accuracy with DT and RF with 96.81 % and developed a web app. The model is compared to other studies, and the results show that the suggested model has a higher accuracy range of 2.71 % to 13.13 %.

Sivashankari et al.;[36] designed a stacked ensemble model.Pima diabetes dataset is applied for experimental purposes. The author used RF, KNN, DT, LR, GBM and Gaussian NB. As the single classifier will not be suitable for the unstructured and large datasets hence the proposed stacking-based ensemble model. The proposed model shows accuracy, precision, recall and F1-score value of 93.1%, 84%, 83.9% and 83.5% respectively for a highly categorical dataset.

Birjais et al. [37] designed a system to diagnose diabetes. Experimentation was done on the PIMA diabetes dataset. The author used different techniques like GB, LR and NBC to predict and diagnose diabetes. During analysis, BMI has played an important role. The implementation results show that GB, LR and NB give accuracy of 86%, 79% and 77%, respectively. This shows GB classifier achieved accuracy and sensitivity of 86% and 89.5%, while the LR model of 100% showed the highest specificity.

Khaleel and Al-Bakry [38] created a machine learning model for predicting diabetes on PIMA dataset. Classifiers such as,LR, KNN and NB used. LR model found an

efficient model by achieving a precision of 94%, NB and KNN show the precision value of 79%and 69%. LR shows the highest recall and F1-measure values of 70% and 78%, respectively.

Diwani and Sam [39] have developed a system using J48 and NB. Model development was done on Pima dataset from the UCI machine learning laboratory. The Waikato environment tool has been used for this study. 10-fold cross-validation applied for validation purpose and observed result through analysis shows that NB performs better than the J48 model by achieving an accuracy of 76.30% and 73.82% J48. NB exhibits an excellent value of a ROC value of 0.8.

Khanam and Foo[40] proposed research to predict whether or not a patient has diabetes The study's dataset is the Pima Indian diabetes dataset, which was obtained from the UCI Machine Learning Repository and contains information about 768 patients with 9 attributes. The proposed system used 7 ML classifiers, KNN, LR, DT, RF, NB, AB, and SVM. After removal of outlier and extreme values dataset contain 699 records from them, 466 are non-diabetic and 233 are diabetic. The correlation-based feature selection method was applied. Comparative analysis of various models on train test split accuracy and cross-validation accuracy was done. The result shows that LR and SVM model works well with an accuracy of approximately 77% to78%.

Wang et al.;[41] proposed an Imbalanced data prediction algorithm based on machine learning Experiments were carried out on Pima Indians diabetes dataset. The proposed method addresses both pre-processing and classification phases. NB method is applied for missing values computation, while class imbalance can be achieved by an adaptive synthetic sampling method, the ADASYN approach. RF classifier is used. K-fold applied for validation of classifier. Due to the insensitivity of accuracy, various performance indicators are considered for evaluation purpose. The DMP MI algorithm outperformed for K=5 with accuracy, precision, F1-score, and AUC values of 87.1 %, 80.6 %, 83 %, and 92.8 % respectively.

Krishnamoorthi et al.;[42] The author proposed an intelligent machine learning and big data-based framework for diabetes prediction. They used four different machine learning methods on the PIMA diabetes dataset: LR, RF, SVM, and KNN. Using machine learning, the author proposed an intelligent diabetes mellitus prediction

framework (IDMPF). The authors describe model assessment strategies and prediction issues using the framework and provide a solution as a result. The results show that RF and SVM gives accuracy of 83 %, RF is suitable for disease prediction, and when hyperparameters are applied to LR, the accuracy of prediction is 86%. The findings show that glucose and BMI have a strong correlation with diabetes. The author also suggested further studies using family history, smoking, drinking, and physical inactivity to predict diabetes.

Dagliati A. et al.;[43] created a model to predict retinopathy, neuropathy, and nephropathy at 3,5 and 7 years after the initial visit. Criteria used for patient's selection: the patient's follow-up duration; complications developed and the registered onset date of complications. Gender, hypertension, age, glycated haemoglobin (HbA1c),BMI, smoking behaviour, length of diabetes, and body mass index are characteristics that were examined. A stepwise feature selection algorithm used by LR yields an accuracy of 83.8%. While hypertension raises the risk of retinopathy and nephropathy, the duration of diabetes and BMI are risk factors for retinopathy and neuropathy.

3.3 Survey on Feature Selection Techniques

Chaki et al.;[44] reviewed ML models used in diabetes detection, diagnosis and self-management. The review also highlighted three research challenges. The author included 107 studies according to the model, the dataset, and pre-processing methods. Utilization of 4 different features selection, namely, color, texture, shape and text features, and their performance. The authors observed that texture, text and shape features produced better outcomes. Also, they found that SVM and deep neural network (DNN) delivered better classification outcomes, followed by RF and ensemble classifiers.

Chen et al.;[45] used three diverse datasets: car evaluation, bank marketing and human activity recognition using smartphones. The used feature selection methods for this study are Boruta, RF methods varImp() and RFE. The author evaluated and compared classifiers RF, SVM, KNN, and LDA and showed RF feature selection method gives the best result among all groups. For human activity recognition using a smartphone dataset RF-RF, the technique displayed an accuracy of 98.57% with 561 features and 93.26% with six features; for the car dataset with six features,

accuracy is 93.31% and with four features, it becomes 93.36%. While in the banking dataset with 16 attributes, accuracy is 90.88%, with seven, it is 90.99%.

Kumar et al. [46] classified medical diseases by combining genetic algorithm-based feature selection techniques and SVM. The best feature subset for the SVM classifier is found using the proposed Genetic-SVM model. A comparison of the SVM model with the proposed model was done on five different datasets and showed better performance in the proposed Genetic-SVM technique. The proposed Genetic-SVM model effectively worked on all other datasets. Compare to conventional SVM, Genetic-SVM shows good result.

Sneha N.et.al;[47] studied various classification algorithms for designing a prediction algorithm and finding an optimal classifier for generating results near clinical outcomes. Classification algorithm RF, KNN, DT, NB and SVM were used in the study. In order to aid in the early diagnosis of diabetes, the author created a model using an optimal feature selection algorithm. Optimal feature selection is based on correlation values. Analysis shows that the DT and RF models have 98.20% and 98.00% specificity. Naïve Bayesian produced an accuracy of 82.30%, as shown in the experimental result. The highest disease prevalence is found in SVM at 45.7%. For enhancing classification accuracy, optimum feature selection from the dataset is generalized using this research.

Hasan MD.et.al;[48] Proposed framework used PIMA dataset for diabetes prediction with 768 female patients, 268 of whom have diabetes, and 500 of whom do not, along with 8 different characteristics: age, pregnancy, BMI, pedigree, triceps, pressure, glucose, and pressure. Designed framework consists of various pre-processing steps: outlier rejection, missing values replacement with mean and data standardization with three different feature selection techniques. The author developed a weighted ensemble model of two boosting classifiers to boost the model performance. Grid search is used for maximizing performance. The result showed that the ensemble classifier is the best, with correlation-based feature selection having the highest AUC value of 0.950 and specificity of 93.4%. The showed result outperformed by 2.00% in AUC than other studies.

Kaur H.et.al;[49] conducted a study to forecast and analyze diabetes trends using the PIMA dataset. The R data manipulation tool was utilized and developed a predictive

model using five supervised machine learning algorithms: RBF-SVM, linear-SVM, KNN, ANN and multi-factor dimensions reduction (MDR). The study highlighted that the boruta wrapper algorithm works well than manual selection. Optimization can be done through hyperparameter tuning. The study shows that the Linear SVM model provided the best accuracy of 89% and precision of 88%, while the KNN model showed the best recall and f1-score values of 90% and 88%. The highest Auc-value of 0.90 and 0.92 found in linear SVM and KNN models.

ML approach was used by Maniruzzaman et al. [50] to categories and predict diabetes. The author used four ML algorithms NB, DT,RF and AdaBoost. Additionally three different protocols (K2, K5, and K10) with 20 trials were used to implement these techniques. The author used dataset from the National Health and Nutrition Examination Survey, which included 6561 people; 657 of them had diabetes and 5904 became non-diabetic. A logistic regression model based on p-value is used for feature selection. For the K10 protocol highest accuracy of 94.25% and AUC value of 0.95 was observed with logistic regression and random forest combination compared to conventional techniques.

Maniruzzaman Md et al.;[51] have developed a ML-based risk-stratification system which replaces outliers or missing values with a median, which is optimized. The dataset used for this study is PIMA Indian diabetes dataset (PIDD). Different classifiers used are, NB, SVM, Adaboost, LR, DT, RF, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), Gaussian process classification (GPC), and ANN are used in combination with six feature selection techniques to optimize the features. Experimental result shows that RF-RF gives an accuracy of 92.26% for K10 and 99.99~100% for JK protocols. Other performance measures include sensitivity of 95.96% and an AUC of 0.93.

Haq et al.;[20] On the basis of the clinical diabetes data set, a ML based diagnosis system was proposed. filter-based DT, Ada Boost, and RF were used for feature selection. For this study, DT was used as a classifier. Hold out, K-fold, and leave one subject out (LOSO) are used for validation purposes. The outcome of the comparison showed that the GL, PDF, and BMI features are more significant. With k-fold and LOSO, the proposed DT (ID3)-DT achieved maximum accuracy of 99.8% and 99.9 %, respectively. When compared to state-of-the-art, the DT-DT performs well in terms of accuracy.

Dutta et al.;[52] designed a system for discovering the critical elements for the risk of diabetes. To predict whether the blood test detection is the positive or negative author used three different classifiers, LR, SVM and RF. To achieve this collected dataset of 392 records had around 262 class 0, and 130 for class 1.Feature importance and correlation using heatmap applied for selecting the significant features. Five important features observed using the feature importance technique are glucose, insulin, age, BMI and diabetes. The experimental result shows that RF is an ideal classifier for predicting Diabetes, with an accuracy of around 84%.

Mishra et al.;[53] system is designed on PIMA dataset using four distinct filter-based methods. The performance of filter methods is assessed using the RBF IBK and JRip classifiers. The Weka tool was used to conduct the study. The outcome demonstrates that the Correlation method has the lowest error rate of 0.0013 and Matthews Correlation Coefficient value of 0.667 and the Chi-Square method has f-Score value of 0.843, while the Information Gain method gives the highest accuracy of 98.78 % and Kappa Statistics value of 0.667. The study found that filter-based techniques improve the performance based on comparative analysis.

S Sivaranjani [54] study based on PIMA dataset . Principal component analysis as well as step-forward and step-backward was used for feature selection and analyzed their performance. Significant features were applied to the model and compared the model performance of RF and SVM. The experimental result showed that the best accuracy showed by RF of 83% compared to SVM, which has an accuracy of 81.4%.

Deberneh and Kim [55] using collected Korean population data from 2013 to 2018, an ML model was developed to forecast the prevalence of diabetes in the ensuing year. Key features were selected based on chi-squared tests, ANOVA tests and RFE. Five different classifiers are used for predicting an outcome, namely, LR, RF, SVM, Boost and ensemble algorithm. The result shows that soft voting and random-forest-based approaches achieved a prediction accuracy of 73%, while XGBoost's accuracy was 72%. The author discovered that when using data from the previous four years, cross-validation accuracy increases to 81 % from 77%.

Jian et al.; [56] developed a study to find eight different complications due to diabetes. Collected dataset from UAE. For selecting the top 5 and 10 features, feature selection was performed for each complication and compared performance on LR,

RF, DT, AdaBoost, and XGBoost classifier. Results showed that for diabetic foot complications. XGBoost model gives accuracy of 97.8%.

MINH LE et al.;[57] designed a model to predict diabetes patients' early onset. Adaptive Particle Swam Optimization (APSO) and Grey Wolf Optimization (GWO) with Multilayer Perceptron (MLP) classifier were used in the study to select features based on wrappers. Traditional ML algorithms like SVM, DT, KNN, NB, RF, and LR are used to compare performance. Higher prediction accuracy could be attained, with 96 % and 97 % for GWO - MLP and APGWO - MLP.

Gill and Pathwar[58] proposed study containing 390 people's records containing 18 features with 60 diabetes and 330 non-diabetes. Feature selection techniques employed in the study were Mutual Information, ANOVA and Genetic Algorithm (GA). Best accuracy was observed with a GA-RF of 93.95%.

Saxena et al.;[59] proposed a comparative study using four classifiers MLP, DT, RF and KNN, with Correlation, Information Gain and PCA feature selection techniques. Weka 3.9 was used to run experiments on the PIMA dataset. Hyperparameter optimization and pre-processing techniques were used. They contrasted the results with and without feature selection and discovered that RF has a 79.8 % accuracy rate.

Ismail et al.;[60] Study contain 35 different ML classifiers were tested, both with and without feature selection, and a taxonomy of risk factors for the onset of diabetes was provided. For the purpose of predicting type 2 diabetes, nine feature selection techniques were used. PIMA, UCI, and MIMIC III, three real-world diabetes datasets, were used by the author to assess and compare performance. According to experimental findings, the Bagging-LR classifier is accurate for a balanced dataset (PIMA) with and without feature selection, while RF is accurate for an unbalanced dataset (UCI and MIMIC III). PIMA's execution time is cut by up to 2.5 times, while UCI's execution time is cut by 1.23 times.

K Kannadasan et al.;[61] proposed a Deep Neural Network (DNN) based framework using stacked autoencoders. The proposed framework has experimented on the Pima dataset with 768 records and eight attributes for each record. Fine-tuning is done using a back propagation network. Results showed that with an accuracy of 86.26 %, this model outperforms other models. The model has a precision of 90.66

% and a recall of 87.92 %., specificity value of 83.41% and f1-score value of 89.27%, respectively.

3.4 Survey on Type-2 Diabetes Prediction Using Machine Learning

3.4.1 Type-2 Diabetes Prediction for Stress Parameter

In their review paper, Falco et al.; [62] have described the types of stress based on the physiological responses. According to the authors, a stress reaction is classified into two types. The sympathetic nervous system, which works quickly, controls stress, whereas the neuroendocrine axis known as hypothalamo-pituitary adrenal mechanism has a delayed response to stress-related changes. Cortisol is a hormone that plays an essential role in stress-related effects on metabolism, especially on glucose metabolism. This hormone causes an increase in the level of blood glucose by stimulation of glucose synthesis in the liver and also by inhibition of insulin action. The levels of stress may differ from person to person. The stress level is based on the ability of the person to evaluate life events. Anxious people may anticipate the challenges and amplify the feeling of danger even in everyday situations.

Zamani-Alavijeh et al.;[63] conducted a study to assess the effect of stress management on glucose control. Study contain, 230 patients splitted into two groups. One group was given a stress management training program and the other acted as the control group. In the second group, only standard care was provided. Prior to the intervention, no significant difference in perceived stress scores between the two groups. Stress score was significantly lower after the stress control management programme. Similarly, before the intervention, there was no difference in HbA1c between the two groups. There was a significant reduction in HbA1c in the intervention group after the intervention. Overall, the findings of this study reveals that stress management interventions based on social cognitive theory can help to reduce stress, improve coping ability and improve glycemic control.

Harris et al.;[64] performed a 12-year-long study to understand the link between perceived stress and type 2 diabetes onset. Study include 12,844 women. This study

showed an association between moderate/high-stress levels and 2.3 times higher odds of diabetes after three years. Moreover, there was no significant difference between perceived stress's direct and indirect effects on diabetes. In this study, women with moderate/high levels of perceived stress also had lesser physical activities than without stress.

In a comprehensive review article, Kelly and Mubarak Ismail [65] have discussed the contribution of stress on type 2 diabetes. Chronic exposure to stress factors such as the socioeconomic condition of the person, mental health status and aggressive behavior contribute to the progression of type 2 diabetes. In addition, also suggests the association between stressful work environments or past traumatic events resulting in depression. The link between diabetes and psychosocial factors has been extensively discussed in the literature published across the world.

Another review by Luthra [66] discusses the link between stress and diabetes. Interestingly, this link is bilateral. In patients with diabetes, the presence of stress can affect blood glucose levels by two mechanisms. Firstly, patients with stress may not be able to take care of their physiological requirements. For example, they may eat non-healthy food by a quick fix. Sometimes, they may be at increased risk of consuming alcohol. Many people with stress use food to comfort themselves.

Additionally, anxiety and stress can influence their physical activities. These people tend to ignore physical health and perform less exercise. This can have harmful effects on their blood glucose level and diabetes. The presence of stress may lead to ignorance of checking their blood glucose level, or even they may forget to check. Stress hormones are known to cause alteration in the blood glucose levels by their direct action. These hormones have antagonistic action with insulin, which reduces blood glucose levels.

Llyod [67] and colleagues have also discussed these links between diabetes and stress in their review paper. These authors have also focused on the changes in the behavior with an unhealthy lifestyle. Stressors can increase the risk of smoking which is also linked to increase in their blood glucose and an increased risk of developing diabetes. Psychosocial features can influence these systems, suggesting their impact on the immune system. The authors have discussed the challenges related to the assessment of stress levels, which is a comparatively not well-

developed area. It is essential to identify the life events in these people who act as stressors and design strategies to mitigate the stress response, which in turn can help to decrease the likelihood of developing type 2 diabetes.

Kaur and Bedi [68] conducted a prospective cross-sectional study on people with recently diagnosed hypertension and type 2 diabetes. In this study, they included 80 patients. The results of this study suggested a strong association between exposure of a patient to stress and diabetes. The relative risk was 1.51 with a 95% confidence interval of 1.09-2.10. There was a positive linear relation was also observed between stress exposure and hypertension. The authors of this study have also stressed the relationship between certain hormones and neurotransmitters contributing to the diseases like diabetes and hypertension.

Sui H. et al.;[69] evaluated the relationship between work-related stress and the risk of developing type 2 diabetes mellitus by conducting a systematic review and meta-analysis. The studies published up to September 2014 were selected. Although overall results did not show direct association, a subgroup analysis showed job strain as a one of the risk factor in women. P-value=0.04, the result was significant. The possible mechanism suggested more chance of health impact due to chronic psychosocial stress on psychoneuroendocrine activation in women compared to men. The authors have mentioned the limitations of the analysis. The work-related stress assessment was evaluated based on the self-administered questionnaire, which may lead to some bias.

Brunner et al.;[70] evaluated the effects of social and work-related stressors and resources at work on health-related productivity in Swiss people. This study included a cross-sectional and longitudinal dataset. In this study, the investigators showed that loss of health-related productivity is proportional to the task-related and social job-related stress factors. On the contrary, more availability of the resources at the job and more personal resources decrease such losses. People with higher levels of job stressors and less personal resources have more chances of improved productivity with more resources at the job. Moreover, productivity losses are related to work stress leading to absenteeism from the job.

Bhui et al.;[71] conducted qualitative interviews to understand and categorize the reasons for stress at work. The study included 51 people working in different

organizations. Difficult working conditions and management practices emerged as common causes of work stress among the study participants. Interestingly, the working condition was the leading cause of stress irrespective of the type of work sector they belonged to. The type of job was also a contributing factor to the work-related stress reported by more people from private organizations and NGOs. Especially jobs with high, more unpredictable profiles of work and jobs that needed odd hours or working reported more job-related stress. Some strategies recommended to reduce work-related stress included physical exercise and appropriate planning of work-related tasks.

Kiranashree B.K. et al.;[72] Author introduced a system for stress detection based on physiological factors using machine learning techniques. Generally, stress can be detected using physiological factors and signals. Based on the detailed analysis found that NB, RF and SVM model gives the best accuracy. Based on various measures, it is found that SVM provides an accuracy of 96.67% compared with other classifiers, but SVM works only for a small dataset. The analysis also suggested that for large data ANN model is the most efficient and accurate for stress detection in employees.

Reddy et al.;[73] performed a study by applying ML techniques to analyze stress patterns in 2017 working people from the technology industry. In this study, they also tried to enlist the factors which strongly affect stress levels. Different ML techniques were applied to mental health survey responses to train the model after cleaning and pre-processing the recorded data. Analysis and comparison of other methods showed the highest accuracy with boosting. The use of decision trees showed gender, family history and health benefits at the workplace to be the important factors affecting stress. Higher accuracy of 75.13% was achieved with boosting, whereas the least accuracy of 69.43 was observed with bagging.

Ahuja and Banga [74] performed a student centric study. This study helps to calculate the mental stress of students before examination and spending time on the internet. The dataset used for this study was 206 records from JIIT. The dataset was based on a perceived stress scale (PSS) with 14 questions that should be scaled from 1 to 5 as never, almost never, sometimes, fairly often and very often. Based on this, students were to be categorized as highly stressed, stressed and normal. Four classifiers used for the study are linear regression, NB, RF and SVM. Performance

measures were be done through sensitivity, specificity and accuracy. Experimental result showed that SVM gives a specificity, accuracy, sensitivity value of 100%, 85.71% and 75%, respectively.

Sanchez W. et al.; [75] developed a model to identify job stress in the early stage through a Fitbit sensor, highlighting stress's importance in job-related functions. This methodology was divided into 3 phases: data collection through the labor check app, data pre-processing and model development. The proposed model uses various classifiers such as KNN, ZeroR, RF, NB, J48, and AdaBoost. Feature selection based on correlation, information gain, PCA, wrapper and Spearman correlation were done on 870 records. The evaluation showed that the information gain-RF model showed the best result by achieving an accuracy, recall, precision and f1-measure value of 83% each.

Kumar P.et.al; [76] designed a predictive model for the occurrence of anxiety, depression and stress using DASS42 and DASS21 datasets. Eight classifiers are to be categorized into a probabilistic, neural network, nearest neighbor and tree based. Experimentation was done using the WEKA tool with 75% data for training. Hybrid method works well on both datasets.

Pabreja K.et.al; [77] study is based on working professionals in India and addresses to rectify the factors that trigger stress and predict stress levels. To achieve this collected dataset from 209 participants on a scale of 1 to 5. The question can be categorized as personal, organization specific, work environmental related and job profile specific. Four machine learning algorithms have been experimented and show that the SVM repressor model is the best performance model among all other models by achieving an R-squared value of 0.9164. The study also reveals that among all respondents17.26% have medium stress, 53.8% have high level stress and 28.94% have low level stress.

Patil and shah [78] developed a stress-based model using stress-based and demographic features. The stress prediction model uses three different classifiers: RF, LR and SVM. The experimental result showed that SVM gives the best accuracy compared to other classifiers, with an accuracy of 80.17%.

3.4.2 Type-2 Diabetes Prediction for Lifestyle Parameter

Rajappa T. et al.; [79] conducted a study on diabetes patients' for awareness of diet, exercise, and lifestyle changes in rural populations. In this study, 100 diabetes patients were involved. According to the study, 74 % of participants were aware of food item avoidance, 54 % were aware of food proportion, and 29 % were aware of lifestyle modifications and practises, with diabetes patients making up the remaining 15 %. For control of diabetes, this study highlights the awareness of dietary practices among diabetic patients. This study also reveals a need for awareness activities in both urban and rural areas of India for controlling purposes.

Kolb and Martin [80] discussed an overview of environmental factors for preventing type 2 diabetes. The review points out that factors like diet quality and quantity, physical activity, prolonged screen time or continuous sitting, noise exposure, disturbed sleep, smoking, depression ,stress and low socioeconomic status all raise the risk of developing diabetes. These elements raise body mass index, hence there is a need for some strategies for diabetes prevention and simultaneously improving the resistance of humans.

Deo and Panigrahi [81] study used NHANES US-based dataset of 140 records with 11 predictor variables. Due to data imbalance and missing values on a small dataset, some statistical measures were applied, SMOTE was used for class imbalance, and features were extracted using the PCA method. Classifiers used for the study are linear-SVM and bagged trees. The results are validated using, 5-fold CV and holdout validation shows that SVM gives the highest accuracy and AUC value of 90.82% and 0.91. This study further recommends that validate the model on a large and diverse dataset.

Nai-aruna et al.; [82] designed a system to classify the risk of diabetes mellitus using various classifiers. After removing missing values, the study contains 30,122 records, with 19,145 non-diabetic and 10,977 diabetic populations of 1 year. The study involves 11 features along with the class variable; for the class variable, FBG <100mg/dl is considered a normal group and FBG between 100 to 125mg/dl is regarded as a risk group. Checking robustness of model ROC and accuracy calculated and compared. Results suggest that the RF model is the best performer

for diabetes risk classification. Based on model performance developed a web application for the prediction of the risk of diabetes.

Jahan N. et al.;[83]designed an intelligent diagnosis system for diabetes prediction. To find diabetes risk level (low, medium, high) proposed system contain 555 records with nine features. The proposed approach was implemented using the weka tool and various models considered for the study are MLP, DT and IBK algorithms. The result shows that IBK gives the best classification accuracy of 98.73% for 12-fold cross-validation. The author also developed mobile app for awareness purpose.

Olivera. A.R. et al.;[84] designed a predictive model for detecting undiagnosed diabetes based on an ELSA-based dataset from Brasil. The study is based on both clinical examination, interview and laboratory tests. The designed model is 4 step process with various models involved in the study: LR, ANN, NB, KNN and RF. The wrapper strategy is used to generate a subset of variables. They identified that all predictive models produced similar results. The highest AUC value is achieved by an LR of 74.41%.

Ayush and Divya [85] designed a predictive model based on personal indicators for establishing a relationship between lifestyle activities and the risk of diabetes. Lifestyle activities include sleeping habits, eating habits, physical activities, BMI, and waist circumference. The collected dataset contains a total of 180 records. According to the author's CART model, based on lifestyle has accuracy rate of 75%, blood pressure is a determinant of the risk of developing diabetes, hence it is necessary to do some measures in daily lifestyle patterns.

A diabetes mellitus disease prediction model was created by Tigga and Garg [86]. A dataset was gathered for the investigation comprising 952 instances and 18 attributes related to lifestyle, health and family background. The comparative analysis purpose study also uses a PIMA dataset of 768 records with nine features. RStudio was used for implementation and the result shows that the highest accuracy was obtained for RF, with a value of 94.10% for the data that was collected and 75% for the PIMA dataset. essential parameters are found to be age, family history of diabetes, physically active, regular intake of medicine and gestation diabetes for the risk of diabetes.

Wang et al.; [87] designed an XGBoost based ensemble model for predicting the risk of diabetes. To conduct this study, collected questionnaire survey from china containing 380 middle-aged and older individuals personal information, diet, exercise, and genetic disorder. To improve the prediction effect, compared the model with SVM, KNN and RF algorithms. The result shows that XGBoost had the best accuracy of 89.09%, with an AUC value of 0.9182.

3.5 Impact of COVID-19 Related Stress on Blood Glucose Level for Type-2 Diabetes patients

Marcovecchio and Chiarelli [88] demonstrate that stress has a negative impact on diabetic patients' ability to control their blood sugar levels. In patients with diabetes, both direct and indirect effects contribute to glycaemic impairment. The stress hormones have direct effects, and changes in lifestyle and behavior have indirect effects.

In addition to more conventional stressors, the coronavirus disease (COVID-19), become a significant source of stress for individuals, their families, and the community over the past year. Following COVID-19 infection, poor outcomes are anticipated for diabetes.Schofield J.et.al;[89], Explored the association between severity of illness and blood glucose and the impact of the pandemic on diabetes.

Hussain A.et.al;[90] review on characteristics of coronavirus (SARS-CoV-2) and (COVID-19) in people with diabetes and management. Diabetes has emerged as a significant risk contributers for the emergence of complications in patients with COVID-19 and the interaction between COVID-19 and diabetes is still in progress.

Jonensen et al.;[91] performed a questionnaire based study involving 2430 adult participants to understand psychological concerns among patients with diabetes. Analysis performed by logistic regressions revealed several risk factors for concerns related to COVID and diabetes. These factors included female gender, presence of type 1 diabetes, complications, feeling lonely, and changed behaviours being more concerned about diabetes and COVID-19.

A cross-sectional study by Sankar et al.; [92] During the COVID-19 lockdown, people from South India reported having unhealthy eating habits, experiencing

stress, and having trouble sleeping. However, the same study found no appreciable variation in overall glycaemic control between type 2 diabetes patients during the lockdown. According to the study, the COVID-19 lockdown had no appreciable impact on overall glycaemic control in patients with type 2 diabetes.

Ruissen M.et.al;[93] conducted a study to identify how the COVID-19 pandemic affected each person's life and led to significant lifestyle changes. According to a study involving 435 patients, people with diabetes reported feeling more stress and anxiety. The study has reported an association between increased stress with difficulty in glycemic control.

Kautzky-Willer [94] suggests that there is a need for a study focusing on sex and gender differences among diabetes groups. In general, male patients are more likely than female patients to develop COVID-19 disease complications. Similar to this, hospitalized diabetic patients with COVID-19 show a male bias.

Yoshida et al.; [95] conducted a study to assess how gender affects clinical outcomes and presentations in COVID-19 hospitalized patients. According to their study, women had a significantly higher prevalence of diabetes than men (61.4%).

Dasappa et al.; [96] performed a study in urban slums of Bangalore for age group 35 and above in 2013 for several risk factors. Observation showed that diabetes and prediabetes prevalence was 12.33% and 11.57%, respectively and also reported a higher prevalence of diabetes in females compared to males. The study also suggested recommendations for controlling diabetes.

Hilliard et al.; [97] synthesized the effect of stress and HbA1c on diabetic people. Neuroendocrine changes and the dysregulation of physiological systems can result from prolonged stress. Patients who have diabetes may experience either acute or chronic stress, depending on how long the stressors last. Due to its brief duration, acute stress may have little impact on HbA1c, a measure of glucose control over a period of several weeks.

Peric and Stulnig [98] did a review on diabetes and COVID-19. As per observation, patients with diabetes having some comorbidities have to take more preventive measures to avoid COVID-19 infections. Suboptimal glycaemic control during infectious diseases has to maintain and suggested to implement some strategies to prevent disease complications.

Tanji et al.; [99] performed a retrospective study from Japan and found that HbA1c levels significantly increased after the COVID-19 outbreak compared to their levels prior to the pandemic. The patient's physical and mental health have changed over this time period. The level of HbA1c in these patients may change due to behavioral changes. Study involves patients older than 65 years showed a significant rise in HbA1c when compared by age. Additionally, patients with a BMI of more than 25 kg/m2 had significantly higher HbA1c levels than those with a lower BMI.

Faulenbach et al.;[100] evaluated how acute stress affected the glycemic control of 30 T2DM patients with age of 60 years. Study involved Trier social stress test for measuring psychological stress among type 2 diabetes patients. In this study, postprandial blood glucose levels significantly increased when postmeal stress was experienced.

Vasanth et al.; [101] performed a cross-sectional study of Chennai. The study involves 400 people of age more than 30. A perceived stress scale is used to examine stress levels. The study reveals that According to research from India, stress levels and FBS and PPBG levels are positively correlated.

Singh et al. [102] suggested that Diabetes has also been identified as a risk factor for COVID-19 patient mortality and severity, hence the need to control blood glucose levels. Telemedicine treatment was also suggested and Wong et al.; [103] showed stress and anxiety are responsible for the development of diabetes.

Salari et al.; [104] showed Stress may further exacerbate the glycemic control problem.COVID-19 pandemic increases the psychological burden among different communities. Therefore, during the COVID-19 pandemic, it is crucial to address the psychological problems of vulnerable groups. Despite making lifestyle changes, patients with diabetes may become frustrated by their experience with hyperglycemia. [105].

Erener[106] performed an exact cause and effect relationship between stress and glycemic parameters cannot be established because this research was retrospective. Another restriction is a single-center study with a minimum data size. COVID-19 might be a factor in the emergence of hyperglycemia.

3.6 Research Gap and Findings

The previous literature showed that the general research trend for prediction of type 2 diabetes was on the PIMA dataset. Moreover, the number of features used in most of the available literature is minimal, as it does not include all the factors for the risk of development of diabetes. Moreover, there is an explicit limitation when it comes to the feature included in the study and the nature of these features. Overall research gaps are listed below.

- Most of the studies work on PIMA Indian diabetes dataset.
- PIMA dataset is only female-centric.
- Limited study on stress as a risk factor and that to be using only DASS scale.
- Limited studies on lifestyle parameters at the ground level.
- Improvement in prediction model.
- Even the latest literature also suggests persistence in the gap.
- Integration of optimization techniques for improving prediction model.
- Very few studies used optimization techniques.
- Very few studies used the application of various feature selection techniques for the development of a model.
- Lack of Nationwide framework, especially in the case of India and other developing nations.
- The lack of adoption of novel technologies can help analyze the root cause of disease at an early stage.
- Scope of improvement in the performance of various ML algorithms for early diagnosis of diabetes in terms of prediction accuracy and worthy of further research for improvement.
- Limited Studies on auto encoder-based feature extraction.
- Prediction using a single algorithm is insufficient and the selected dataset may not fit that algorithm, leading to less accuracy.

3.7 Summary

This chapter reveals a review on machine learning based type 2 diabetes prediction. It also focuses on recent developments in Machine Learning which have significantly impacted the prediction of diabetes. This review gives a feature

selection based on ongoing research and focuses on reviews elaborating stress and diabetes as well as lifestyle patterns and diabetes. This study also reveals the relation between stress and diabetes and the impact of COVID-19 on type 2 diabetes patients. This review helps to understand the prediction of diabetes mellitus limitations by using various approaches. To overcome the issues and to provide a competent model for the development of risk of type 2 diabetes based prediction model, gaps were identified, which are explained in further chapters. It is necessary to introduce an auto encoder-based ensemble model for the efficient classification of diabetes mellitus.

CHAPTER-4

PROPOSED FRAMEWORK FOR ANALYZING IMPACT OF VARIOUS PARAMETERS ON TYPE 2 DIABETES PREDICTION

CHAPTER- 4 **PROPOSED FRAMEWORK FOR ANALYZING IMPACT OF VARIOUS PARAMETERS ON TYPE 2 DIABETES PREDICTION**

4.1 Introduction

Diabetes mellitus is now regarded as leading causes of death globally. Diabetes mellitus is the most common endocrine disease influenced by stress and lifestyle factors such as drinking, exercise, diet, and others. According to the review, lifestyle modification and medication interventions can help prevent complications associated with the diabetes. To propel early disease detection, machine learning is a technique that allows the training and testing of classification systems with available datasets. Automatic learning has fetched greater interest in the healthcare sector; rather than wasting time on detection utilizing time for patient care. The main focus of this research is the use of machine learning methods that combine feature selection for the early prediction of type 2 diabetes based on personal indicators. This chapter presents a proposed framework for analyzing the impact of various parameters on developing a type 2 diabetes risk predictive model. The proposed methodology is divided into nine sections. Section 4.2 elaborates on the data collection process, which emphasizes various datasets for study and validation purpose. 4.3. Discussed proposed research framework followed by various performance metrics mentioned in section 4.4. Section 4.5 explains the methodology adopted for stress parameter Section 4.6 COVI_stress methodology. Section 4.7 describes the methodology used for lifestyle parameters. Section 4.8 discussed mathematical model is for D_Stress and Diabe_Lifestyle methodology. Section 4.9 Concludes the chapter.

4.2 Data Collection Process

To maintain the integrity of research, accurate data collection is necessary. To examine the efficacy of ML algorithms at the earlier stages of predicting the risk of diabetes, collected datasets basically from three sources: one data set is a questionnaire based dataset collected through a web-based approach. Another dataset is collected from MGM Hospital Aurangabad, which contains hospitalized diabetes patients during a pandemic and the other two datasets are from the UCI repository and from the researcher. The following datasets were used in this study are,

1. Self-collected questionnaire based dataset – LS_diabetes dataset (Appendix III A)
2. Hospitalized diabetes patient dataset –COVI_Stress dataset (Appendix III B)
3. UCI repository diabetes dataset - PIMA Indian diabetes dataset (Appendix III C)
4. Kaggle Source/ Collected from researcher - Diabetes _2019 dataset (Appendix III D)

Self-collected questionnaire based dataset for the study:

The dataset was developed by web-based questionnaires. LS_diabetes dataset comprised 374 people with 35 features; the questions were related to demographic information, dietary pattern, lifestyle, and pathological and stress related factors.

To estimate the effect of COVID-19 related stress on blood glucose of type 2 diabetes patients, collected COVI_Stress dataset comprising 199 hospitalized diabetes patients with 27 different questions was collected.

4.2.1 Development of Machine Learning Model Using LS_Diabetes Dataset

LS_Diabetes Dataset:

Data was collected using a questionnaire-based survey. The questionnaire was designed and developed in consultation with a qualified healthcare professional. The developed questionnaire was validated and pre-tested before using on an extensive sample data set. The questionnaire used in the dataset used for research was administered web based as well as personally through face-to-face interviews. It has

374 instances collected from 35 features. Below, Table (4.1) describes the characteristics of the LS_diabetes dataset.

Table 4.1: LS_diabetes dataset characteristics

Dataset	Training data records	Testing data records	Total no. records	Total No. of attributes	Output class
LS_diabetes	299	75	374	35	2

Class wise distribution of the dataset based mentioned in Table (4.2).

Table 4.2: Class wise LS_diabetes dataset distribution

Dataset	Diabetic	Non-Diabetic	Total Number of Records
LS_diabetes dataset	87	287	374

The features are a combination of demographic, stress related, lifestyle related and other diagnostic results. The dataset consists of both numerical and categorical values from the questionnaire answers. Collected data will be compiled for analysis shown below in figure (4.1)

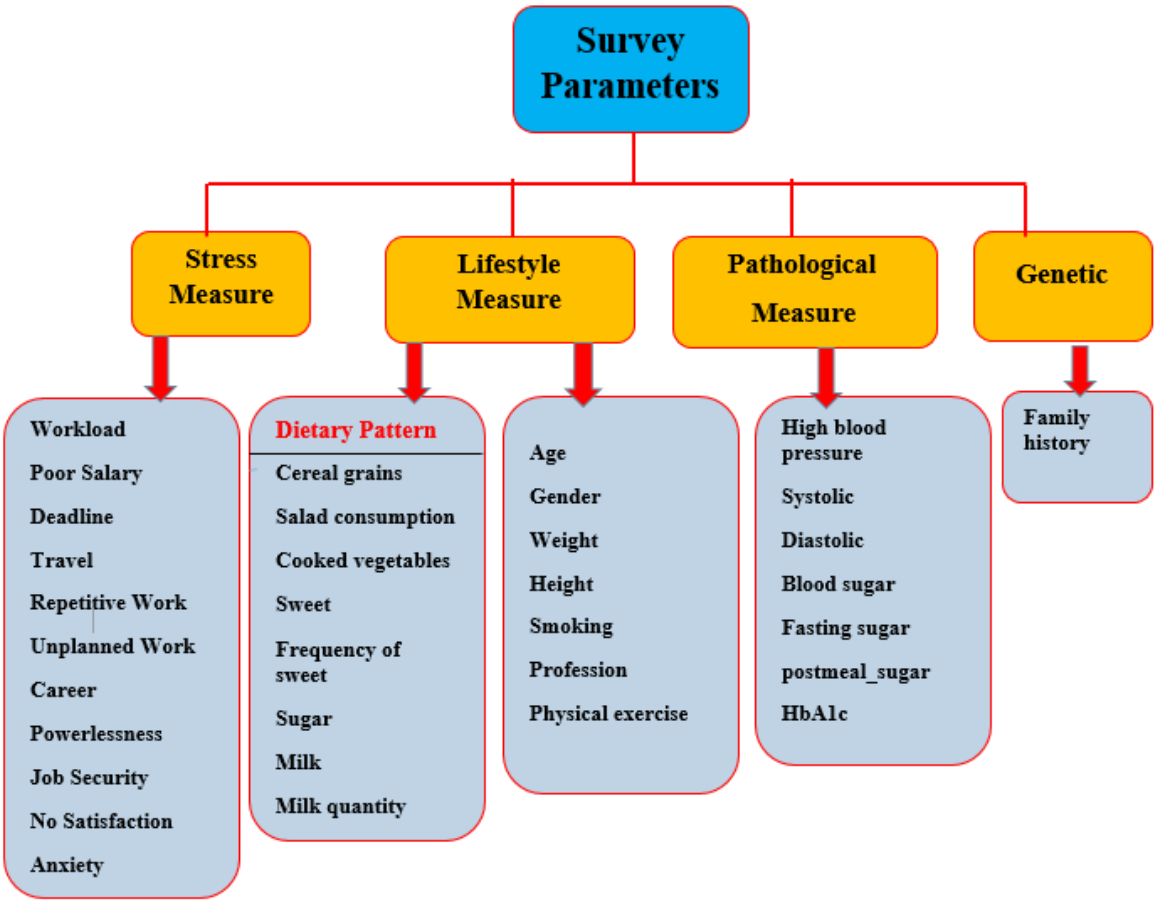


Figure 4.1: LS_diabetes dataset parameters

4.2.2 Development of Model Using COVI_Stress Dataset

Due to COVID-19 human life is adversely affected, including physical health, psychological and emotional health, education, travel and economic aspects. To find the impact of stress on blood sugar levels during the COVID-19 Period, we collected a dataset of patients having type 2 diabetes on anti-diabetes medications. This study is a single-center questionnaire based study from Maharashtra, with patients of both genders having diabetes and receiving anti-diabetes medications with a history of hospitalization for the treatment of COVID-19 from May 2020 to July 2020, shown in Table (4.3). Patients with available investigations related to glycemic parameters (current and previous three months) and whose readings for glycemic parameters, i.e., FBG level, PPBG level, were available. The research included a current or past history of asymptomatic or symptomatic COVID-19 infection. Type 1 diabetes patients, newly diagnosed cases of T2DM not receiving any anti-diabetic medication, patients with T2DM with no history of hospitalization for COVID-19 treatment and patients with gestational diabetes were excluded.

Table 4.3: Characteristics of COVI_Stress dataset

Dataset	Total no. records	Total No. of attributes
COVI_Stress	199	25

The records of patients in diabetic groups were screened for the following parameters shown in Table (4.4).

Table 4.4: Attribute description of COVI_Stress dataset

Features	Description
1. Age	Age of a person
2. Gender	Male or Female
3. Weight	Weight in kg
4. Height	Height in cm
5. BMI	Body mass index
6. Residence	Specifies the residence as Rural/Urban
7. Family status	Specifies Single/married/widow
8. Profession	Service/Business/Retired/ Student/Housewife/Others
9. Type 2 diabetes	Since (years)
10. On treatment	Since (years)
11. Current HbA1c%	Specifies HbA1c value
12. Current FBG mg/dl	Specifies current FBG value
13. Current PPG mg/dl	Specifies current PPG value
14. Did COVID-19 affect glycemic control	Yes/No
15. HbA1c % before COVID-19	Specifies past HbA1c value
16. FBG mg/dl before COVID-19	Specifies past FBG value

17. PPG mg/dl before COVID-19	Specifies past PPG value
18. Did COVID-19 increase psychological stress on the patient?	Yes/No
19. Was there major change in diet during COVID-19 pandemic?	Yes/No
20. Was physical activity decreased during COVID-19 pandemic	Yes/No
21. Was treatment for diabetes taken regularly during COVID-19 pandemic?	Yes/No
22. What causes you stress in your profile? 1]I found it hard to wind down 2]I tended to over-react to situations 3]I found myself getting agitated 4]I was intolerant of anything that kept me from getting on with what I was doing	Never , Sometimes, Often, Almost always

4.2.3 Development of Machine Learning Model to Test on PIMA Indian Diabetes Dataset

PIMA Diabetes Dataset: Standard Dataset:

The analysis and prediction of diabetes are performed on the Pima Indian diabetes dataset which was downloaded from the UC Irvine machine learning repository, consist of 768 records with a female aged 21 and above [107]. This dataset is female-centric only. The dataset comprises of numerically-valued eight attributes with the class label as diabetic or non-diabetic. Data contain both medical examination data as well as personal health data. It includes 268 diabetic instances and 500 non-diabetic instances shown in Table (4.5).

Table 4.5: PIMA Indian diabetes dataset characteristics

Dataset	Diabetic	Non-diabetic	Total no. records	Input attributes	Class Variable	Total No. of attributes
PIMA	268	500	768	8	2	9

Age, Body mass index(BMI), Diastolic blood pressure, Number of times pregnant, 2-h serum insulin, Plasma glucose concentration at 2 h in an oral glucose tolerance test, Skin fold thickness, Pedigree function and target class is shown in below Table (4.6).

Table 4.6: Attribute description of PIMA Indian diabetes dataset

Sr.No.	Features	Description
1	Preg	Specifies count of pregnancy
2	Plas	Glucose Concentration in an oral glucose tolerance test
3	DPres	Diastolic blood pressure
4	Skin_T	Specifies skin fold thickness
5	Insul	Specifies 2-Hour serum insulin
6	BMI	Body mass index
7	Pedig	Specifies diabetes pedigree function
8	Age	Individual age
9	Class	Labelled as positive or negative

4.2.4 Development of Machine Learning Model to Test on Diabetes_2019 dataset

Diabetes_2019 Dataset:

Another research dataset collected for analysis and prediction purpose was self-prepared [86]. The questionnaire was designed based on constraints that could lead to the risk of the development of diabetes. The dataset comprises 952 participants

aged 18 and above; among them372 are females and 580 are males. It contains both 267 diabetic instances and 685 non-diabetic instances. The dataset comprises numeric-valued 18 attributes out of that 17 independent predictor variables and one outcome or variable mentioned in Table (4.7).

Table 4.7: Characteristics of Diabetes_2019 dataset

Dataset	Diabetic	Non-diabetic	Total no. records	Input attributes	Class Variable	Total No. of attributes
Diabetes_2019	267	685	952	17	2	18

Age, Gender, BMI, Family history, High blood pressure, Blood pressure level, Walk/run/physically active, Smoking, Alcohol consumption, Sound sleep hours, sleep hours, Daily medicine intake, Junk food intake, Stress, Pregnancies count, Gestation diabetes, Frequency of urination, Diabetic. All these features are listed in Table (4.8)

Table 4.8: Attribute specification of Diabetes_2019 dataset

Sr.No	Features	Description
1	Age	Age of an individual 18 or above
2	Gender	Male/Female
3	Family history	Yes/No
4	High blood pressure	Yes/No
5	Walk/run/physically active	None / Less than 30 min / More than 30 min / 01 hour or more
6	BMI	Body mass index
7	Smoking	Yes/No
8	Alcohol consumption	Yes/No
9	Sleep hours	Specifies sleep hours
10	Sound sleep hours	Specifies hours of sound sleep

11	Daily medicine intake	Yes/No
12	Junk food intake	Yes/No
13	Stress	Specifies Not at all / Sometimes /Often /Always
14	Blood pressure level	Specifies low/ normal/high
15	Number of pregnancies	Specifies the number of times pregnancies
16	Gestation diabetes	Yes/No
17	Urination frequency	Specifies Not much / Quite much
18	Diabetic	Yes/No

4.3 Proposed Research Framework

The Proposed Methodology will covered extensive work on two levels: data collection and the development of predictive models on the collected data. Figure (4.2) shows model development divided into four phases: Data collection, Predictive model directing, Model construction and lastly Model validation.

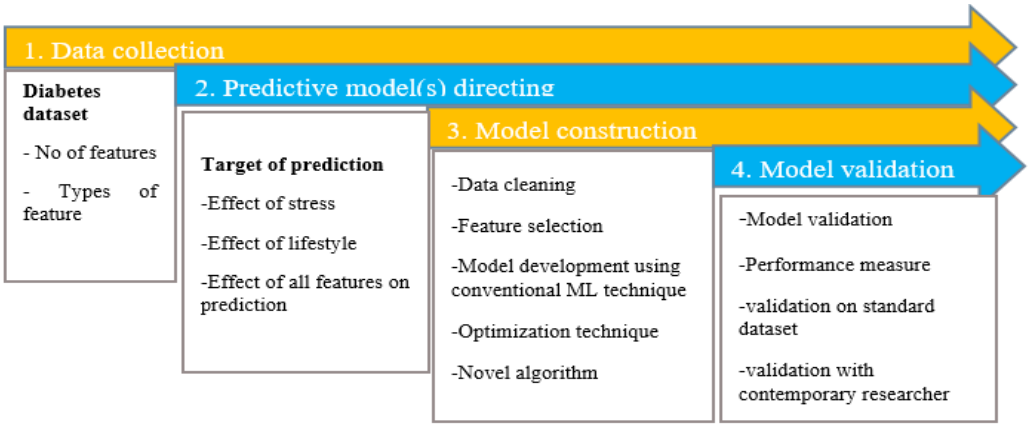


Figure 4.2: Flow of model development

The research methodology works on the LS_diabetes dataset, which contains both populations i.e., Diabetic and non-diabetic. The detailed research framework is shown in below figure (4.3).

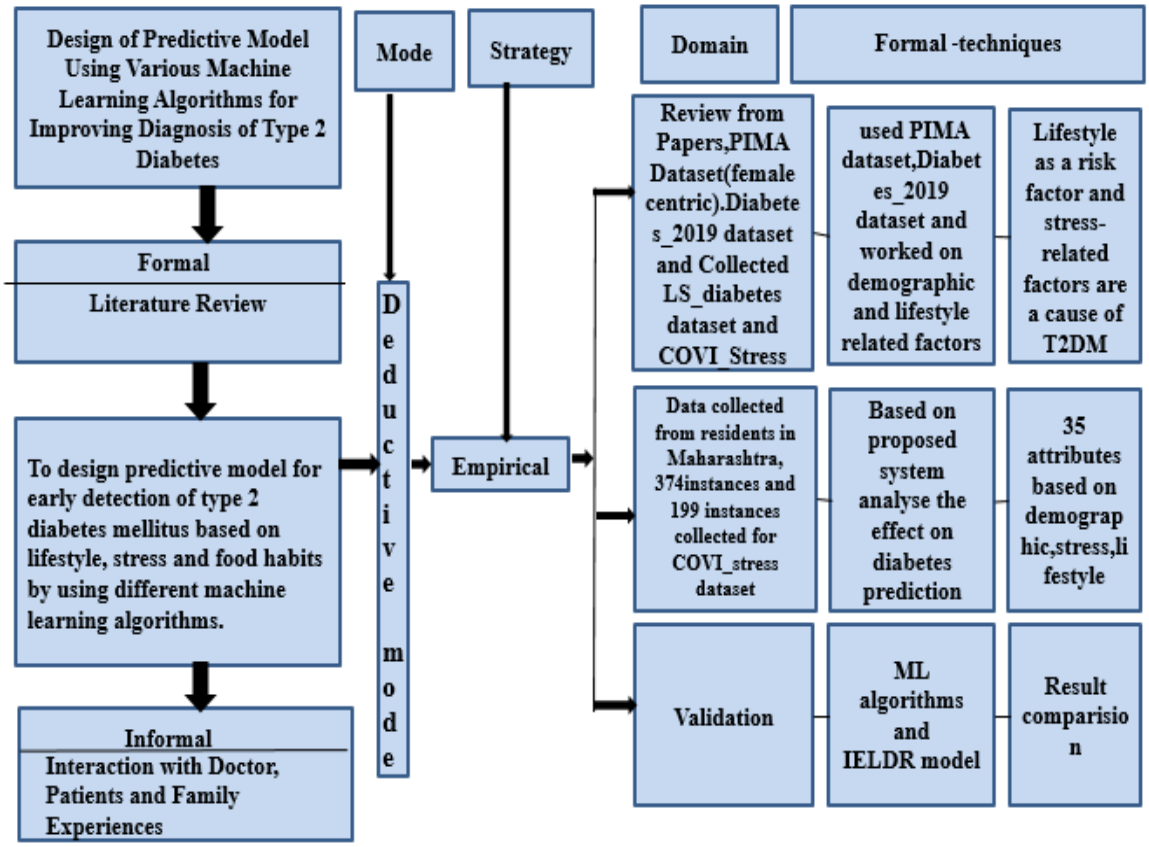


Figure 4.3: Research framework

The methodology adopted for phase I and phase III has been summarized as, after the data collection process, pre-processing of data was done. This data were divided into train and test data. Various classification methods such as gradient boosting classifier (GBC), K-nearest neighbour (KNN), logistic regression (LR), random forest (RF), decision tree (DT) and support vector machine (SVM) were applied. By applying a suitable model, both accuracy and cross-validation accuracy of different models was calculated. Comparative analysis was made using different classifiers. Applying different feature selection methods and optimization using Grid search performance of the model performance were measured. Overall methodology steps are summarized as follows,

1. Experiments have been conducted on LS_diabetes self-collected dataset.
2. Pre-processing steps apply to LS_diabetes datasets.

3. Utilization of dimensionality reduction techniques for diabetes prediction.
4. use of machine learning algorithms viz. RF, LR, SVM, GBC, DT and KNN have been proposed for the prediction of the risk of diabetes.
5. Utilization of tuning of parameters using grid search techniques for diabetes prediction.
6. Performance evaluators such as classification accuracy, precision, recall, f1-score and AUC-ROC value have been considered for testing the proposed methodology's efficiency.

Phase III study finds the impact of COVID-related stress in diabetes people. Statistical analysis methods and the use of the DASS-12 scale specify the impact of stress on blood glucose levels with the help of the COVI_stress dataset. Phase IV designed an autoencoder and ensemble based novel framework. The methodology is classified into four phases shown in figure (4.4).

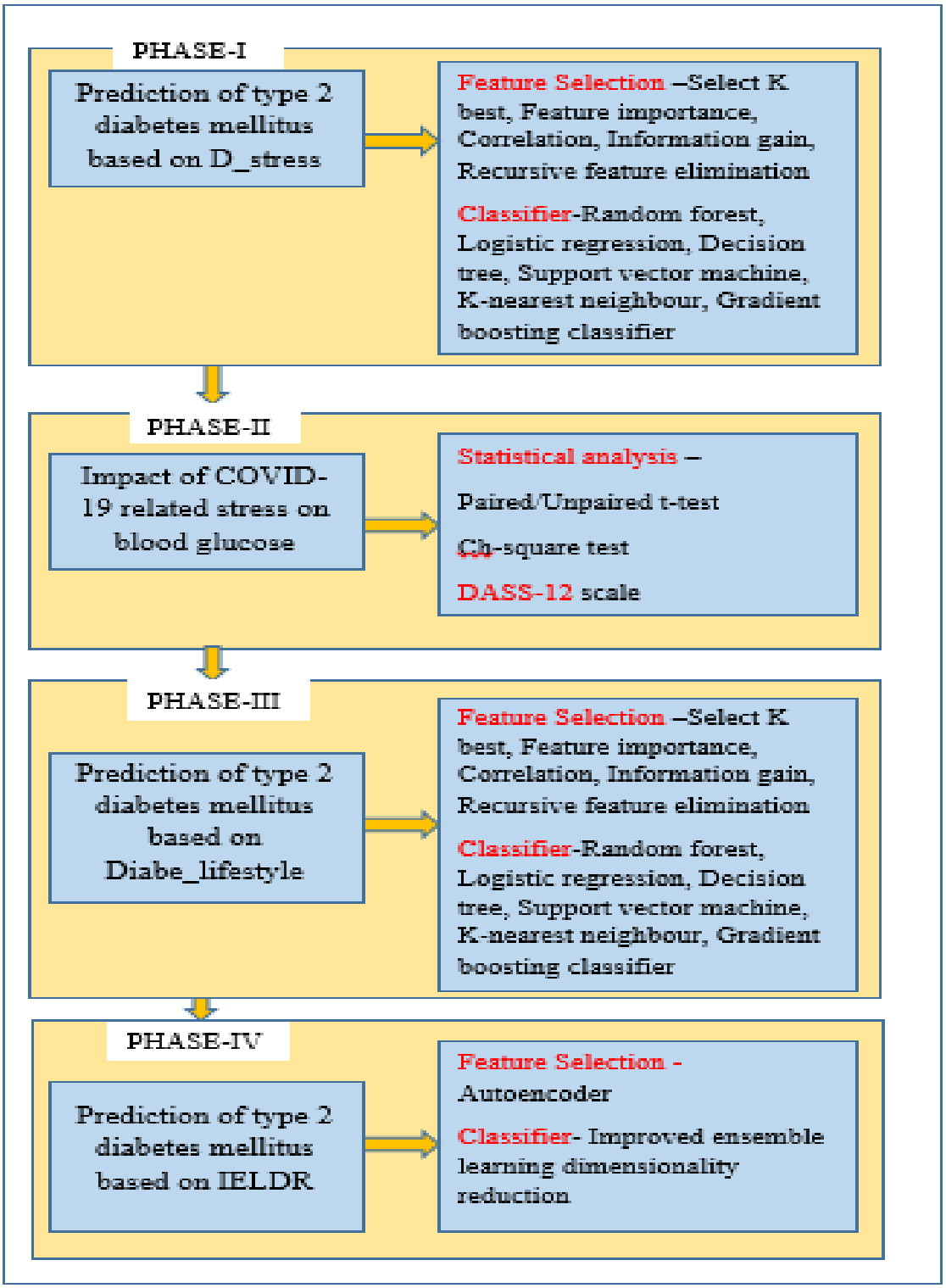


Figure 4.4: Proposed research methodologies

4.4 Performance Metrics

Model evaluation is a technique used to understand the ML model performance. It is the process that quantifies the quality of a system's prediction. Any error in the representation of data van degrades the model's performance. After model construction, model performance can be evaluated based on its correctness. A confusion matrix serves as the foundation for the classification metrics. A confusion matrix is the result of the prediction model. It is a 2x2 contingency table that contains rows and columns for binary classification problems, as shown in Table (4.9). A predicted class is represented on the X-axis, and the actual class is represented on the Y-axis. In this case, TP stands for True Positive, TN for True Negative, FP for False Positive and FN for False Negative. FN and FP are the prediction errors.

Table 4.9. Confusion Matrix

		Predicted Class	
Actual Class	Total Population	Positive Prediction	Negative Prediction
	Diabetes	True Positive (TP)	False Negative (FN)
	Non-diabetes	False Positive (FP)	True Negative (TN)

Where,

- True positive (Tp) –Tp refers to the proportion of patients who are both diagnosed with and likely to develop diabetes.
- False positive (Fp) –FP refers to those individuals who are predicted to be healthy but who actually have diabetes.
- True negative (Tn) –TN refers to the number of patients who are predicted to be healthy and do not have diabetes.
- False negative (Fn) –FN refers to the proportion of patients who are mistakenly diagnosed with diabetes despite not having the disease.

Accuracy, precision, sensitivity, specificity, f1-score, mcc, and area under the curve (AUC) were all used to evaluate the performance of a model. Accuracy is the ratio of the total number of correct predictions to all predictions as shown in Eq. (4.1). precision, i.e., Positive Predictive Value is the number of selected samples being pertinent as in Eq. (4.2). Sensitivity/Recall is the number of relevant samples being pertinent as in Eq. (4.3). Specificity defines the rate of real negative samples can be detected correctly as in Eq. (4.4). Overall effectiveness of the machine learning classifier assessed through accuracy. Estimation of model performance on different classes can be specified through specificity and sensitivity measure. The f1-score is the harmonic mean of precision and recall as in Eq. (4.5) when f1-score is one considered the best value. The error rate is given as in Eq. (4.6). These measures are derived using the following equations.

$$Accuracy = \frac{Tp + Tn}{\mathrm{Tn + Tp + Fp + Fn}} \tag{4.1}$$

$$Precision = \frac{Tp}{Tp+Fp} \tag{4.2}$$

$$Recall/Sensitivity = \frac{Tp}{Tp+Fn} \tag{4.3}$$

$$Specificity = \frac{Tn}{Tn+Fp} \tag{4.4}$$

$$F1score = 2 \times \frac{Precision*Recall}{Precision+Recall} \tag{4.5}$$

$$Error\ Rate = \frac{Fn+Fp}{Tp+Tn+Fn+Fp} = (1 - Accuracy) \tag{4.6}$$

Matthews Correlation Coefficient (MCC) is the correlation coefficient of predicted result and the corresponding ground truth. It shows the classifier predictability with values between (-1 to +1). If a value is -1, the classifier generates wrong predictions and if the value is +1means ideal prediction, and if MCC is 0, then the classifier produces the wrong prediction as in Eq. (4.7).

$$MCC=\frac{(Tp\ X\ Tn-Fp\ \ X\ Fn)}{\sqrt{(Tp+Fp)(Tn+Fp)(Tp+Fn)(Tn+Fn)}}X\ 100\% \tag{4.7}$$

Receiver operating characteristic-Area under curve (ROC-AUC) is a graphical tool used to compare true positive rate and false positive rate. X-axis represents sensitivity and Y- axis indicates specificity as in Eq. (4.8). AUC characterizes the ROC of the model. A classifier performs well if the AUC value is between 0.5 to 1 and below 0.5 model shows weak performance. Below is the figure (4.5).

$$ROC = \frac{Specificity+Sensitivity}{2} \tag{4.8}$$

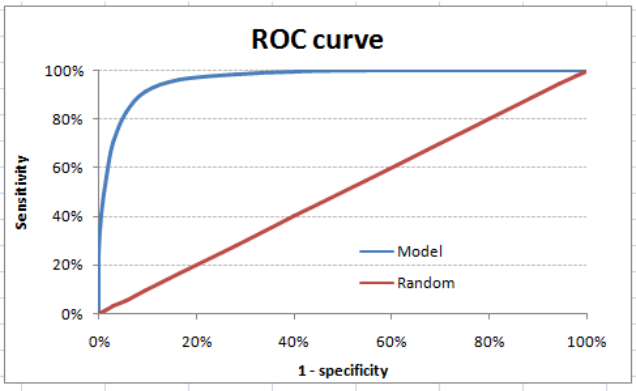


Figure 4.5: ROC curve

(Source: analyticsvidhya. com)

4.5 Impact of Stress Parameter on Type-2 Diabetes Prediction

4.5.1 Description of Stress Data

This study used a pre-designed, pre-validated questionnaire to collect data from the general adult population over the age of 18. Gender, age, weight, height, BMI, workload, deadline, poor salary, unplanned work, repetitive work, travel, career, job security, powerlessness, no satisfaction and anxiety with a outcome variable were among the items on the questionnaire. Among these features, four variables are numerical variables (age, weight, height, BMI) and the rest of the other variables are categorical variables [78]. A total of 16 features are considered stress parameters with one class label. The dataset consists of a total of 374 records of both diabetes and non-diabetes populations.

Table 4.10. D_Stress methodology dataset description

Sr.No.	Attribute	Description
1	Age	Individual age
2	Weight	Weight (in kg)
3	Height	Height (in cm/m)
4	BMI	Body Mass Index (Kg/m2)
5	Anxiety	Yes/No
6-16	Stress	Workload/ Poor Salary / Deadline / Travel/ Unplanned Work/Repetitive Work/ Powerlessness/ Career/ Job Security/No Satisfaction Specified as, Very high /high /average /low /very low
17	Class	1- Diabetic/ 0- Nondiabetiv

4.5.2 Data Preprocessing

As raw data is prone to missing values noise, and inconsistency, making it consistent need to be pre-processed, reflecting the data quality and thereby providing reliable and accurate results. Data cleaning, transformation, integration and reduced data size are different data pre-processing steps. In our research to analyze the final scores, the dataset was pre-processed. As machine learning can understand only numerical data and in our dataset, most of the attributes are categorical features, label encoding is adopted for conversion.

4.5.3 Optimal Feature Selection Using Preferred Techniques

Feature selection is an important step for reducing the dimensionality of data. It is the process of choosing a subset of features that contribute to the outcomes.

Classification accuracy also depends upon the subset of extracted features extracted from the dataset. In order to improve classification accuracy, ML approach in combination with feature selection techniques were used. For selecting relevant features from the dataset, the proposed methodology used five techniques of feature selection, namely, Select k best, Feature importance, Correlation, Information gain, recursive feature elimination (RFE).our research reveals the fact that model accuracy is significantly improved by applying feature selection technique on the dataset.

4.5.4 Diabetes Prediction for Stress Parameter Using Preferred Classifier

After the pre-processing step, the entire dataset was divided into training data and testing data in the ratio of 80:20. The training set is used to train each model and testing is used to evaluate the model performance. The trained model learns the patterns from the dataset and is used to predict new data for presence or absence of disease. Due to a sedentary lifestyle, early detection of diabetes is required hence developing a predictive model. ML classifiers such as LR, SVM, DT, RF, GBC and KNN are applied for model training. The flow of D_Stress methodology and the corresponding algorithm is shown in Figures (4.6) and Figure (4.7), respectively. The work of research objective one is divided into three different sections as follows:

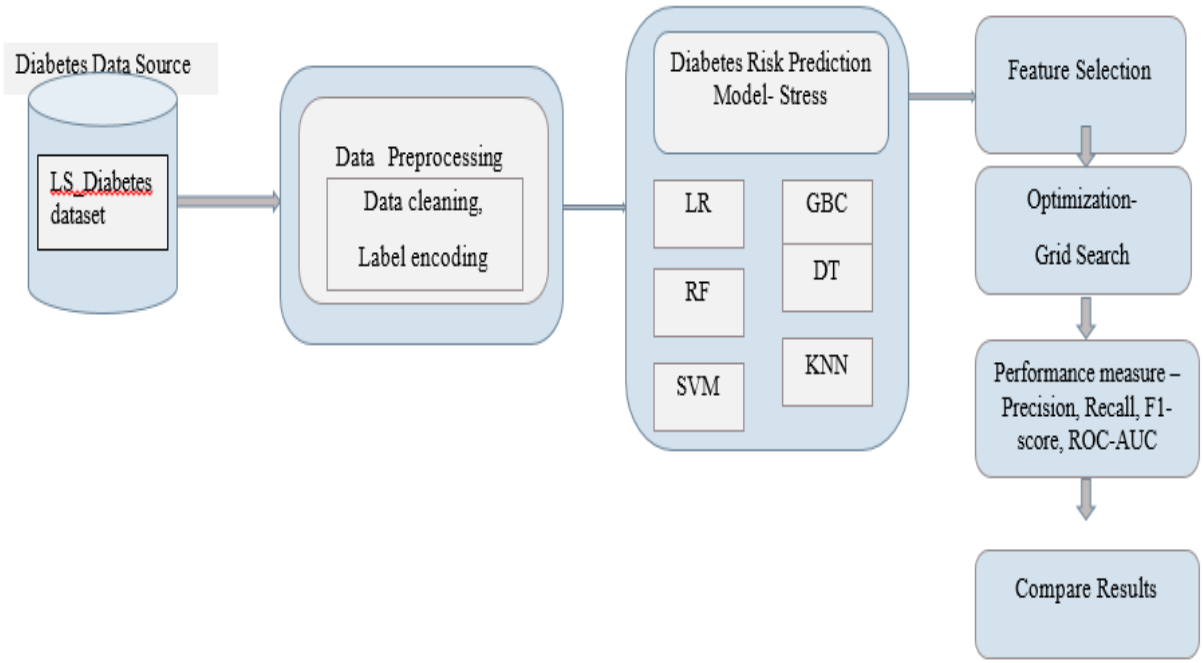


Figure 4.6: A proposed methodology for D_Stress

Data pre-processing was completed in phase I. All classifiers were trained on 80% of the data and tested on 20% of data. Training and testing accuracy was measured for each classifier. To increase the model performance, a 10-fold cross-validation technique was used and model accuracy was calculated.

The study used feature selection techniques in phase II, including select-K best method, feature importance, information gain, correlation and RFE. For each classifier best selected features provided the highest accuracy using all five feature selection methods were recorded. 10-fold was used to improve model performance. The results of classifier evaluation were compared with and without the inclusion of feature selection.

In phase III, all selected classifiers' models were optimized using the grid search method. The best hyperparameters were chosen and ranked. based on their different measure and accordingly, the effective model was determined for the risk of development of type 2 diabetes. The performance of the various ML algorithms was analyzed. Model performance was calculated using metrics such as precision, recall, f1-score and accuracy.

4.6 Impact of COVID-19 Related Stress on Blood Glucose Level for Type2 Diabetes Patients

Diabetes already puts people at risk for psychological problems, and people with severe mental illness are more likely to have diabetes than the general population. Pandemics and natural disasters invariably have profound and long-lasting psychological consequences. Coronavirus disease (COVID-19) has emerged as a significant stressor on people, families, and the community in the last year [90]. COVID-19 is a rapidly spreading viral disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-COV-2). The COVID-19 outbreak began in December 2019 in Hubei Province, China. Since then, the virus has spread rapidly to a number of countries around the world, including India, in a relatively short period of time. Diabetes patients are concerned about COVID-19-related complications. More than half were concerned about being overly affected by diabetes if infected with COVID-19, and about one-third were concerned about being classified as a risk group for diabetes and being unable to manage diabetes if infected. Psychological responses are especially strong in high-risk individuals [91].

With the dramatic increase in confirmed COVID-19 cases and the rising death toll in India, timely and effective management of severe patients, particularly those with diabetes, is critical. However, little attention has been paid to clinical characteristics and outcomes of patients, data on which are limited but critical to lowering mortality. The study explored the effect of perceived stress on blood glucose levels in diabetes patients discharged from the hospital after their COVID-19 treatment.

4.6.1 Proposed COVI- Stress Methodology

To evaluate and compare glycemic control among hospitalized COVID-19 positive diabetes patients. In view of the problem selected for the study and the objectives to be accomplished, combined demographic observations and investigation approaches were considered to be appropriate. The research design selected for this study was conducted in MGM hospital, Aurangabad, a tertiary care center.

To accomplish this objective, data were collected from the medical record section of MGM Medical College & Hospital Aurangabad [MS], India. The duration of the study was one month admitted, from 1st May 2020 to 31st July 2020 will be analyzed. Cases will be selected using purposive sampling. A total of 199 patients with demographic details gender, age, weight, BMI and duration of type 2 diabetes were noted. History and clinical examination, type 2 diabetes since (years), on treatment since (years), DASS-12 [108] stress sub-scale and laboratory findings such as fasting blood glucose before admission and after admission for COVID-19, post-prandial blood glucose before admission and after admission for COVID-19 and glycosylated hemoglobin before admission and after admission for COVID-19 were also considered. The methodology (COVI-Stress) used is shown in figure. (4.7)

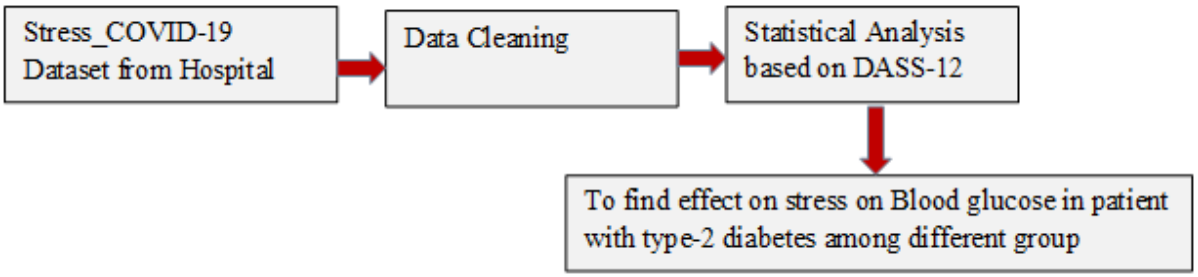


Figure 4.7: COVI- Stress methodology

4.6.2 Data Cleaning

Data compilation was done in the master chart, i.e., in the MS-EXCEL sheet. Representation of categorical data in the form of numbers and percentages while continuous data are presented as mean, standard deviation.

4.6.3 Preferred Statistical Methods

DASS was originally developed for people having an age 17 or older. To assess the severity of depression, anxiety and stress. To evaluate the severity of core symptoms of depression, anxiety and stress over the week. The DASS-12 stress sub-scale was used to determine the level of stress, as shown in Table (4.11) [108].

Table 4.11: Depression, anxiety &stress scale (DASS)

Version scale	Subscale	Questions per subscale	Total no of questions
DASS-42	Depression Anxiety Stress	14 14 14	42
DASS-21	Depression Anxiety Stress	7 7 7	21
DASS-12	Depression Anxiety Stress	4 4 4	12

DASS-12 scale has four questions related to depression, anxiety and stress. In our research, our focus is to concentrate on stress. The questions related to stress are as below.

1] I found it hard to wind down

2] I tended to over-react to situations

3] I found myself getting agitated

4] I was intolerant of anything that kept me from getting on with what I was doing

These stress questions scale is categorized as below:

- Did not apply to me at all – NEVER (0)
- Applied to me to some degree, or some of the time - SOMETIMES (1)
- Applied to me to a considerable degree, or a good part of the time - OFTEN (2)
- Applied to me very much, or most of the time - ALMOST ALWAYS (3)

Based on the stress scores, patients were categorized (0 to 4) as normal, 5-moderate, 6-severe and >= 7 extremely severe groups. Before and after discharge, HbA1c, FBG, and PPBG were measured to estimate the difference in glycaemic parameters. based on the severity of stress; if a significant difference is found, then group wise comparison will be done as below.

- Patients with mild stress vs. Patients with moderate stress
- Patients with mild stress vs. Patients with severe stress
- Patients with moderate stress vs. Patients with severe stress

To estimate the difference in the severity of stress, a comparison of residence wise and age group wise (34 years, 35-49 years, and >50 years)[109] and gender wise was performed. The paired t-test was used to compare glycemic parameters before and after COVID-19-related admission. The difference in glycemic parameters between groups was compared using an unpaired t-test. The chi-square test was used for categorical variables. A statistically significant p-value was defined as less than 0.05.

4.7 Impact of Lifestyle Parameters on Type-2 Diabetes Prediction

4.7.1 Dataset Description

To analyze the effect of lifestyle related parameters, the dataset used for this research contain 374 cases, 87 of which were diabetes patients and 287 of which were not with 17 features. It must be emphasized that no experiment or intervention was performed on human subjects in this study. With their permission, only their personal information was collected for analysis. The dataset contains food habit related features and demographic and lifestyle features from the dataset. Along with class labels, Cereal grains, Salad consumption, Cooked vegetables, Grain, Sweet,

Frequency of sweet, Sugar, Milk, Milk quantity, Smoking, Profession, Physical exercise, age, weight, height, and gender are applied to analyze the effect are specified in Table (4.12).

Table 4.12. Diabe_lifestyle methodology dataset description

Sr.No.	Features	Description
1	Gender	Male or Female
2	Age	Age of a person
3	Height	Specifies height of a person in cm
4	Weight	Specifies weight of a person in kg
5	BMI	Specifies body mass index in Kg/m2
6	Profession	Business involves moderate work/business involves sitting job profile/housewife/office executive/involves heavy exercise/other
7	Smoke	Yes/No
8	Exercise	Less than 30 min/30 min-1 hour/1 -2 hour/2-4 hour/dance/martial arts/no exercise
9	Cereal grains consumption	Once/twice/thrice/more than 3 times in a day
10	Salad consumption	Salad quantity consumption
11	Cooked Vegetables	Once/twice/once in 2-3days/once a week/none
12	Sweet eating	Yes/No
13	Sweet_freq	Once/once a week/once in 15-30 days/rarely
14	Sugar	Yes/No
15	Milk	Specifies consumption of milk and milk products Yes/No
16	quantity	Milk quantity consumption
17	Class Label	Diabetes / No diabetes specifies Yes/No

4.7.2 Data Preprocessing

As the received data is in raw format, so dataset needs some pre-processing activities. Initially, the nature of the dataset is examined. Non-numeric attributes are

inherently abstract in value; hence, after data collection, some pre-processing steps such as data cleaning and label encoding technique were applied. From the used features, numerical variables are age, weight, height and BMI. BMI is calculated by considering the value of weight and height. The remaining features are categorical in nature; therefore, the processing of categorical attributes to numeric attributes is done by using label encoding.

4.7.3 Optimal Feature Subset Selection Using Preferred Techniques

To extract a subset of relevant features from the dataset features selection approach is used. It helps to reduce the dimensionality of the dataset by selecting the appropriate features from the original feature set. In this research, five feature selection techniques, such as RFE, Feature importance, K-best method, Correlation, Information gain were applied to extract important features. The optimal feature subset selection process is divided into four phases.

Phase 1: Using a training data set to predict diabetes, five different feature selection methods were applied to see how well they performed.

Phase 2: Application of all these methods on LS diabetes data set of lifestyle parameter, helps to determine the most efficient and accurate feature selector.

Phase 3: To carry out this phase, first determine the classifier and then apply five methods and analyze which feature selector method performs best for the chosen classifier.

Phase 4: This phase determines which feature selection method provides the best performance by taking into account various performance metrics.

4.7.4 Diabetes Prediction Using Preferred Classifier

A machine learning approach is used to create a predictive model for diabetes early diagnosis. The overall methodology was divided into 3 phases. Figure (4.8) shows the proposed Diabe_lifestyle framework.

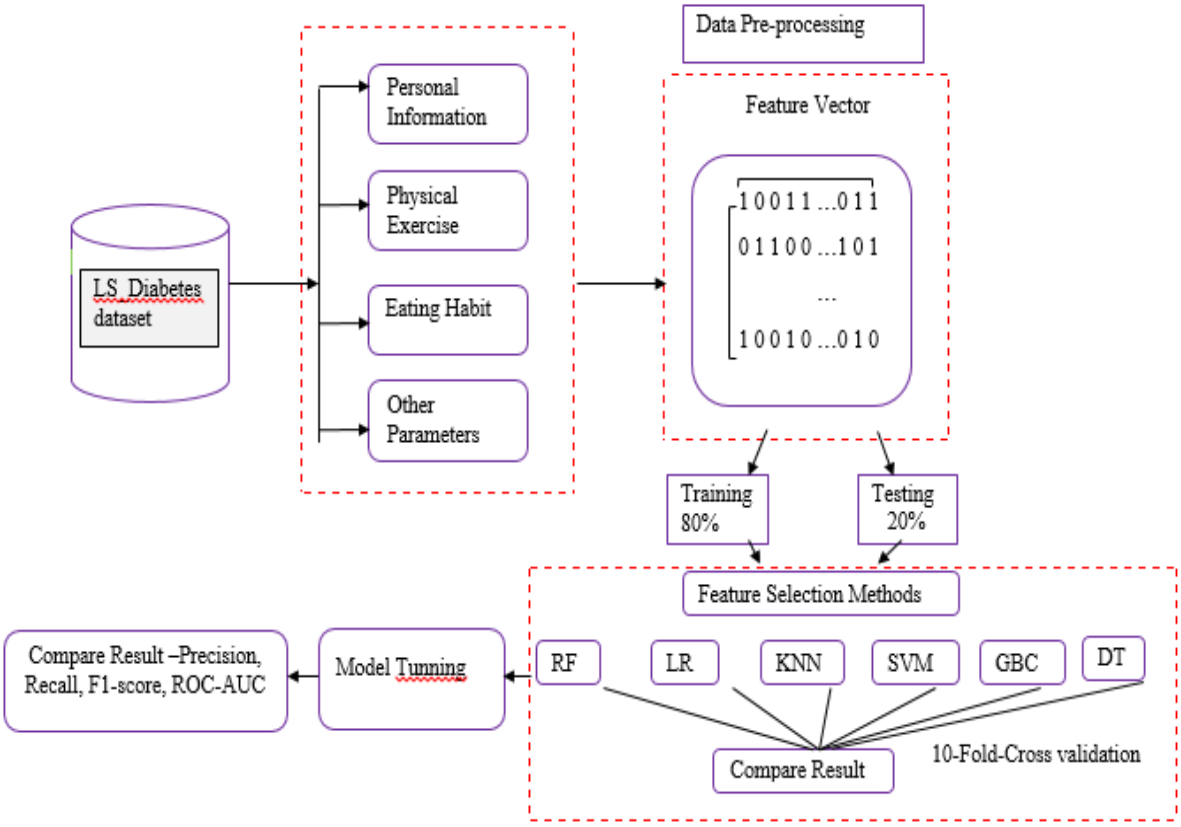


Figure 4.8: Proposed Diabe_lifestyle framework

Data pre-processing was completed in phase I. A predictive model was built using six ML classifiers: LR, SVM, DT, RF, GBC, and KNN. Model training was done on 80% of data and rest of the data used for testing purpose. Training and testing accuracy was calculated. K-fold method was applied for assessment of model performance; here, we considered K=10.

In phase II, for extracting important features we applied five feature selection techniques namely, K best method, feature importance, information gain, correlation and recursive feature elimination (RFE) techniques. We calculated the performance of models using various selected features.The number of features that provided the highest accuracy for each model were recorded. A comparison of the feature selection technique was performed.

In phase III, all selected classifiers' models were optimized using the hyperparameter tuning-based grid search method. The data in the grid search was cross-validated ten times. The best hyperparameters were chosen based on “Area Under the Receiver

Operating Characteristic Curve (ROC- AUC curve)."and accuracy. The algorithmic steps mentioned below in figure (4.9) are based on lifestyle related parameters

4.8 Algorithm for D_stress and Diabe_lifestyle methodology

Proposed D_Stress and Diabe_lifestyle procedural steps are listed in Table (4.13).Dataset of these methodologies are different

Table 4.13. Procedure for D_Stress and Diabe_lifestyle methodology

Algorithm: D_Stress and Diabe_lifestyle methodology
1: **procedure** Label Encoder *(Diabetes dataset)*
2: **return** *(Diabetes dataset)*
3: **procedure** Split_ Data *(Diabetes dataset)*
4: *Train_data,Test_data= split(attributes, label_var)*
5: **return** *Train_data,Test_data*
6: **procedure**L1_Crossvalidation*((Diabetes dataset, clf, K=10)*
7: *clf= (RF | LR | KNN |GBC | SVM |DT)*
8: **return**
9: **procedure** Feature Selector*(Diabetes dataset, Fs)*
10: *Fs= (Correlation |Select k best | RFE | Feature importance | Information gain)*
11: **return** *Fs*
12: **procedure** L2_Crossvalidation *(clf, Fs)*
13: **return**
14: **procedure** grid search estimator*(clf, Fs, parameter grid)*
15: **return** *best estimator*
16: **procedure** L3_Crossvalidation*(clf, Fs , best estimator)*
17: **return**

4.9 Mathematical Model for D_Stress and Diabe_lifestyle

Input: Training Dataset: $|Dtrain|_{X,Y} \leftrightarrow |X,Y|_{i,m}$

Output: Best Classifier with learner *S*

LS_Diabetes dataset is expressed as,

D= {f1, f2, f3,….fp}

The dataset is expressed as,

$$\boldsymbol{F}(\boldsymbol{X},\boldsymbol{Y}) = \{(\mathbf{Xi}\,,\mathbf{Yi})|\ \mathrm{X_i} \in \mathrm{R}^{n}\,,\mathrm{Y_i} \in 0,\mathbf{1}\}\} \tag{4.9}$$

Where,

X={x1, x2, x3,….xn} and Y={ y0, y1}

Feature Set,

freduced = { fs1| fs2| fs3| fs4| fs5}

Where, fs1 =Select K-Best, fs2 =Feature Importance

fs3 =Correlation , fs4 =Information Gain

fs5=Recursive Feature Elimination

Classifier:

Cf = Classifier {freduced |K=10 }

Optimized to:

$$x^{*} = \underset{x\epsilon X}{argmin}\, g(x) \tag{4.10}$$

Where,

g(x)=Objective score on the validation set

x*= Set of hyperparameters

4.10 Summary

This chapter reveals the proposed research methodology of the complete research work. The functioning process of the proposed methodology is clearly explained. In this section, we provide the details of all data sources along with the description of all datasets. We also provide a brief overview of the train and test data set used in this study. In addition, the phases involved in the research methodology are also discussed. It also describes the adopted phase wise implementation methodology. The proposed methods are used to predict the impact of various parameters on the risk of development of diabetes to achieve a higher classification accuracy on the self-collected LS_diabetes dataset and COVI_Stress dataset. In the next chapter auto encoder based ensemble learning methodology adopted for the diabetes prediction model has been discussed in detail.

CHAPTER-5

DESIGN OF NOVEL MODEL FOR TYPE 2 DIABETES PREDICTION – IMPROVED ENSEMBLE LEARNING WITH DIMENSIONALITY REDUCTION (IELDR)

CHAPTER 5 DESIGN OF NOVEL MODEL FOR TYPE 2 DIABETES PREDICTION – IMPROVED ENSEMBLE LEARNING WITH DIMENSIONALITY REDUCTION (IELDR)

5.1. Introduction

Today's world is extensively affected by the endocrine disease Diabetes Mellitus which is commonly known as diabetes. The requirement for an effective model can able prognoses diabetes with more accurateness as early. Machine learning is a need for automation with minimum human effort. The single algorithm to predict type 2 diabetes is a very sophisticated approach for early diabetes prediction. A hybrid approach with a stacking classifier would give better results. To increase the accurateness of innovative prediction autoencoder based stacking machine learning model (IELDR) is proposed. In this research ensemble approach of machine learning classifiers and autoencoder based feature extraction is implemented for the early diagnosis of type 2 diabetes. The Improved Ensemble Learning with Dimensionality Reduction (IELDR) is employed on the LS_diabetes dataset to obtain efficient classification accuracy and validated on the PIMA diabetes dataset and Diabetes_2019 dataset.

5.2. Type-2Diabetes Prediction System Using Proposed IELDR

5.2.1. Dataset Description

The proposed novel framework uses a total of 374 records of data collected using a questionnaire-based survey. The questionnaire used in the dataset used for research was administered web based as well as personally through face-to-face interviews. The dataset contains a total of 35 features. These features are a combination of demographic, stress related, lifestyle related, genetic and other diagnostic results. Table (5.1) shows a detailed attribute description.

Table 5.1. LS_Diabetes dataset attribute description

Sr.No	Attribute	Description
1	Gender	Male or Female
2	Age	Age of a person
3	Height	Specifies height of a person in cm
4	Weight	Specifies weight of a person in kg
5	BMI	Specifies body mass index in Kg/m2
6	Anxiety	Yes/No
7-16	Stress	Workload / Poor Salary/ Deadline/ Travel/ Repetitive Work/ Unplanned Work/ Career/ Powerlessness/ Job Security/No Satisfaction Specified as, Very low/ Low/Average/ High / Very high
17	Profession	Business involves moderate work/business involves sitting job profile/housewife/office executive/involves heavy exercise/other
18	Smoke	Yes/No
19	Exercise	Less than 30 min/30 min-1 hour/1 -2 hour/2-4 hour/dance/martial arts/no exercise
20	Cereal grains consumption	Once/twice/thrice/more than 3 times in a day
21	Salad consumption	Salad quantity consumption
22	Cooked Vegetables	Once/twice/once in 2-3days/once a week/none
23	Sweet eating	Yes/No
24	Sweet_freq	Once/once a week/once in 15-30 days/rarely
25	Sugar	Yes/No

26	Milk	Specifies consumption of milk and milk products Yes/No
27	quantity	Milk quantity consumption
28	High blood pressure	Yes/No
29	Systolic	Numeric
30	Diastolic	Numeric
31	Fasting Sugar	Specifies Range of FBG
32	Postmeal sugar	Specifies Range of PPBG
33	HbA1c	Specifies Range
34	Family history	Yes/No
35	Class	Diabetes/Non-diabetes specifies Yes/No

5.2.2. Data Pre-Processing

The dataset consists of both categorical and numerical values from the questionnaire answers. The collected dataset is above the age group of 18 years. In our collected dataset, one record was below 18 years, so removed from the dataset. As most of the features are categorical and machine learning works only on numerical data, label encoding is used to convert categorical features into numerical ones.

5.2.3. Autoencoder based optimal feature extraction

The proposed model used an autoencoder based feature extraction method. Autoencoder is used for a compressed representation of input features. It is used in classification problems for predictive model development. The autoencoder is made up of an encoder and a decoder. The encoder learns how to interpret the input and compress it to a bottleneck layer-defined internal representation. The decoder takes the encoder's output and attempts to recreate the input. Once the autoencoder has been trained, the decoder is discarded and the encoder is used to compress data and train a model.

5.2.4. Diabetes Prediction Using IELDR

To achieve this objective, a novel autoencoder based two-level stacking super learner framework is proposed. Stacking is a two-level classification method for the final prediction creates a new training dataset. In the IELDR framework, we are training the base learner in level 0 on the complete dataset and different blocks of dataset using k-fold. Here we have considered K=10. After training probabilities of each class are predicted. The proposed IELDR model trains the base model on different blocks of data, stores a prediction, and then combines the predictions from the base learner into a single vector and passes this vector as input to the meta learner. Meta learner is trained from predicted probability to do the final prediction using k-fold cross validation. The existing stacking model trained the models on complete dataset and combining the predictions using Meta learner. Here we have considered various base learner models at level 0, namely, RF, SVM, LR, DT, GBM, DT, KNN and NB. The multilayer perceptron is used as a Meta classifier at level 1, which shows the final prediction. Based on the prediction, provide three different levels of recommendation. Figure (5.1) shows the proposed Improved Ensemble Learning Dimensionality Reduction Model (IELDR). The algorithm is mentioned below.

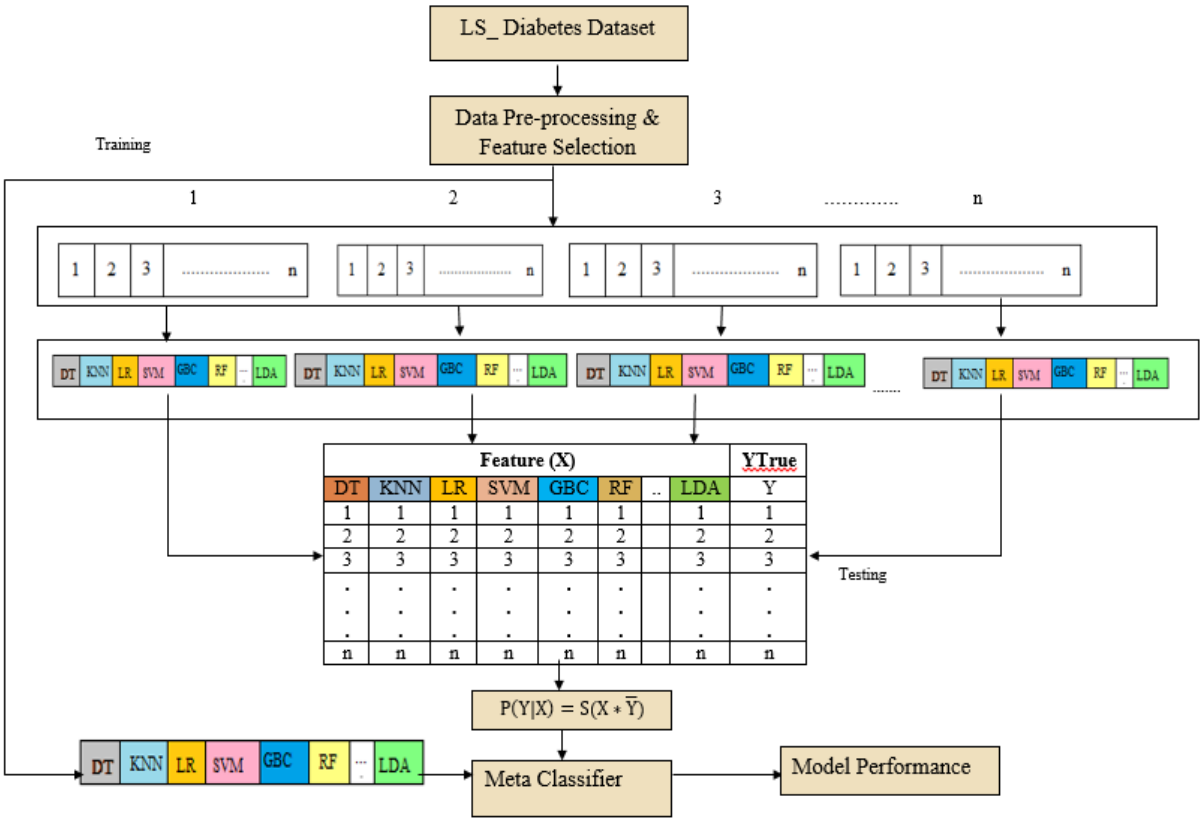


Figure 5.1: Proposed Improved Ensemble Learning with Dimensionality Reduction Model

5.3. Mathematical model for IELDR

Proposed two levels stacking super learner model with autoencoder for feature extraction

Input: Training Dataset: D_{train} = {features(f_i), Class(y_i)}

Output: Stacking ensemble classifier with super learner *S*

IELDR model is divided into two phases:

1) Feature Extraction
2) Learning Ensemble model.

1. Feature Extraction using autoencoder

- Build the autoencoder using the encoder network and the decoder model.

$$\alpha: \mathcal{D} \rightarrow \mathcal{F} \tag{5.1}$$

$$\beta: \mathcal{F} \rightarrow \mathcal{D} \tag{5.2}$$

$$\alpha, \beta = \underset{\alpha,\beta}{argmin} \parallel \mathcal{D} - (\alpha \circ \beta)\mathcal{D} \parallel^2 \tag{5.3}$$

- The encoder model function α in eq. (5.1) maps the actual data D to the latent feature space F. The decoder model function β in eq. (5.2) maps the latent feature space F to the output D. In this case, the output obtained from the decoder is similar to the input function with a reduced feature set as denoted in Eq. (5.3).
- The encoder model represented by eq. (5.4) is defined with the sigmoid activation function.

$$z = \sigma(w\mathfrak{f} + b) \tag{5.4}$$

Where **z** is the latent feature dimension.

- Similarly, the decoder model represented by eq. (5.5) is defined with different weights, biases, and sigmoid activation function:

$$\mathfrak{f}' = \sigma'(w'z + b') \tag{5.5}$$

- The error function $\mathbb{E}$ is represented by eq. (5.6) is defined with encoder and decoder network functions. This error function trains the autoencoder using the standard backpropagation method.

$$\mathbb{E}(\mathfrak{f}, \mathfrak{f}') = \parallel \mathfrak{f} - \mathfrak{f}' \parallel^2 = \parallel \mathfrak{f} - \sigma'(w'(\sigma(w\mathfrak{f} + b)) + b') \parallel^2 \tag{5.6}$$

- The purpose of using an autoencoder for feature selection is to use the information to encode the data efficiently to regenerate the same data with important features on the decoder side.

2. Two-Levels Stacking Super Learner Model

- **Step1:** Select a 10-fold split of the training dataset D_{train}
- **Step 2:** Identify the appropriate list of base Models *M*

- **Step3:** Train the first-level *M* base learners on the training dataset

  for m ← 1 to M do

  i. Perform 5-fold cross-validation on each base learner b_m on D_{train}

  ii. Collect the cross-validated predicted values from each of the M base

  algorithms.

  end for

- **Step 4:** Build the new dataset with the *R* cross-validated predicted score obtained from each M base algorithm and construct a new dataset with the shape as R x M.

  for m← 1 to M do

  Build the new dataset as D_{new} = ({features(f_new_i), Class(y_i)}

  where f_new_i = { $b_1(f_i)$, $b_2(f_i)$, $b_3(f_i)$, ……., $b_m(f_i)$)

  end for

- **Step 5:** Build the second-level meta-model based on out-of-fold predicted values and evaluate the meta-model based on a holdout dataset to make predictions using the following equation:

  $$y=g(f) = B(f)= b' (b_1(f_i), b_2(f_i), b_3(f_i), \ldots\ldots, b_m(f_i)) \quad (5.7)$$

- **Step 6: Prediction on new data.**

First, generate predictions from the base models and provide the obtained predictions to the super learner *S* to generate the ensemble prediction on testing or new data samples.

5.4. Summary

The research aims to propose a novel IELDR framework with the major assumption that a Single classifier showing prediction is replaced by two levels stacking autoencoder based model producing great accuracy. IELDR framework works on the LS_Diabetes dataset used to evaluate the performance. The proposed framework

was validated with two different datasets, PIMA Indian diabetes dataset and the Diabetes_2019 dataset, for its stability. The next chapter reveals the experimental result of all proposed methodologies.

CHAPTER-6

RESULT AND PERFORMANCE ANALYSIS

CHAPTER 6 **RESULT AND PERFORMANCE ANALYSIS**

6.1 Introduction

Diabetes Mellitus needs the support of machine learning for early prediction. The research study aims to optimize machine learning methods for the early prediction of diabetes. This chapter reveals the experimental evaluation of the impact of various parameters on the development of the type 2 diabetes risk predictive model. To find the impact of various parameters, the proposed research compares the performance of various preferred conventional machine learning to provide a new state-of-art prediction classification system for diabetes. The proposed research builds a prediction model based on an autoencoder based two-level stacking model that has been tested on the PIMA and Diabetes_2019 datasets. Results are analyzed in terms of various measures and expressed through ROC Curves. The observed results are tabulated and prove that the proposed IELDR models give more noticeable results than traditional algorithms on diverse datasets, which we have considered for validation. Section 6.2 discusses the experimental setup required for model development. Section 6.3 demonstrates the performance analysis based on stress. Statistical analysis of COVID-19 related stress is discussed in section 6.4. Section 6.5 describes performance analysis of lifestyle parameters. In section 6.6 comparison between Diabe_lifestyle methodology and one of the existing methods. IELDR model is discussed in section 6.7.Validation on two datasets is discussed in section 6.8 and section 6.9.Overall summary of research is explained in section 6.10.

6.2 Experimental Setup

The experiments are implemented using Google Colab and Python Jupyter IDE and the platform has an x64-based processor, AMD A9-9420 RADEON R5, 5 COMPUTE CORES,4GB RAM and Windows 10 Pro OS. This methodology used quantitative and qualitative experiments to evaluate the proposed model's performance on the LS_Diabetes dataset.

6.3 Performance Analysis of Type-2 Diabetes Prediction for Stress Parameters

The dataset contains a total of 374 records. The data is split into training and testing portions in an 80:20 ratio. The target is designated as "Outcome" and is represented as: 0 for "No Diabetes" and 1 for "Diabetes." With this as the class label, an analysis of the current states was conducted that can be used to make predictions about the future based on the input parameters. A total of 186 were male and 188 were female. 88 had diabetes and 286 were non-diabetes. The proposed D_Stress methodology uses six different machine learning classifiers.

Table 6.1: Training and Testing Accuracy of Six Classifiers

Classifier	Training Accuracy (%)	Testing Accuracy (%)
LR	81	84
RF	100	82.67
GBC	98	85.33
DT	82	80
SVM	83	81.33
KNN	82	81.33

Employing stress parameters on the LS_Diabetes dataset using six different classifiers shows that RF gives the highest training accuracy, whereas all other classifiers show an accuracy above 80%. The Testing accuracy of LR, RF, GBC, DT, SVM and KNN was 84%, 82.67%, 85.33%, 80%, 81.33% and 81.33% respectively shown in Table (6.1).

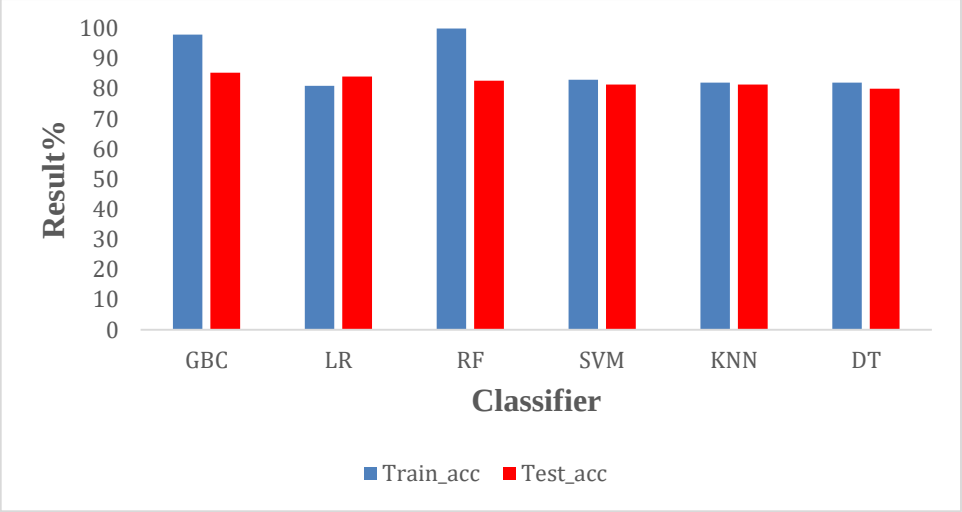


Figure 6.1: Comparison of train_test classification accuracy

In figure (6.1) the graphical representation of training and testing accuracy was shown. GBC shows good results compared to others by achieving an accuracy of 85.33%.

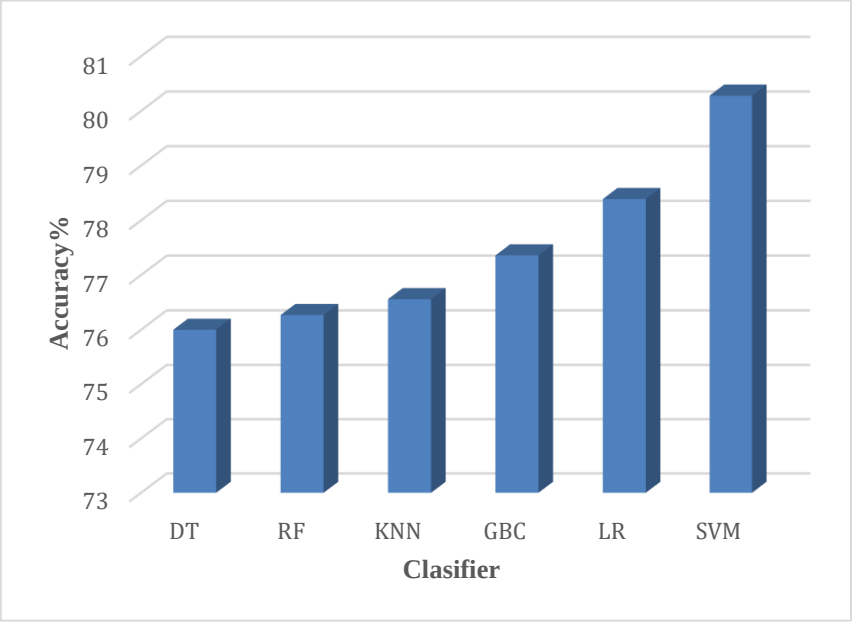


Figure 6.2: Comparative analysis of k-fold accuracy

To improve model performance, we have used the K-fold cross-validation method. Here we have used K=10 and calculated the accuracy of all six classifiers shown in figure (6.2).Result shows that DT, RF, KNN, GBC, LR and SVM shows an accuracy of 76%,76.27%,76.56%,77.36%,78.39% and 80.28% respectively. The highest K-fold accuracy is shown by the SVM classifier of 80.28%.

6.3.1 Analysis of Feature Selection Methods

Feature selection is an essential step for model improvement. Analysis of Different feature selection techniques based on the number of selected features and corresponding classifier accuracy is given in Table (6.2).

Table 6.2: Analysis of feature selection methods based on the number of features.

Method Used	No of Selected features	Accuracy (%)
DT	16	76
K-Best +DT	**6**	**77.08**
Correlation + DT	10	75.22
Feature importance + DT	8	75.75
Information gain + DT	7	75.23
RFE + DT	8	77.06
RF	16	76.27
K-Best +RF	**5**	**80.28**
Correlation +RF	6	78.68
Feature importance +RF	3	77.9
Information gain +RF	7	77.35
RFE +RF	15	78.95
KNN	16	76.56
K-Best +KNN	11	80
Correlation +KNN	**10**	**80.26**
Feature importance +KNN	8	78.14
Information gain +KNN	7	80
GBC	16	77.36

K-Best +GBC	5	78.68
Correlation +GBC	6	80.01
Feature importance +GBC	**6**	**81.08**
Information gain +GBC	5	79.77
RFE +GBC	6	80.55
LR	16	78.39
K-Best +LR	12	80.79
Correlation +LR	7	81.07
Feature importance +LR	6	80.8
Information gain +LR	7	81.07
RFE +LR	**9**	**81.34**
SVM	16	80.28
K-Best +SVM	**11**	**82.14**
Correlation +SVM	8	81.6
Feature importance +SVM	10	81.35
Information gain +SVM	8	81.33
RFE +SVM	9	81.35

A comparative analysis of five different feature selection methods on six different classifiers is discussed in Table (6.3).

Table 6.3: Comparative analysis of different feature selection methods

Feature Selector	LR	RF	GBC	DT	SVM	KNN
Select K-Best	80.79	**80.28**	78.68	**77.08**	**82.14**	80
Feature importance	80.8	77.9	**81.08**	75.75	81.35	78.14
Correlation	81.07	78.68	80.01	75.22	81.6	**80.26**
Information Gain	81.07	77.35	79.77	75.23	81.33	80
RFE	**81.34**	78.95	80.55	77.06	81.35	-

By applying five feature selection techniques with 10-fold cross-validation LR model with RFE gives an accuracy of 81.34%, the RF model with select k-best gives 80.28% accuracy, the GBC model with feature importance method shows an accuracy of 81.08%, the DT model with select k-best method shows an accuracy of 77.08%, SVM classifier shows the highest accuracy for the select k-best method by achieving an accuracy of 82.14%, KNN classifier provided an accuracy of, 80.26% with correlation feature selection method.

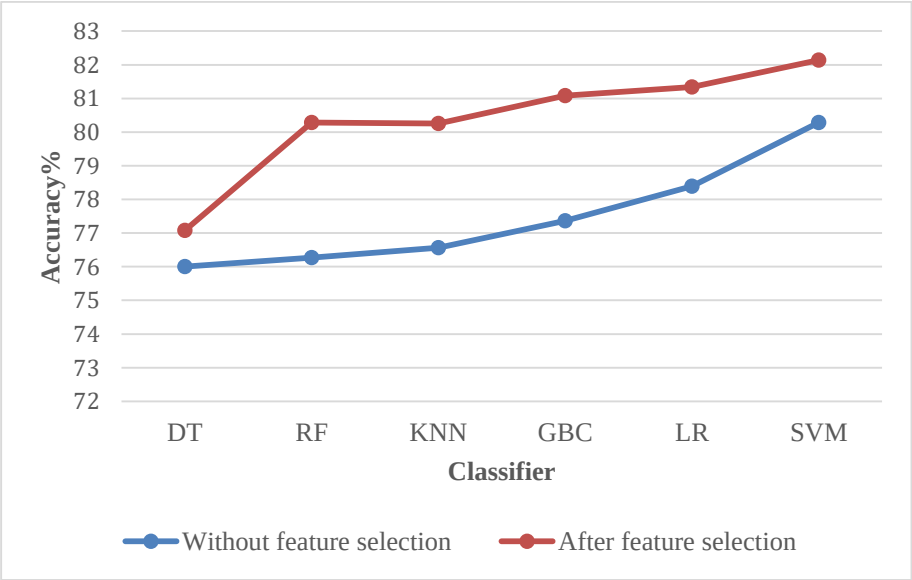


Figure 6.3: Analysis with feature selection and without feature selection method

Figure (6.3) illustrates feature selection and without feature selection and the result shows that with feature selection, classifier accuracy has been significantly improved. Among all classifiers, SVM shows the highest accuracy of 82.14%.

6.3.2 Experimentation Result

To improve the model performance, we have opted hyperparameter tuning optimization technique. Hyperparameter tuning using the grid search method applied to six different classifiers is shown in Table (6.4).

Table 6.4: Hyperparameter tuning classification accuracy

Classifier	Accuracy (%)	Error Rate (%)
KNN	80.26	19.74
RF	81.33	18.67
DT	82.67	17.33
LR	85.33	14.67
GBC	**86.67**	13.33
SVM	**86.67**	13.33

Comparative analysis by applying feature selection technique and hyperparameter tuning accuracy is shown graphically below.The proposed method shows that overall improvement in model performance in terms of accuracy.

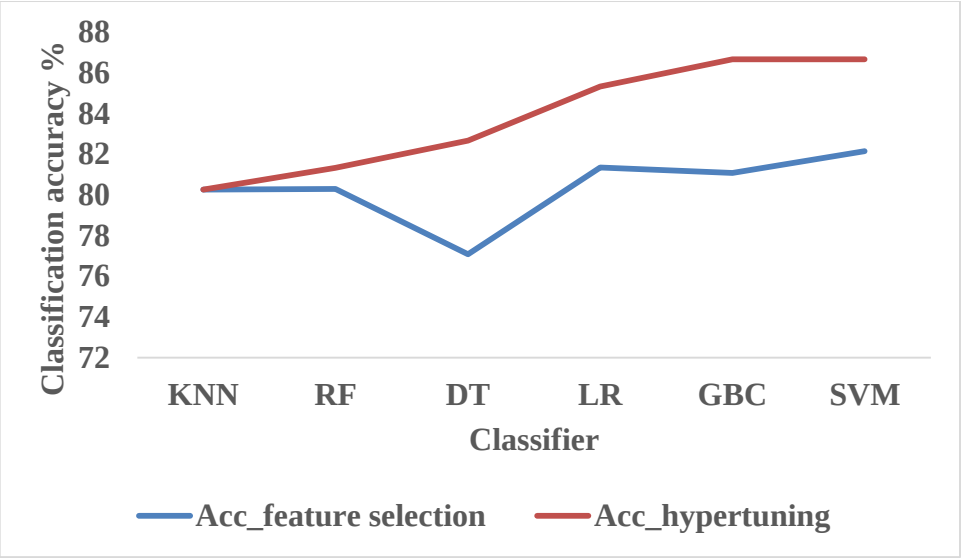


Figure 6.4: Accuracy based comparative analysis of feature selection and hyperparameter tuning

Figure (6.4) shows comparative analysis. Experimental analysis shows that applying hyperparameter tuning using the grid search method significantly improves model accuracy. All classifiers achieved accuracy above 80%. Among six different classifiers, the SVM and GBC classifier gives an efficient result in classification accuracy of 86.67% and an error rate of 13.33%.

The accuracy comparison of K-fold, feature selection and hyperparameter tuning is shown in figure (6.5) below.

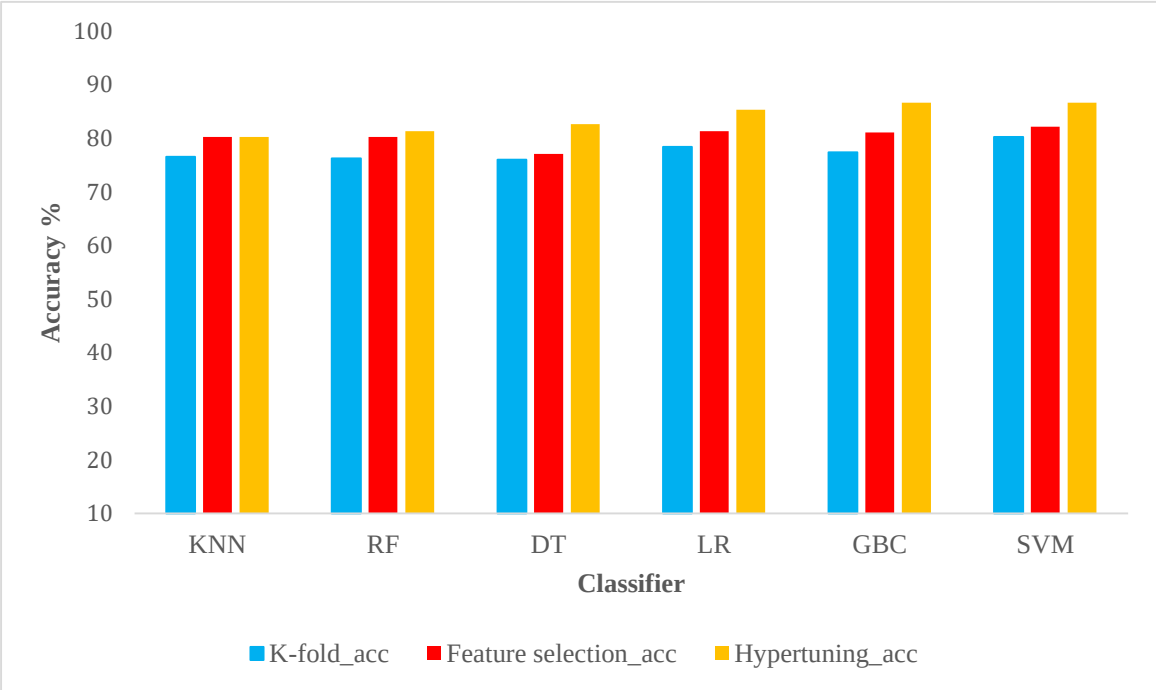


Figure 6.5: Accuracy comparison of classification algorithm

According to experimental observations hyperparameter tuning method improved accuracy by 1% , 5.59% , 3.99% , 5.59% , 4.53% for RF,DT,LR,GBC,SVM classifier respectively as compared to feature selection accuracy. A similar observation was seen between k-fold and hyperparameter tuning accuracy. The rise was by 3.7%, 5.06%, 6.67%, 6.94%, 9.31% and 6.39% for KNN, RF, DT, LR, GBC, SVM classifiers respectively.

Table (6.5) shows the selected best hyperparameter based classifier based performance.

Table 6.5: Selected hyperparameter based classifier performance

ML Method	Best hyper parameter	Performance
RF	criterion= entropy max_depth=4 max_features= sqrt n_estimators= 200	**91.17%**
GBC	learning_rate=0.01, max_depth=2 max_features=sqrt n_estimators=2000 subsample=1	86%
LR	C= 3,multi_class= ovr,penalty=l2 solver= liblinear	90.87%
K-NN	algorithm=auto,metric=minkowski leaf_size=30,n_neighbors=7,p=2 weights=uniform	81.24%
DT	criterion= 'gini', max_depth= 1, min_samples_split= 2	76.31%
SVM	C=0.1, kernel=linear,degree=3,gamma=1	**91.18%**

Above table shows, best performance for the diabetes prediction based on stress features from the proposed pipeline using SVM model is achieved when the c value is equal to 1, the linear kernel has a degree value of 3 and the gamma is equal to 1.

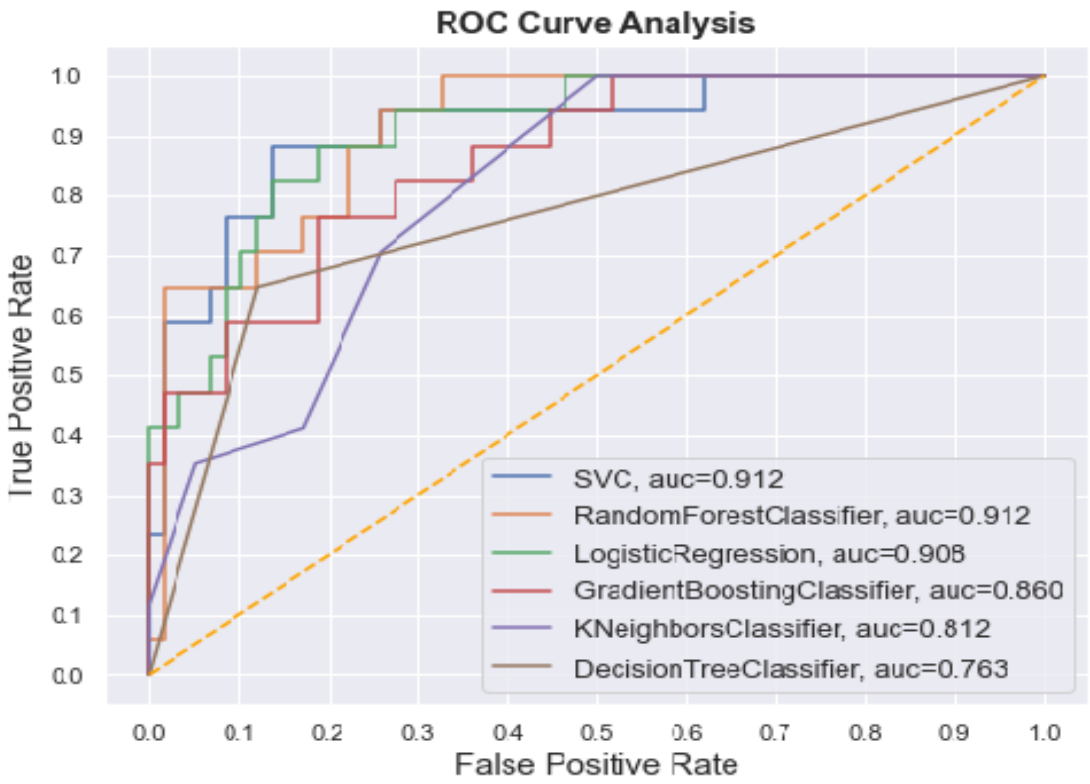


Figure 6.6: ROC curve analysis

Comparison of ROC curve analysis of various classification algorithms shown, figure (6.6).Experimental analysis showed that the SVM classifier performs best with an AUC value of 0.912. Analysis showed, the best performance was observed in SVM and RF with ROC value of 0.918 and 0.917.

Performance evaluation of these classification algorithms can be done by measures such as accuracy; positive predicted value can also be called precision, recall and F-measure. Comparative analysis of different classifiers on various measures is shown in Table (6.6).

Table 6.6: Comparative analysis of different classifiers on various measures

Classifier	Result (%)				
	Accuracy	Precision	Recall	F1-score	Roc_Auc Score
LR	85.33	85	85	84	90.87
RF	81.33	81	81	77	**91.17**
GBC	**86.67**	**87**	**87**	**85**	86
DT	82.67	83	83	83	76.31
SVM	**86.67**	**87**	**87**	**85**	**91.18**
KNN	80.26	80	81	79	81.24

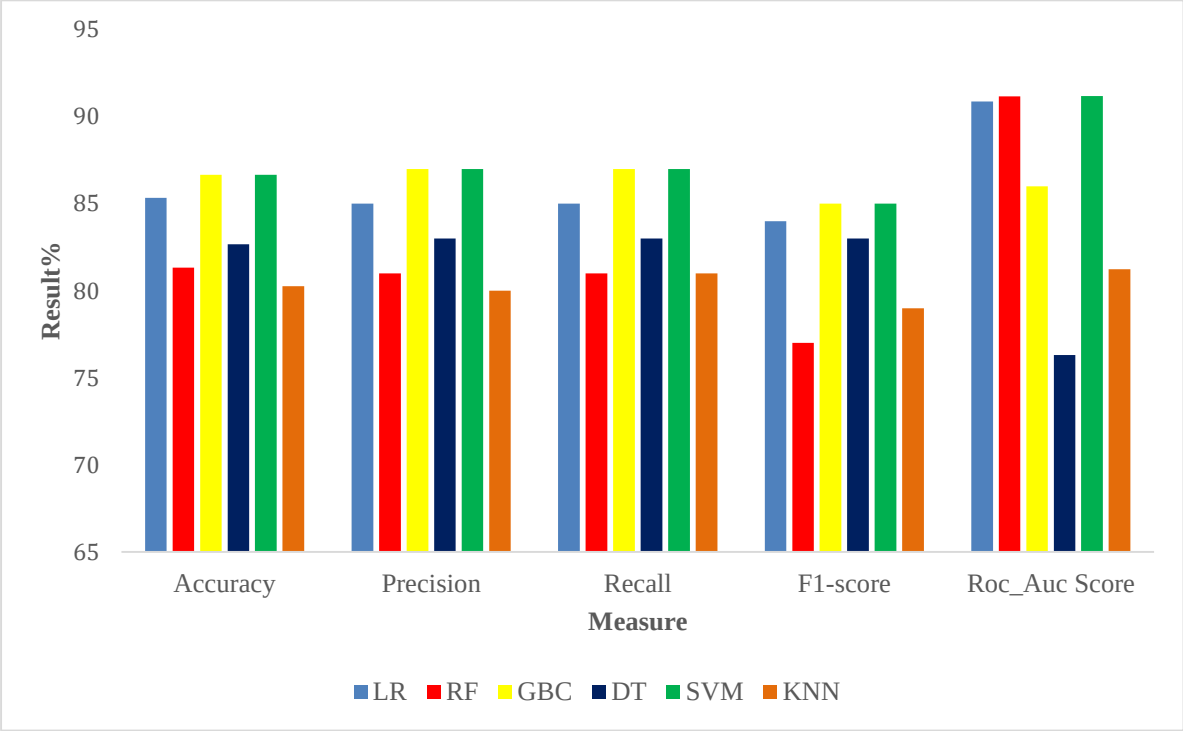


Figure 6.7: Performance evaluation with D_Stress

The proposed evaluation of various classifiers is done using various performance metrics. Figure (6.7) depicts a graphical representation of all performance measures mentioned above. SVM and GBC showed highest precision and recall values of 87 %. The highest f-1 score of 85%, was seen with GBC and SVM. Overall, the precision, recall, and f-1 scores are highest for GBC and SVM. In contrast, the lowest

precision found in KNN of 80% and the lowest recall and f1-score value showed by the RF classifier by 81% and 77%, respectively.

Table (6.7) shows experimental validation with the various study based on stress related parameters.

Table 6.7: Experimental validation

Ref. No.	Dataset	No. of Instances Considered for Study	Risk factor	Best Model	Accuracy (%)
[73]	OSMI Mental Health Survey	750	Stress & Demographic	Boosting	75.13%
[74]	Self Collected	206		SVM	85.71%
[75]	Self Collected-Fitbit app	870		RF	82%
	LS_Diabetes	374		**SVM,GBC**	**86.67%**

The result shows that the increasing stress raises the risk of developing type 2 diabetes. Various stress factors that must be considered in our research showed that stress is one of the strong risk element contributing to the risk of developing type 2 diabetes.

6.4 Statistical Analysis for Impact of COVID-19 Related Stress on Blood Glucose Level for Type-2 Diabetes Patients

6.4.1 Data Analysis

To find the impact of COVID-19 related stress, a hospitalized dataset of 199 diabetes patients was collected. A total of 199 patients with M$\pm$SD age 54$\pm$12.8 years were included, of whom 122 (61.30%) were females and 77 (38.7%) were males. Patients' M+SD weight and BMI were 78.1$\pm$14.6 kg and 30.9$\pm$7.5 kg/m^2. A total of 116

(58.3%) patients were from the urban area, whereas 83 (41.7%) were from rural areas Table (6.8**)**. A total of 114 (57.3%) patients were housewives.

Table 6.8: Demographics characteristics of participants

Parameter	**Result**
M±SD age Age range (minimum, maximum)	54±12.8 years 30-92 years
Gender n (%) Male Female	 77 (38.7%) 122(61.3%)
M±SD weight Range of weight	78.1±14.6 kg 30-114 kg
M±SD BMI Range of BMI	30.9±7.5 Kg/m^2 10.6-51.5 Kg/m^2
Residence n (%) Rural Urban	 83 (41.7%) 116 (58.3%)
Profession n (%) Business Service Housewife Other	 7 (3.5%) 10 (5.03%) 114 (57.3%) 68 (34.2%)
M±SD duration of diabetes Range of duration of diabetes (minimum, maximum)	5±3.6 years 0.08-25 years

Out of 199 participants, 54 (27.1%) patients had moderate stress. Severe or extremely severe stress was observed in 19 (9.6%) patients.Thus, out of 199 participants, 73 (36.7%) had perceived stress and 126(63.3%) patients had no stress, as shown in Figure (6.8). In our study, out of 73 patients with stress, 46.6% of patients were female; hence the prevalence of severe stress was more common in females.

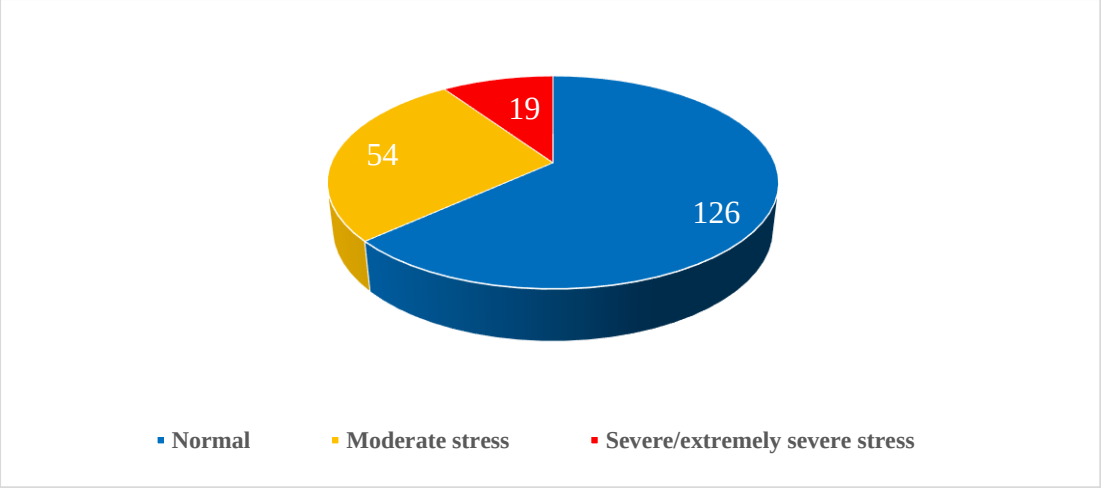


Figure 6.8: Distribution based on the severity of stress as assessed by DASS-12 score

In a group of moderate stress out of 54 patients, 23 (42.6%) were females and 31 (57.4%) were male. In a group of patients with severe/extremely severe stress out of 19, a total of 11 (57.9%) were females and 08 (42.1%) were male. Table (6.9) shows the significant difference found between FBG and PPG values before and at the time of discharge.

Table 6.9: Blood glucose parameters comparison

	Before admission	**After discharge**	**P value**
Mean (SD) fasting blood glucose mg/dl	168.4 (30.6)	195.9 (28.8)	**<0.001**
Mean (SD) postprandial blood glucose mg/dl	312 (62.3)	351.6 (61.9)	**<0.001**

We observed a significant increase in mean FBG and PPBG in T2DM patients at the time of discharge compared to before admission for COVID-19 treatment.

6.4.2 Intra and Inter Group Analysis

A comparison of glycemic parameters based on the severity of stress is presented in Table (6.10).

Table 6.10: Intragroup comparison of glycemic parameters based on the severity of stress

	No stress (n=126)	**Moderate stress (n=54)**	**Severe/Extremely severe stress (n=19)**
M±SD age in years	52.6±13	57.3±12.5	54.2±10.6
M±SD weight in Kg	80±13.9	74±15.1	77.2±16.5
M±SD BMI in kg/m^2	32.6±7.3	27.8±6.7	28.3±7.5
Gender n (%) Male Female	 38 (30.16%) 88 (69.84%)	 23 (42.59%) 31 (57.41%)	 8 (42.11%) 11 (57.89%)
M±SD **fasting blood glucose (mg/dl)** Before admission After admission	 164.7±30 195.1±27.8	 171.3±24.5 198.3±30.8	 185±43.2 194.2±30.8
P value	<0.001	<0.001	0.43
M±SD **post-prandial blood glucose (mg/dl)** Before admission After admission	 309.2±62.3 347.3±64.7	 315.8±59 359±52.6	 319.7±72.8 358.7±68.1
P value	<0.001	**<0.001**	0.06

There was a significant difference in mean FBG before and after admission in patients without stress and those with moderate stress. However, there was no significant difference in mean FBG in patients with severe/extremely severe stress (p=0.43). Similar observations were seen for PPBG (without stress p<0.001; moderate stress p<0.001; severe/extremely severe stress p=0.06). Intragroup analysis for FBG and PPBG is shown in Figure (6.9)

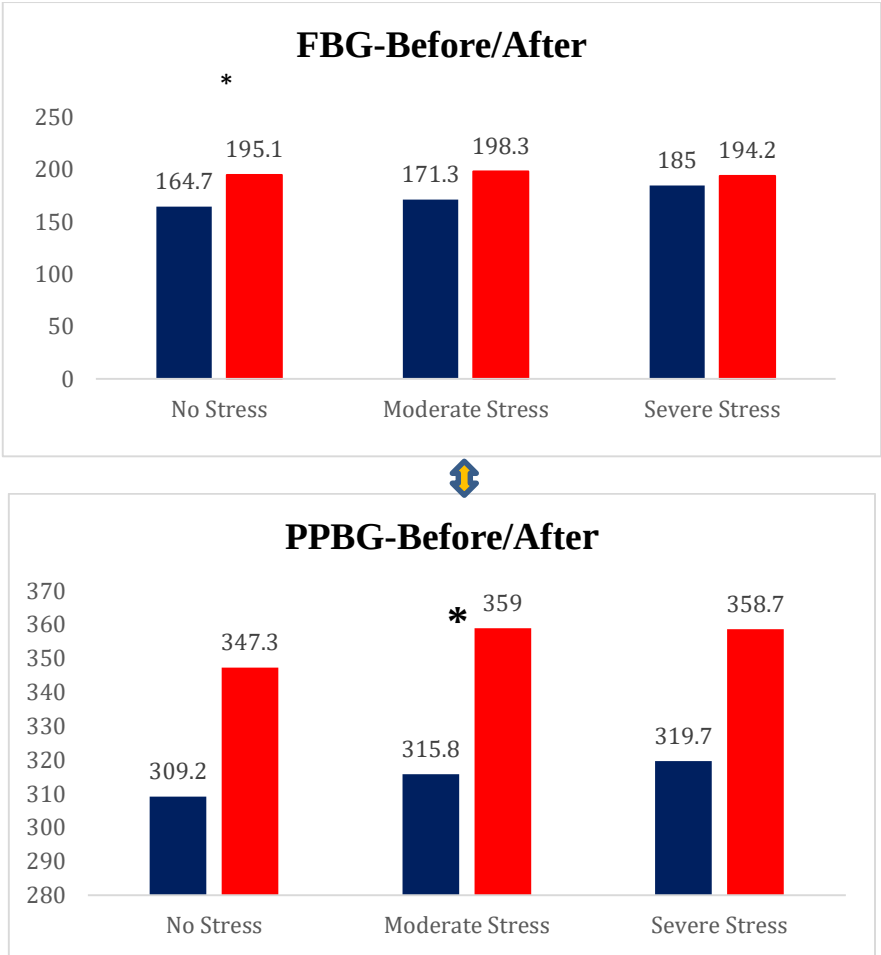


Figure 6.9: Intra-group analysis

Intergroup analysis:

Intergroup analysis showed no difference in the FBG or PPBG in patients without stress versus moderate stress, moderate versus severe/extremely severe stress, and no stress versus severe/extremely severe stress before admission or after discharge shown in Table (6.11).

Table 6.11: P values of intergroup comparison

	Without stress versus moderate stress	**Moderate versus severe/extremely severe stress**	**Without stress versus severe/extremely severe stress**
Fasting blood glucose before admission	0.15	0.09	0.01
Fasting blood glucose at the time of discharge	0.497	0.68	0.98
Postprandial blood glucose before admission	0.51	0.82	0.50
Postprandial blood glucose at the time of discharge	0.24	0.99	0.48

Overall, the study population showed a rise in fasting and postprandial glucose levels in type-2 diabetes patients discharged after COVID-19 treatment. The surge was significant in type-2 diabetes without stress and those with moderate stress.

6.4.3 Group wise Analysis

In patients with stress, compared age wise, gender wise and residence wise and found a significant difference in the FBG and PPBG based on their age, gender and residence, i.e., urban versus rural population, shown in Table (6.12).

Table 6.12: Group wise comparison

Groups	**P-Value**
Age-Stress	**< 0.05**
Gender-Stress	**< 0.05**
Urban-Stress	**< 0.001**
Rural-Stress	**< 0.001**

Overall age wise and gender wise showed the results were found to be statistically significant ($p < 0.05$) and if we consider residence wise, then found that both rural and urban were found to be statistically significant ($p<0.001$).

6.4.4 Age wise Analysis

To perform age wise analysis, we considered three different categories of age group, namely old age (>=50), middle age (between 35- 49) and young age(<=34)[109].

Table 6.13: Age group wise p values comparison

Age Group	p-value
Old age(>=50) --- Stress	**P<0.05**
Middle age (35-49)--- Stress	p>0.05
Young age (<=34)--- Stress	p>0.05

The analytical result with respect to age group wise comparison showed a significant difference in the severity of stress ($p<0.05$) among the old age group shown in Table (6.13).

6.5 Performance Analysis of Type-2 Diabetes Prediction for Lifestyle Related Parameters for LS_diabetes dataset

LS_diabetes lifestyle dataset contains a record of 374 with 18 features, including class labels. For training to the model, 80% of data is used means 299 records are used for training purposes, reaming 20% of data is used for testing purposes, i.e., 75 records used for testing. Based on the class label, an analysis was made regarding the present states, which can be used to make future predictions based on input parameters. The proposed Diabe_Lifestyle methodology used six machine learning classifiers, namely, LR, SVM, GBC, RF, DT and KNN classifiers.

Table 6.14: Comparison of training and testing accuracy

Classifier	Training Accuracy (%)	Testing Accuracy (%)
DT	100	81.33
SVM	90.64	81.33
KNN	83.94	82.66
RF	100	82.66
LR	81.93	85.33
GBC	98.66	85.33

Table (6.14) shown employing lifestyle parameters on the LS_Diabetes dataset using six different classifiers showed that the highest training accuracy was observed in the tree-based model, i.e., RF and DT, with an accuracy of 100%. During testing classification accuracy of LR, GBC,RF, DT, SVM and KNN was 85.33%,85.33%,82.66%,81.33%,81.33% and 82.66% respectively.

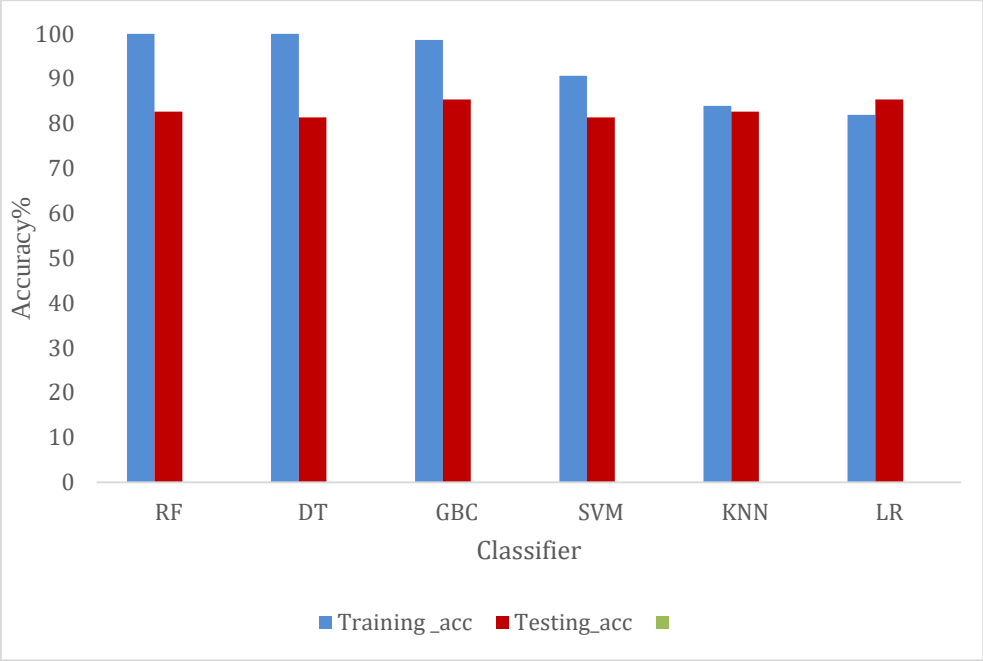


Figure 6.10: Training and testing classification accuracy

In figure (6.10) showed graphical representation of training and testing accuracy. Observation showed that overall, LR and GBC give the highest training accuracy

compared to other classifiers of 85.33%. All other classifiers also showed an accuracy above 80%.

Model performance can be improved using the K-fold cross-validation method; hence K=10 fold and the measured accuracy of all six classifiers are shown in figure (6.11).

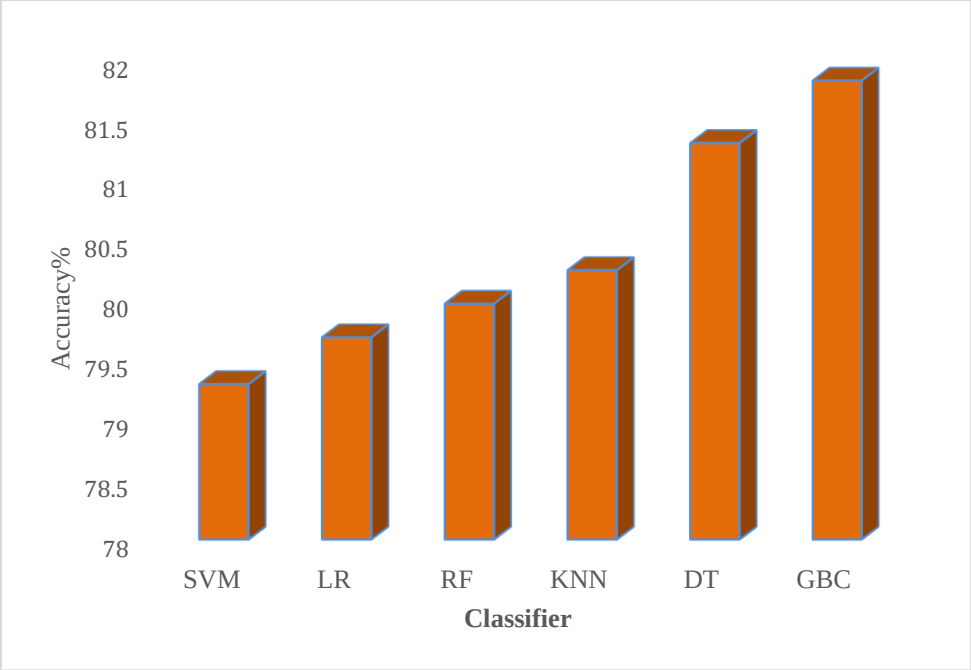


Figure 6.11: Analysis of k-fold accuracy

Figure(6.11) showed that SVM, LR, RF, KNN, DT and GBC gives an accuracy of 79.29%, 79.68%, 79.96%, 80.24%, 81.3% and 81.82% respectively. SVM gives the lowest accuracy, while the highest k-fold accuracy is shown by the GBC classifier of 81.82%.

6.5.1 Analysis of Feature Selection Methods

To enhance the model performance, feature selection is an important step that extracts the important feature. Feature selection techniques such as filter method-select-k best method, correlation, feature importance, information gain and RFE are applied for model development. Based on the number of selected features provided, the best classification accuracy is shown in Table (6.15).

Table 6.15: Classification accuracy based on the number of selected features

Method used	No of selected features	Accuracy (%)
DT	17	81.3
DT+ K-Best	8	76.22
DT+ Information gain	12	76.74
DT +Feature importance	13	78.07
DT +Correlation	11	77.04
DT+RFE	**10**	**81.57**
RF	17	79.96
RF+ K-Best	11	80.24
RF+ Information gain	12	80.51
RF +Feature importance	13	80.53
RF +Correlation	11	78.62
RF+RFE	**9**	**80.75**
KNN	17	80.24
KNN+ K-Best	**10**	**80.24**
KNN+ Information gain	12	80.24
KNN +Feature importance	14	80.24
KNN+ Correlation	13	80.24
GBC	17	81.82
GBC+ K-Best	11	79.16
GBC+ Information gain	12	80.23
GBC +Feature importance	11	79.96
GBC+ Correlation	12	78.62
GBC+RFE	**10**	**82.92**
LR	17	79.68
LR+ K-Best	8	79.95
LR+ Information gain	6	80.48

LR+ Feature importance	**5**	**80.48**
LR+ Correlation	11	80.48
LR+RFE	14	80.24
SVM	17	79.29
SVM + K-Best	8	81.31
SVM + Information gain	**11**	**82.39**
SVM +Feature importance	12	81.85
SVM + Correlation	10	81.31
SVM +RFE	10	80.23

A comparative analysis of the feature selection method is shown in figure (6.12).

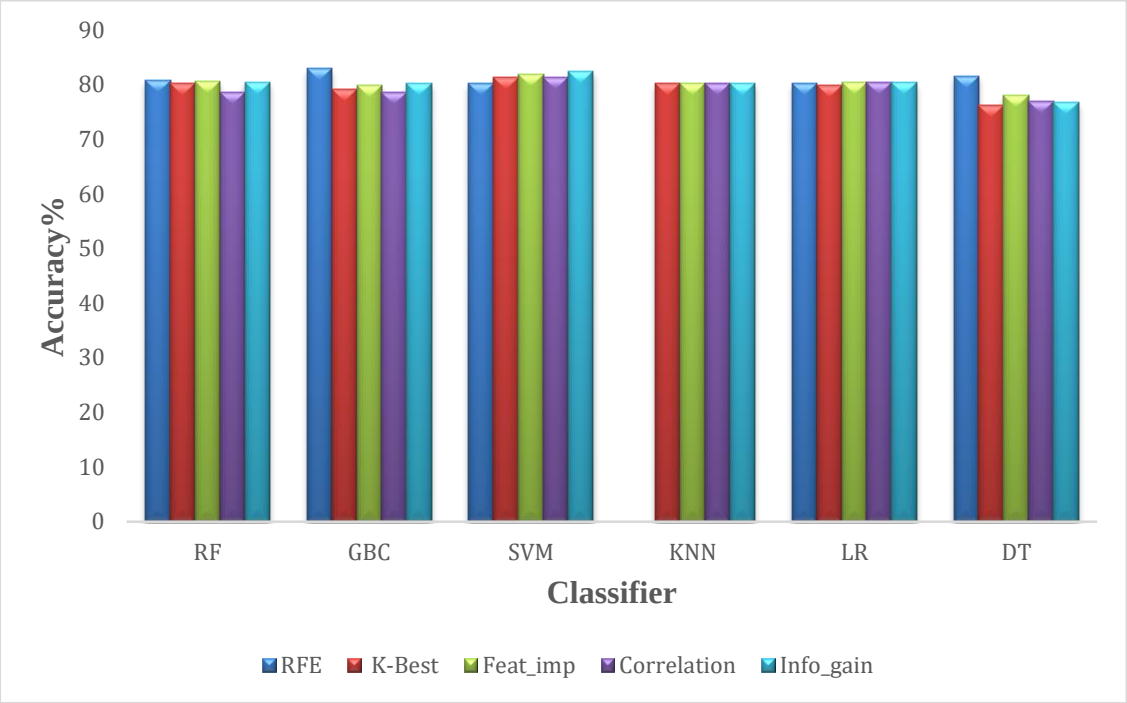


Figure 6.12: Comparative analysis of different feature selection methods

Figure (6.12) showed that applying five feature selection techniques with 10-fold cross-validation, the RFE method showed the best result on three different classifiers, DT, RF and GBC, by achieving an accuracy of 81.57%, 80.75% and 82.92%, respectively. KNN provided an accuracy of 80.24% with the select k-best method. SVM showed the highest accuracy for the information gain method by achieving an accuracy of 82.39% and LR showed the best accuracy with the feature

importance method of 80.48%. Among all classifiers, the highest accuracy showed by GBC of 82.92%.

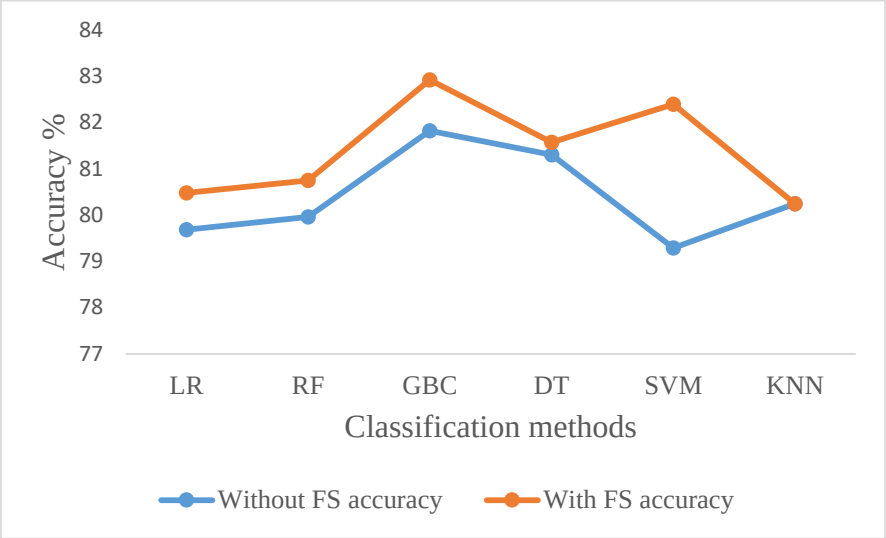


Figure 6.13: Comparison with feature selection and without feature selection

Figure (6.13) illustrates graphical analysis by applying feature selection and without feature selection accuracy. The comparable result showed that by applying the feature selection technique classification accuracy had been significantly improved.

6.5.2 Experimentation Result

Model tuning helps to improve performance. To improve the classifier performance, we have used the hyperparameter tuning optimization technique. The classification accuracy of various classifiers after applying grid search based hyperparameter tuning optimization showed in Table (6.16).

Table 6.16: Classification accuracy of hyperparameter tuning

Classifier	Accuracy (%)	Error Rate (%)
KNN	85.33	14.67
RF	85.33	14.67
DT	84	16
LR	85.33	14.67
GBC	84	16
SVM	81.33	18.67

Table showed that the highest accuracy of 85.33% and the minimum error rate of 14.67% were found by RF, LR and KNN classifiers. The comparison of feature

selection technique and hyperparameter tuning accuracy is shown graphically in figure (6.14) below.

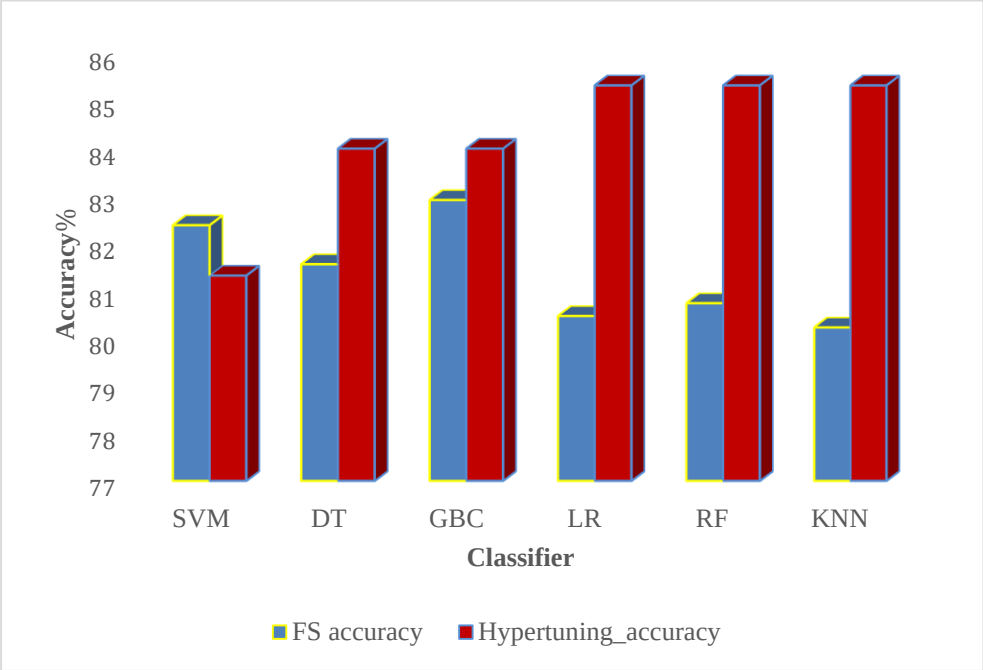


Figure 6.14: Comparison of feature selection and hyperparameter tuning accuracy

Experimental analysis showed that applying using grid search based hyperparameter tuning method showed significant improvement in terms of model accuracy. Highest feature selection accuracy observed in GBC of 82.92% while highest parameter tuning accuracy showed by LR,KNN and RF of 85.33%.The proposed method showed that overall improvement in model performance in terms of accuracy.

Accuracy based comparison of the K-fold, hyperparameter and feature selection method is shown in the below figure (6.15).

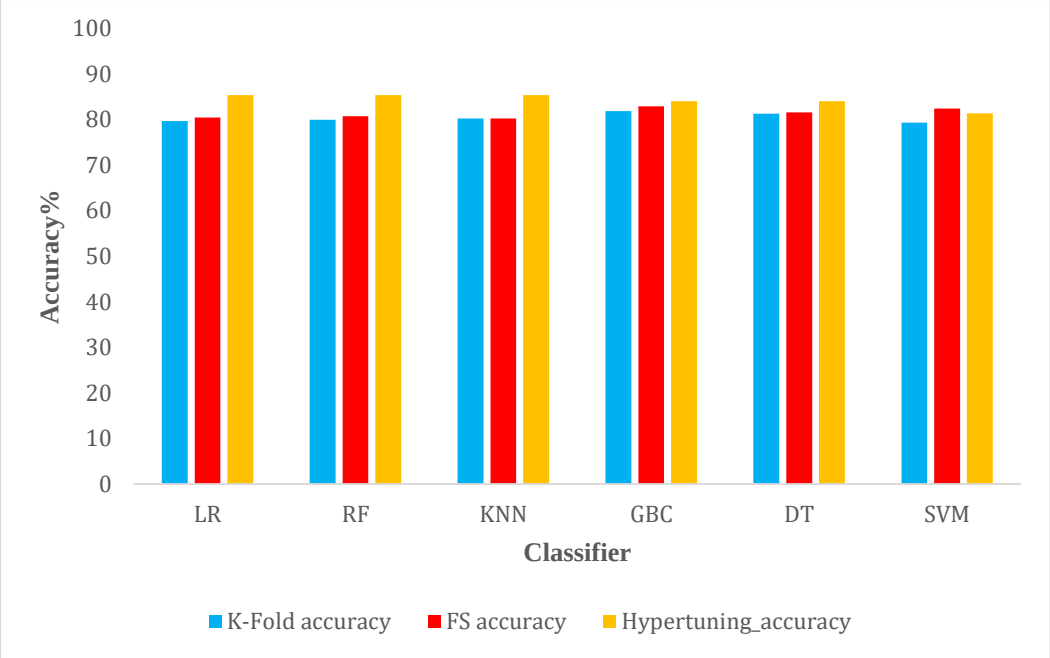


Figure 6.15: Comparison of K-fold, feature selection and hyperparameter tuning

Figure (6.15) showed accuracy was improved by 4.85%, 4.58%, 5.09%, 1.08%, 2.43% for LR, RF, KNN,GBC ,DT classifier as compared to feature selection accuracy. A similar observation was seen between k-fold and hyperparameter tuning accuracy. The rise was by 5.65%, 5.37%, 5.09%, 2.18%, 2.7% and 2.04% for LR, RF, KNN, GBC, DT,SVM classifiers respectively. The outcome demonstrates that hyperparameter tuning enhances classifier performance.

Table (6.17) shows the best hyperparameter based model performance.

Table 6.17: Classifier based best hyperparameters

ML Method	Best hyperparameter	Performance
RF	Criterion= gini max_ depth=4 ,max_features= auto, n_estimators= 200	**92.8%**
GBC	learning_rate=0.1, min_samples_split=37min_samples_leaf=6, max_features=sqrt n_estimators=30, subsample=0.8	92.49%
LR	C= 1, multi_class= ovr,penalty=l2 solver= newton-cg	87.52%
K-NN	Algorithm=auto,metric=minkowski leaf_size=30,n_neighbors=7 p=2, weights=uniform	85.55%
DT	Criterion= 'gini', max_depth= 2, min_samples_split= 2	82.5%
SVM	C=10, kernel=rbf,degree=2, gamma=0.001	81.74%

The experimental result showed that the best parameters observed when a criterion is gini with a max depth is 4 and the number of estimators is 200. These hyperparameters performs best on RF classifier with an AUC value of 0.928.Comparison of ROC curve analysis of various classification algorithms shown, figure (6.16).

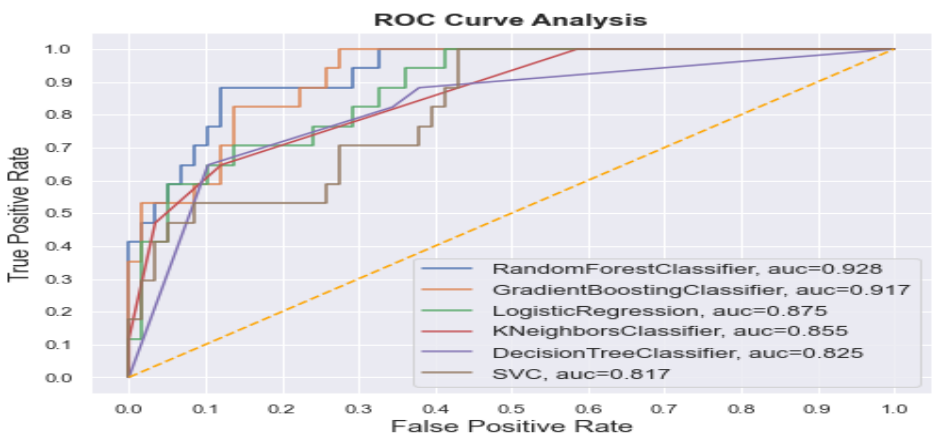


Figure 6.16: Comparison of ROC curve analysis

Figure (6.16) showed GBC and RF gives an AUC value of 0.9249 and 0.928. RF model provides the best performance Observation showed RF and GBC gives best performance. All classifiers showed the AUC value is above 0.8.

Performance measures was done by precision, accuracy, recall, f-measure and ROC-AUC curve. Analysis of six classifiers on various measures for analyzing the effect of lifestyle parameters is shown in Table (6.18)

Table 6.18: Performance metric of classifiers

Classifier	Result (%)				
	Precision	**Recall**	**F1-score**	**Roc-Auc score**	**Accuracy**
RF	**88**	**85**	82	**0.928**	**85.33**
GBC	84	84	81	0.9249	84
LR	86	**85**	83	0.8752	**85.33**
K-NN	85	**85**	**84**	0.8555	**85.33**
DT	84	84	**84**	0.825	84
SVM	80	81	80	0.8174	81.33

Table (6.18) showed evaluation of various classifiers using performance metrics shows that RF, LR and KNN model gives an accuracy of 85.33%. Precision is nothing but a measure of correctness.RF gives the highest precision value of 88%. Recall is the measure of exactness or completeness; here, LR, KNN and RF models showed a good recall value of 85%, KNN and DT show the highest f1-score value with a value of 84%.

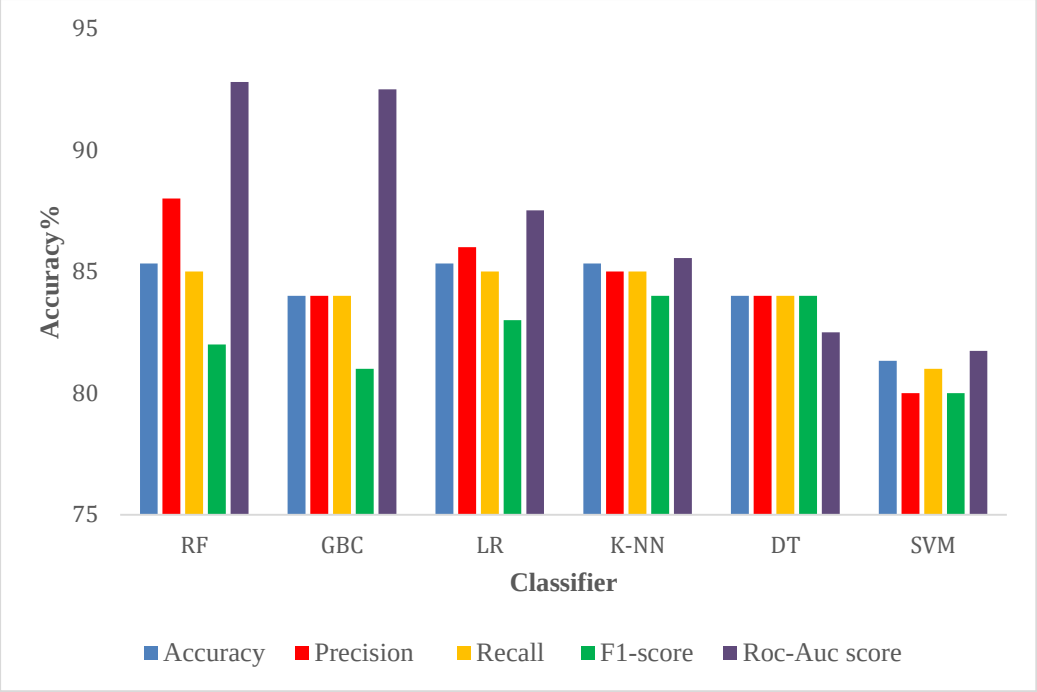


Figure 6.17: Graphical analysis of various measures

Figure (6.17) depict a graphical representation of performance measure.All model showed high predictive performance, with the roc-auc value ranging from 0.928 to 0.817. Compared with other models, the best performance showed by the RF model, with an AUC value of 0.928. The outstanding performance of the RF model helps in the assessment of the risk of developing type 2 diabetes.

6.6 Comparative Analysis of Type-2 Diabetes Prediction for LS_diabetes Dataset and Diabetes_2019 Dataset for Lifestyle parameters

6.6.1 Experimental Validation

The proposed Diabe_Lifestyle method was applied to the Diabetes_2019 dataset and we have compared it with the existing researcher adopted methodology result and our result is shown below in Table (6.19).

Table 6.19: Comparison of existing methodology vs. proposed methodology

Classifier	Proposed Method	[86]
DT	**94.24**	84
RF	**96**	94.1
SVM	**92.67**	86.5
LR	**85.86**	85.7
KNN	**93.71**	77.3
GB	94.24	-

In table (6.19) we observed that our methodology gives an exceptional result in comparison with existing researcher adopted method on the similar dataset.

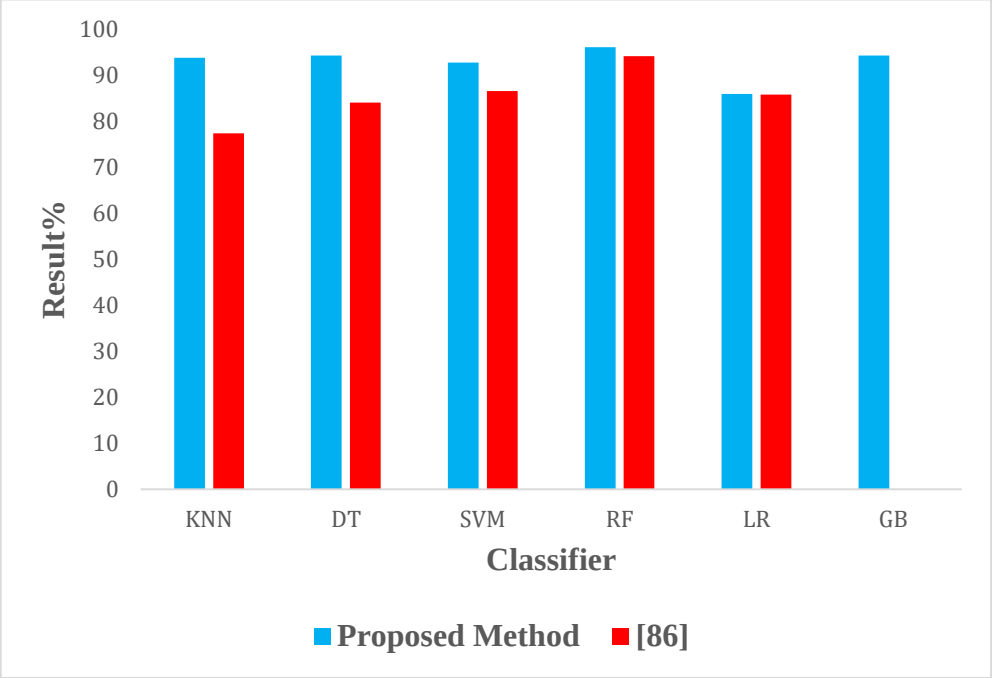


Figure 6.18: Accuracy comparison of proposed Diabe_lifestyle method with existing method

Figure (6.18) showed the comparative analysis. Experimental analysis showed that the proposed Diabe_Lifestyle methodology improved accuracy by 16.41%, 10.24%, 6.17%, 1.9%, 0.16%. in KNN,DT,SVM,RF and LR classifiers.

6.7 Performance Analysis of Type-2 Diabetes Prediction Using IELDR Model for LS_Diabetes Dataset

6.7.1 Experimentation Result

The use of a single algorithm approach for the prediction of diabetes would not provide a reliable and efficient prediction system; hence to improve the model accuracy, we have proposed a novel autoencoder based ensemble framework – IELDR. IELDR model applied on LS_diabetes dataset containing 374 records with 35 features. Accuracy based comparison with different classifiers on LS_diabetes dataset shown in Table (6.20).

Table 6.20: Accuracy and error rate comparison vs. model

Classifier	Accuracy (%)	Error Rate (%)
IELDR	**98.67**	**1.33**
MLP	90.67	9.33
SVC	90.67	9.33
LR	88.00	12.00
LDA	86.67	13.33
Adaboost	86.67	13.33

Top five model-based comparisons showed in figure (6.19).

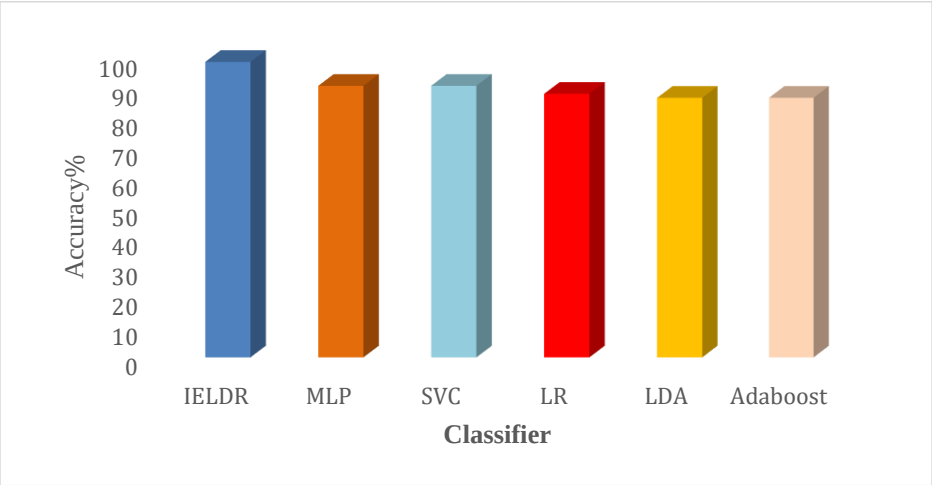


Figure 6.19: Accuracy comparison of top five model

The proposed IELDR model achieved the highest accuracy of 98.67% The result showed that the proposed IELDR model showed lowest error rate of 1.33%. In comparison with MLP and SVC, IELDR improved accuracy by 8%.

This research work uses LS_diabetes datasets associated with non-diabetes and diabetic population and employs different approaches using evaluation metrics like accuracy, recall, precision,f1-score, Mathew correlation coefficient, roc-auc for the performance evaluation, and comparison among selected classifiers. Performance comparison of different classifiers on various measures is shown in Table (6.21).

Table 6.21: Comparison of IELDR model with different classifier models on various measure

Model	Accuracy	Precision	Sensitivity	Specificity	f1 - score	ROC	Log_Loss	Mathew_ corrcoef
IELDR	**98.67**	0.9524	**100**	0.9818	**0.9756**	**0.9909**	**0.46**	**0.9670**
MLP	90.67	0.8421	0.8	0.9455	0.8205	0.8727	3.22	0.7580
SVC	90.67	1.0000	0.65	1.0000	0.7879	0.8250	3.22	0.7594
LR	88.00	0.9231	0.6	0.9818	0.7273	0.7909	4.14	0.6797
LDA	86.67	0.8125	0.65	0.9455	0.7222	0.7977	4.61	0.6428
Adaboost	86.67	0.8125	0.65	0.9455	0.7222	0.7977	4.61	0.6428
RF	85.33	1.0000	0.45	1.0000	0.6207	0.7250	5.07	0.6124
KNN	85.33	1.0000	0.45	1.0000	0.6207	0.7250	5.07	0.6124
GBM	84	0.7857	0.55	0.9455	0.6471	0.7477	5.53	0.5623

Table (6.21) illustrates the comparative performance measures of various classifiers, including MLP, SVC, LR, LDA, Adaboost, RF, KNN, GBM and IELDR classifiers. The results suggest that the IELDR classifier yields the highest accuracy of 98.67%. If the precision and specificity are considered, SVC, RF and KNN classifier yields a value of 100%. If the recall is considered, the IELDR classifier yields 100%. The IELDR classifier provides better result of f1-score, ROC value, Log_loss and MCC value of 97.56%, 0.9909, 0.46 and 0.967 respectively.

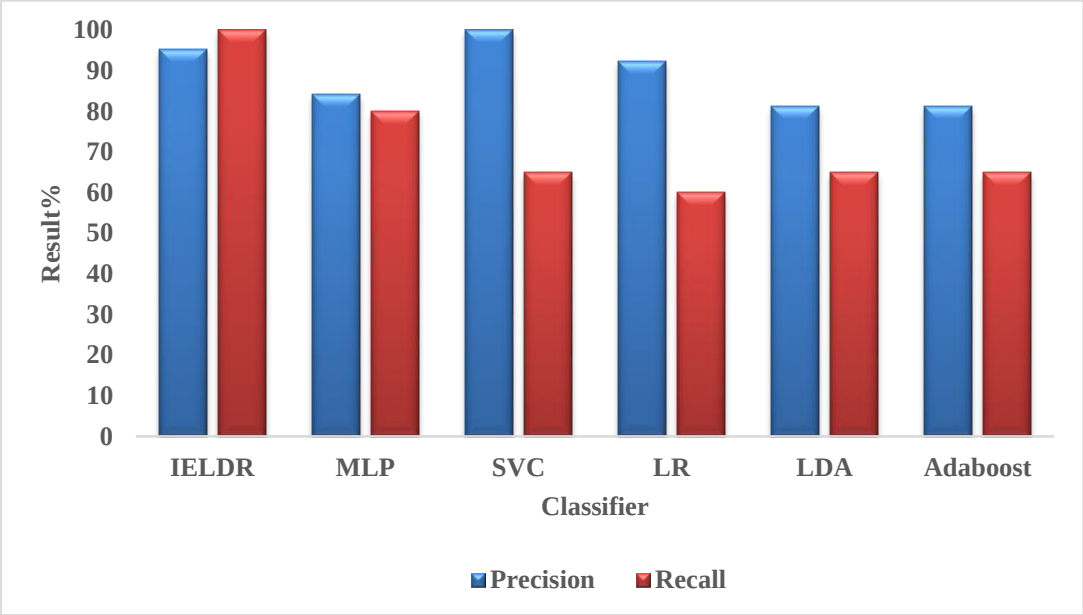


Figure 6.20: Comparison of precision vs. recall

Figure (6.20). Illustrates the results of the precision and recall comparison of the MLP, SVC, LR, LDA, and the discussed IELDR classifier. It can be concluded from the results that the discussed IELDR classifier yields much better precision results of 95.24%. If the recall is considered, the discussed IELDR classifier yields recall results of 100%, while the available LR renders just 60%.

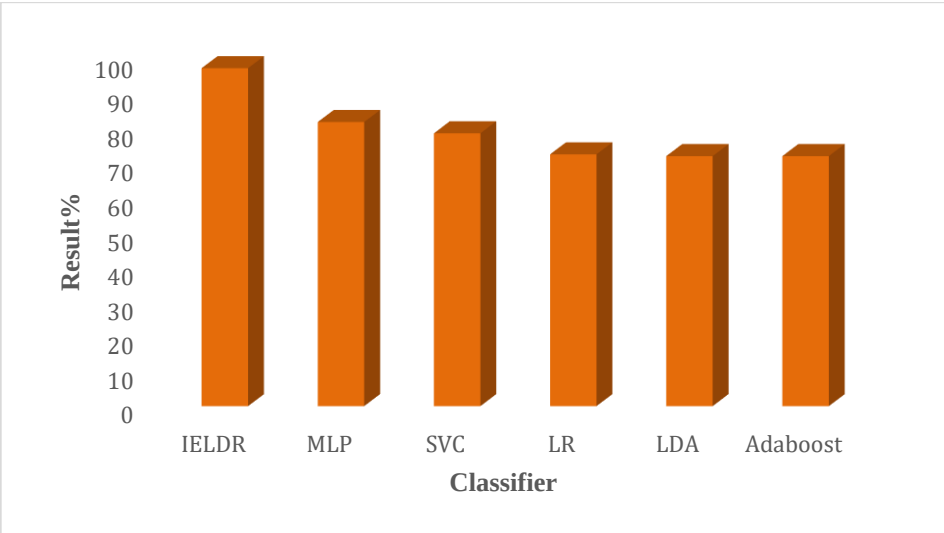


Figure 6.21: Comparative analysis of f1-score for various classifiers

Figure (6.21) illustrates the results of the F1-score comparison of the available MLP, SVC, LR, LDA, Adaboost and the discussed IELDR classifier. It can be concluded

from the results that the discussed IELDR classifier yields better results of 97.56%, while the available algorithm MLP produces 82.05%, SVC gives 78.79%, LR gives 72.73% and %. LDA and Adaboost render just 72.22%.

6.7.2 Recommendation

The system predicts the risk for developing type 2 diabetes in a healthy person based on the person's current lifestyle pattern; the system calculates the risk score and offers three types of recommendations.

1]

```
--------------------------
Diabetes Risk Analysis
--------------------------
Your health score is good;follow the healthy practices
```

2]

```
------------------------
Diabetes Risk Analysis
------------------------
You are Predicted On the way to developing diabetes ,if you don,t change lifestyle.Your
diabetes risk index is 44.54/50
```

3]

```
--------------------------
Diabetes Risk Analysis
--------------------------
You are predicted having diabetes so you are on higher risk,kindly consult with
doctors.Your diabetes risk index is 49.58/50
```

6.8 Comparative Analysis of Type-2 Diabetes Prediction Using IELDR Model with PIMA dataset

6.8.1 Experimental Validation

The proposed IELDR model was evaluated on the PIMA Indian diabetes dataset collected from the UCI repository and the LS_diabetes dataset. IELDR model

validated on PIMA dataset and comparative analysis was done with LS_diabetes dataset mentioned in Table (6.22).

Table 6.22: Comparative analysis of PIMA dataset vs. LS_diabetes dataset

Evaluation Type	Result%	
	PIMA_Diabetes[107]	LS_Diabetes
Accuracy	80.71	**98.67**
Precision	82.09	**95.24**
Sensitivity	59.78	**100**
Specificity	92.59	**98.18**
F1 score	69.18	**97.56**
ROC	76.19	**99.09**
MCC	0.57	**0.97**

Experimental observation showed that best result observed on LS_diabetes dataset with an accuracy, precision, sensitivity, specificity, f1-score,roc and mcc value by 17.96%,13.15% 40.22%,5.59%,28.38%,22.09% and 0.4 respectively.

Figure (6.22) illustrates an accuracy based comparison after applying a novel IELDR framework to the PIMA dataset and LS_diabetes dataset.

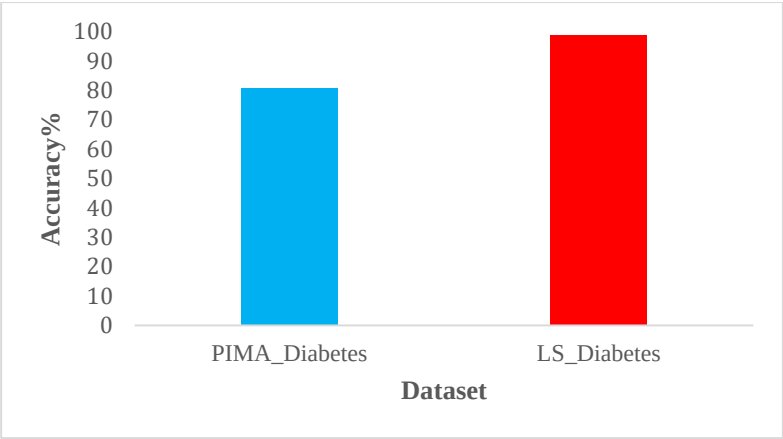


Figure 6.22: Accuracy based comparison with PIMA dataset

The result showed that the IELDR model achieves the highest accuracy on the LS_diabetes dataset at 98.67%.

6.8.2 Comparative Analysis

Case I-

A comparison of the proposed IELDR model and the best model experimented on the PIMA Indian diabetes dataset from the state of the art is shown in Table (6.23).

Table 6.23: Proposed IELDR model vs best model on PIMA dataset

Case1-[PIMA[107]]				
Evaluation Method	**IELDR**	**[28]**	**[34]**	**[110]**
Accuracy	**80.71**	76.3	79.08	75
Precision	**82.09**	75.9	73.13	47
Specificity	**92.59**		83.44	76

The result showed that accuracy, precision and specificity rise by 1.63%, 6.19% and 9.15% from the existing researcher. Highlighted result showed that the proposed IELDR method gives good results.

Figure (6.23) graphical comparison of various measures.

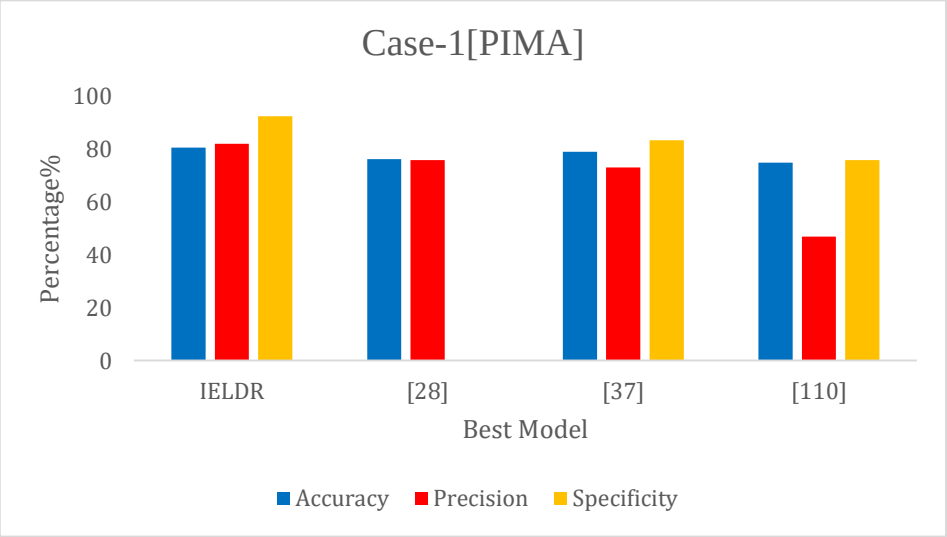


Figure 6.23: Proposed IELDR model vs best model for case-I

Comparative analysis on various measure showed that the IELDR model highlighted the best result compared to other researchers best models employed on the PIMA dataset. The result showed that the IELDR model achieved the highest accuracy, precision and specificity values of 80.71%, 82.09% and 92.59%, respectively.

6.9 Comparative Analysis of Type-2 Diabetes Prediction Using IELDR Model with Diabetes_2019 dataset

6.9.1 Experimental Validation

The proposed IELDR model was evaluated on the Diabetes_2019 dataset collected from the researcher and the LS_diabetes dataset. Below Table (6.24) showed performance measure of two datasets.

Table 6.24: Comparative analysis of LS_diabetes dataset vs. Diabetes_2019 dataset

Evaluation Type	Result%	
	Diabetes_2019[86]	LS_Diabetes
Accuracy	96.85	**98.67**
Precision	96.89	95.24
Sensitivity	98.42	**100**
Specificity	98.42	98.18
F1 score	97.65	97.56
ROC	96.08	**99.09**
MCC	0.929	**0.97**

The experimental result showed that the IELDR model gives the best result on the LS_diabetes dataset with performance measures such as accuracy, sensitivity, roc and mcc value by 1.82%, 1.58%, 3.01%and 0.04 % compared to the Diabetes_2019 dataset.

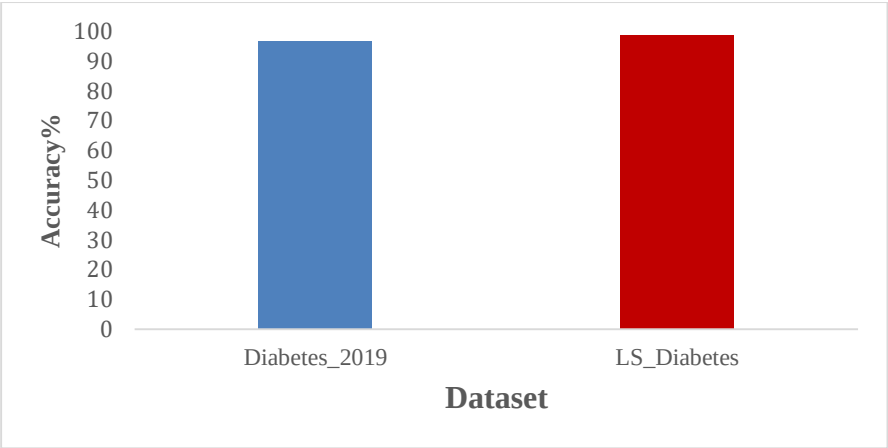


Figure 6.24: Accuracy based comparison with Diabes_2019 dataset

Figure (6.24) illustrates an accuracy based comparison after applying a novel IELDR framework to the Diabetes_2019 dataset and LS_diabetes dataset showed that the IELDR model achieves the highest accuracy on the LS_diabetes dataset of 98.67%, while Diabes_2019 gives an accuracy of 96.85%.

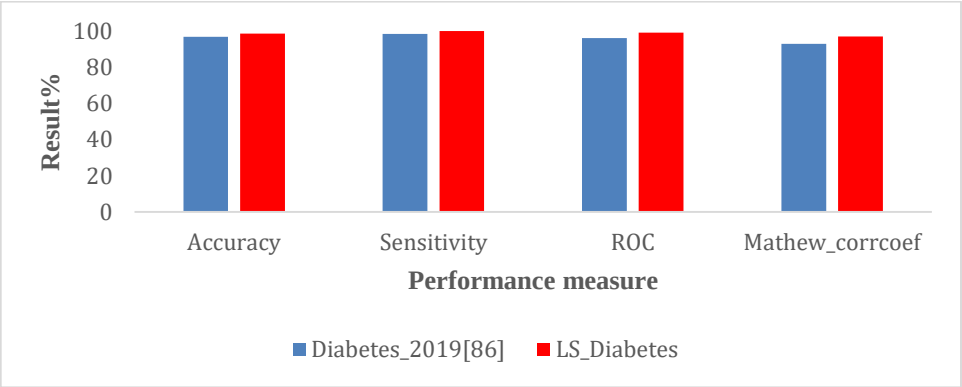


Figure 6.25: Performance comparison of IELDR method on LS_diabetes dataset vs. Diabetes_2019 dataset

Figure (6.25) illustrates a performance measure based comparison after applying a novel IELDR framework to the LS_diabetes dataset and LS_diabetes dataset. The result showed that the IELDR model achieves the highest accuracy, sensitivity, roc and mcc value on the LS_diabetes dataset of 98.67%, 100%, 99.09% and 97% respectively.

6.9.2 Comparative Analysis

Case II-

The proposed IELDR based framework and the framework used by existing researchers on the diabetes_2019 dataset have been compared and found to be IELDR framework, which gives the good result shown in Table (6.25).

Table 6.25: Proposed IELDR methodology vs. methodology used on Diabetes_2019 dataset

Case2-[Diabetes_19][86]			
Evaluation Method	**IELDR model**	**[86]**	**[35]**
Accuracy	**96.85**	94.1	96.81
Precision	**96.89**	97.6	96
Sensitivity	**98.42**	94.3	92.3
Specificity	98.42	93.4	98.52
F1- score	**97.65**	95.9	94.11

The result showed that the IELDR model gives the highest accuracy, precision, sensitivity and f1-score with a value of 96.85%, 96.89%, 98.42% and 97.65% respectively. Specificity value is also approximately similar to an existing researcher.

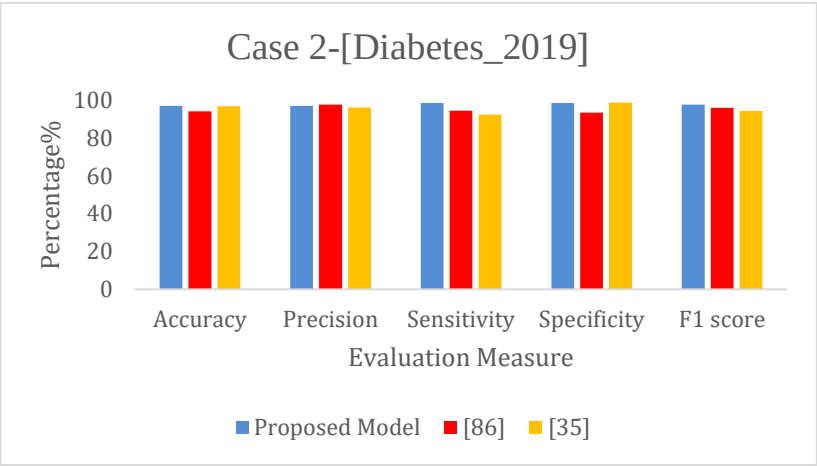


Figure 6.26: Proposed IELDR model vs best model for case-II

Figure (6.26) illustrates a graphical comparison of various measures, highlighting proposed IELDR method gives good results. Comparative analysis showed that the IELDR model showed the best result in comparison with other researchers best model employed on the Diabetes_2019 dataset.

6.10 Summary

This chapter discusses the findings and discussion of the research work and comparative analysis of the proposed methodology with various classifiers. It was observed that reducing the number of dimensions greatly impacts the performance of the models. The results and findings depict stress and lifestyle as one the important risk factors for developing type 2 diabetes. The performance results show that the proposed IELDR is efficient with high classification accuracy for classifying the diabetes dataset, which is applicable for further diagnosis. The conclusion of the thesis and the future direction of the present work are discussed in the subsequent chapter.

CHAPTER-7

CONCLUSION AND FUTURE SCOPE

Chapter 7 CONCLUSION AND FUTURE SCOPE

7.1 Conclusion

Type 2 diabetes is a health concern worldwide, including in India, because of its severe complications, direct and indirect costs and psychological burden associated with it. Uncontrolled disease and delayed treatment can be associated with several adverse implications affecting organs like kidneys, eyes, nerves and heart. The disease affects productive age groups and older people, posing a significant economic burden on the family. With the increasing burden of diseases and complexity of health-related data and other challenges in the diagnosis of disease, applications of ML are evolving. ML can-not replaces manpower but has significant potential for assisting clinicians in the accurate and timely diagnosis of diseases.

The LS_Diabetes dataset was developed by collecting records from 374 people with 35 features. Using this dataset, the proposed framework examined the performance of six different machine learning models, SVM, LR, RF, DT, GBC and KNN, applying five different feature selection methods: RFE, Select-K best method, Feature importance technique, Information gain and Correlation method. Grid search with cross-validation was used to optimize hyperparameters. The proposed D_Stress method was validated against on PIMA database from the UCI repository and the Diabetes_2019 database from another Indian researcher. Using the LS_Diabetes dataset, we used stress related parameters to analyze the effect of stress. Select k best method with SVM classifier showed the highest feature selection accuracy of 82.14%. Machine learning classifiers, GBC and SVM, showed accuracy, precision, recall and f1-score value of 86.67%, 87%, 87% and 85%, respectively, for the prediction of type 2 diabetes. The highest roc value shown by SVM and RF was 91.18% and 91.17%, respectively.

The impact of COVID-19 related stress on blood glucose levels in type-2 diabetes patients was examined using a dataset of 199 diabetes patients. The results of the COVI-Stress dataset showed a significant increase in blood glucose levels in type-2 diabetes patients after discharge of COVID-19 treatment as compared to before

admission. The rise in blood glucose was significant in patients with type-2 diabetes without stress and those with moderate stress. Age wise significant difference was found in the old age group people, suggesting the possible impact of the psychological stress of higher mortality and other complications among elderly type 2 diabetic patients with COVID-19. Overall, gender wise as well as residence wise, found a significant difference in FBG and PPBG.

The effect of lifestyle based parameters was also analyzed using the LS_Diabetes dataset. The proposed Diabe_Lifestyle framework examined the performance of six different machine learning models, SVM, LR, RF, DT, GBC and KNN, by applying five different feature selection methods: select-k best method, feature importance technique, information gain, correlation technique and RFE. Hyperparameter optimization was done using grid search cross-validation. The highest feature selection accuracy of 82.92% was observed in the GBC with the RFE method. Based on the comparative analysis, RF and GBC showed the highest ROC value of 92.8% and 92.49%.

A novel- Improved Ensemble Learning with Dimensionality Reduction Model (IELDR) was designed. Autoencoder based feature extraction method along with an ensemble learning algorithm was used. The research work was validated on two benchmark datasets. IELDR model achieved highest accuracy, precision, sensitivity, specificity, f1 score, roc and mcc value by 17.96%, 13.15% 40.22%, 5.59%, 28.38%, 22.09% and 0.4 compared to PIMA dataset. Comparative to the Diabetes_2019 dataset, the IELDR model provided the highest accuracy, sensitivity, roc and mcc value by 1.82%, 1.58%, 3.01% and 0.04 on the LS_Diabetes dataset.

This proposed IELDR system can effectively predict the risk of type 2 diabetes development in healthy people based on the current lifestyle pattern. The system calculates the risk score and offers a recommendation.

Considering the rising recognition of ML in healthcare, the findings of this study may be useful in predicting type 2 diabetes. Overall, the results of our research and validation with other datasets provide significant insights for predicting type 2 diabetes development. Stress and lifestyle should be considered strong risk factors

for the development of DM and appropriate strategies should be designed to control these risk factors.

7.2 Future Scope

Studies with a larger sample size of type-2 diabetes patients under stress may be necessary to shed more light on the distinction between those under moderate to severe/extremely severe stress and those who are not. Further large scale studies covering wider population in India are needed to confirm these observations. Studies for estimation of the duration required for the development of type 2 diabetes based on the lifestyle of a person and prediction of organs at higher risk of development of complications can also be planned.

REFERENCES

REFERENCES

[1] Diabetes http://www.who.int/en/news-room/fact-sheets/detail/diabetes accessed on 21st Feb 2019

[2] IDF SEA members https://www.idf.org/our-network/regions-members/south-east-asia/.../94-india.html assessed on 22nd Feb 2019

[3] ICMR Guidelines for management of type 2 diabetes 2018 https://main.icmr.nic.in/sites/default/files/guidelines/ICMR_GuidelinesType2diabetes2018_0.pdf accessed on 23July 2021

[4] Samuel AL. Eight-move opening utilizing generalization learning. (See Appendix B, Game G-43.1 Some Studies in Machine Learning Using the Game of Checkers. IBM Journal 1959; 210–229.

[5] Tom M. Machine learning. First Edition; McGraw- Hill 1997.

[6] Dutt S, Das AK. Machine learning. Pearson Education India.

[7] Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning, and health systems. In the Journal of Global Health 2018; 8. https://doi.org/10.7189/jogh.08.020303

[8] Patil R, Shah K. Chapter 6 Machine learning in healthcare: Applications, current status, and future prospects. Handbook of Research on Machine Learninh. CRC press 2022 India.

[9] Jain V, Chatterjee J. Machine Learning with Healthcare Perspective", Springer Science and Business Media LLC 2020; Vol-13.

[10] Uddin S.Khan A, Hossain ME, et al. Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making 2019; 19(1): 1–16. https://doi.org/10.1186/s12911-019-1004-8

[11] Panesar A. Machine learning and AI for healthcare. In Machine Learning and AI for Healthcare. 2019. https://doi.org/10.1007/978-1-4842-3799-1

[12] Wuest T, Weimer D, Irgens C, et al. Machine learning in manufacturing: Advantages, challenges and applications. Production & Manufacturing Research 2016; 4(1): 23-45.

[13] Rahmani AM, Yousefpoor E, Yousefpoor MS, et al. Machine learning (ML) in medicine: Review, applications, and challenges. Mathematics 2021; 9: 2970. https:// doi.org/10.3390/math9222970

[14] Svalastog AL, Donev D, Jahren KN, et al. Concepts and definitions of health and health-related values in the knowledge landscapes of the digital society. Croat Med J. 2017;58:431-435.

[15] India State-Level Disease Burden Initiative Diabetes Collaborators* The increasing burden of diabetes and variations among the states of India: the Global Burden of Disease Study 1990–2016 Lancet Glob Health 2018; 6: e1352–62.

[16] ICMR-INdia DIABetes[INDIAB] study Phase-I executive summary https://main.icmr.nic.in/sites/default/files/reports/Executive_summary_INDIAB_Phase_I.pdf accessed on 20July 2021

[17] Pradeepa R, Mohan V. Epidemiology of type 2 diabetes in India. Indian J Ophthalmol 2021;69:2932-8

[18] Raji OS. Diabetes mellitus in relation to life style and stress-An analytical study. Kerala University Library Thiruvananthapuram 2013.

[19] Bennasar M, Hicks Y, Setchi R. Feature selection using joint mutual information maximisation. Expert Systems with Applications 2015; 42: 8520-8532.

[20] Haq A ,Li JP, Khan J, et al. Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data. Sensors 2020;20: 2649

[21] Kumar V, Minz S. Feature Selection: A literature Review. *Smart Computing Review* 2014; 4; 3.

[22] Chandrashekar G, Sahin, F. A survey on feature selection methods. Computers & Electrical Engineering 2014;40: 16-28.

[23] Charte D, Charte F, García S, et al. A practical tutorial on autoencoders for nonlinear feature fusion: taxonomy, models, software and guidelines. Inf. Fusion 2018;44:78–96.

[24] Contreas I, Vehi J. Artificial intelligence for diabetes management and decision support: Literature review. J Med Internet Res 2018. 20: e10775.

[25] Rigla M, García-Sáez G, Pons B, et al. Artificial intelligence methodologies and their application to diabetes. J Diabetes Sci Technol 2018;12:303–310.

[26] Perveen S, Shahbaz M, Gurgachi A, et al. Performance analysis of data mining classification techniques to predict diabetes. Proc Comput Sci 2016; 82:115–121.

[27] Jayanthi N, Babu BV, Rao NS. Survey on clinical prediction models for diabetes prediction. J Big Data 2017; 4:26.

[28] Sisodia D. Sisodia D. Prediction of diabetes using classification algorithms: Procedia Computer Science 2018; 132:1578–1585.

[29] Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. Proc. Comp. Sci. 2019;165:292–299.

[30] Zou Q, Qu K, Luo Y, et al. Predicting Diabetes Mellitus With Machine Learning Techniques. Front. Genet. 2018; 9:515.

[31] Lai H, Huang H, Keshavjee K, et al. Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders 2019; 19: 1–9.

[32] Patil BM, Joshi RC, Toshniwal, D. Hybrid prediction model for type-2 diabetic patients. Expert Systems with Applications 2010; 37: 8102-08.

[33] Mahabub A. A robust voting approach for diabetes prediction using traditional machine learning techniques, SN Applied Sciences,Springer 2019; 1:1667.

[34] Kumari S, Kumar D, Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. International Journal of Cognitive Computing in Engineering 2021;2:40-6.

[35] Ahmed N, Ahammed R, Islam M, et al. Machine learning based diabetes prediction and development of smart web application. International Journal of Cognitive Computing in Engineering 2021; 2: 229-41

[36] Sivashankari R, Sudha M, Hasan MK, et al. An empirical model to predict the diabetic positive using stacked ensemble approach. Front. Public Health 2022; 9:792124.

[37] Birjais R, Mourya AK, Chauhan R, et al. Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Applied Sciences 2019;1:1112.

[38] Khaleel FA, Al-Bakry AM. Diagnosis of diabetes using machine learning algorithms. Mater. Today Proc. 20121. doi: 10.1016/j.matpr.2021.07.196

[39] Diwani SA. Sam A. Diabetes forecasting using supervised learning techniques. Adv. Comp. Sci. Int. J. 2014; 3:10–18.

[40] Khanam JJ, Foo SYA, A comparison of machine learning algorithms for diabetes prediction. ICT Exp. 2021; 7: 432–439.

[41] Wang Q, Cao W, Guo J, et al. DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values. In *IEEE Access* 2019; 7: 102232-38.

[42] Krishnamoorthi R, Joshi S, Hatim Z, et al. A novel diabetes healthcare disease prediction framework using machine learning techniques. Journal of Healthcare Engineering2022, Article ID 1684017, 10 Pages,2022. https://doi.org/10.1155/2022/1684017

[43] Dagliati A, Marini S, Sacchi L, et al. Machine learning methods to predict diabetes complications. Journal of Diabetes Science and Technology 2018; 12: 295-302.

[44] Chaki J, Ganesh ST, Cidham SK, et al, Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review, Journal of King Saud University – Computer and Information Sciences 2020; https://doi.org/10.1016/j.jksuci.2020.06.013

[45] Chen R, Dewi C, Huang S, et al. Selecting Critical Features for Data Classification Based on Machine Learning Methods. J Big Data 2020; 7:1-26.

[46] Kumar GR, Ramachandra GA, Nagamani K. An efficient feature selection system to integrating SVM with genetic algorithm for large medical datasets. International Journal 2014; 4: 272- 77.

[47] Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection: J Big Data 2019; 6:13.

[48] Hasan K, Alam A, Das D, et al. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 2020; 8.

[49] Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics 2022;18:90-100.

[50] Maniruzzaman Md, Rahman Md J, Ahammed B, et al. Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems 2020; 8:1-14.

[51] Maniruzzaman Md, Rahman Md J, Al-MehediHasan Md, et al. Accurate diabetes risk stratification using machine learning: Role of missing value and outliers. Journal of Medical Systems 2018; 42: 92.

[52] Dutta D, Paul D, Ghosh P. Analysing feature importances for diabetes prediction using machine learning. IEEE 9th Annual Information Technology,

Electronics and Mobile Communication Conference (IEMCON) 2018; 924-28. DOI: 10.1109/IEMCON.2018.8614871

[53] Mishra S, Chaudhury P, Mishra BK, et al. An implementation of feature ranking using machine learning techniques for diabetes disease prediction. Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies. 2016.

[54] Sivaranjani S, Ananya S, Aravinth J, et al. Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction, India: Amrita Vishwa Vidyapeetham, 2021.Conference: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)

[55] Deberneh HM, Kim I. Prediction of type 2 diabetes based on machine learning algorithm. Int. J. Environ. Res. Public Health. 2021; 18: 3317.

[56] Jian Y, Pasquier M, Sagahyroon A, et al. A Machine learning approach to predicting diabetes complications. Healthcare (Basel). 2021; 9: 1712.

[57] Le TM, Vo TM, Pham TN, et al. A novel wrapper–based feature selection for early diabetes prediction enhanced with a meta heuristic. *IEEE Access* 2020; 9: 7869-84.

[58] Gill S, Pathwar P. Prediction of Diabetes Using Various Feature Selection and Machine Learning Paradigms. In: Gunjan, V.K., Zurada, J.M. (eds) Modern Approaches in Machine Learning & Cognitive Science: A Walkthrough. Studies in Computational Intelligence 2022;, vol 1027. Springer, Cham.

[59] Saxena R, Sharma SK, Gupta M, et al. A novel approach for feature selection and classification of diabetes mellitus: Machine learning methods. Computational Intelligence and Neuroscience 2022, Article ID 3820360, 11 pages. https://doi.org/10.1155/2022/3820360

[60] Ismail L, Materwala H, Tayefi M. et al. Type 2 diabetes with artificial intelligence machine learning: Methods and evaluation. Arch Computat Methods Eng 2022; 29: 313–33.

[61] Kannadasan K, Edla D, Kuppili V. Type 2 diabetes data classification using stacked autoencoders in deep neural networks. Clinical Epidemiology and Global Health. 2019;530-35

[62] Falco G, Pirro PS, Castellano E, et al. The relationship between stress and diabetes mellitus. J Neurol Psychol. 2015;3: 7.

[63] Zamani-Alavijeh F, Araban M, Koohestani HR, et al. The effectiveness of stress management training on blood glucose control in patients with type 2 diabetes. Diabetol Metab Syndr. 2018 May 8;10:39.

[64] Harris ML, Oldmeadow C, Hure A, et al. Stress increases the risk of type 2 diabetes onset in women: A 12-year longitudinal study using causal modelling. PLoS ONE 12:e0172126.doi:10.1371/journal. pone.0172126

[65] Kelly SJ, Ismai M. Stress and type 2 diabetes: A review of how stress contributes to the development of type 2 diabetes. Annu. Rev. Public Health 2015. 36:441–62.

[66] Luthra D. A link between stress and diabetes. Webmed Central Psychiatry 2010; 1: WMC001095.

[67] Lloyd C, Smith J, Weinger K. Stress and diabetes: A review of the links. Diabetes Spectrum 2005;18:121-27.

[68] Kaur B, Bedi U. Role of stress in newly detected type 2 diabetes mellitus and hypertension. Int J Med and Dent Sci 2017;6:1493-1497.

[69] Sui H, Sun N, Zhan L, et al. Association between Work-Related Stress and Risk for Type 2 Diabetes: A Systematic Review and Meta-Analysis of Prospective Cohort Studies. PLoS ONE 2016; 11(8): e0159978.

[70] Brunner B, Igic I, Keller AC, et al. Who gains the most from improving working conditions? Health-related absenteeism and presenteeism due to stress at work. Eur J Health Econ 2019; 20:1165-80.

[71] Bhui K, Dinos S, Galant-Miecznikowska M, et al. Perceptions of work stress causes and effective interventions in employees working in public, private and non-governmental organisations: a qualitative study. BJPsych Bull. 2016;40:318-25.

[72] Kiranashree BK , Ambika C, Radhika AD. Analysis on machine learning techniques for Stress detection among employees. Asian Journal of Computer Science and Technology 2021;10: 35-7.

[73] Reddy US, Thota AV, Dharun A. Machine Learning Techniques for Stress Prediction in Working Employees. IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2018, 1–4.

[74] Ahuja R. Banga A. Mental stress detection in university students using machine learning algorithms. International Conference on Pervasive

Computing Advances and Applications (PerCAA 2019), Procedia Computer Science 2019;152: 349-53.

[75] Sanchez W, Martinez A, Hernandez Y, et al. A predictive model for stress recognition in desk jobs. Journal of Ambient Intelligence and Humanized Computing. 2018, https://doi.org/10.1007/s12652-018-1149-9

[76] Kumar P, Garg S, Garg A. Assessment of anxiety, depression and stress using machine learning models: Procedia Computer Science 2020; 171: 1989-98.

[77] Prediction of Stress Level on Indian Working Professionals Using Machine Learning Kavita Pabreja, Maharaja Surajmal InstituteInternational Journal of Human Capital and Information Technology Professionals Volume 13 • Issue 1DOI: 10.4018/IJHCITP.297077 ;2022

[78] Patil R, Shah K. Assessment of risk of type 2 diabetes mellitus with stress as a risk factor using classification algorithms. International Journal of Recent Technology and Engineering 2019; 8: 11273–77.

[79] Rajappa T, Ponniraivan K, Kalyan H, et al. Assessment of degree of awareness about diet, physical exercise, and lifestyle modifications among diabetic patients. Int J Med Sci Public Health 2018;7:481-48

[80] Kolb H, Martin S. Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. BMC Med. 2017;15:131.

[81] Deo R, Panigrahi S. Performance assessment of machine learning based models for diabetes prediction. 2019 IEEE Healthcare Innovations and Point of Care Technologies, (HI-POCT), 2019: 147-50.

[82] Nai-aruna N, Moungmaia R. Comparison of classifiers for the risk of diabetes prediction. 7th International Conference on Advances in Information Technology. Procedia Computer Science 2015; 69:132-142.

[83] Jahan N, Islam A, Mamun AA. Machine learning with factor scoring to predict diabetes risk level In Bangladesh. Int J of Scientific & Technology Research 2020; 9:2863-7.

[84] Olivera AR, Roesler V, Iochpe C, et al. Comparison of machine learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: Accuracy study. Sao Paulo Med J 2017;135:234-6

[85] Anand A, Shakti S. Prediction of diabetes based on personal lifestyle indicators,In :1st international conference on next generation computing technology:IEEE;2015

[86] Tigga N, Garg S. Prediction of type 2 diabetes using machine learning classification methods classification. International Conference on Computational Intelligence and Data Science, Procedia Computer Science 2020; 167: 706-16.

[87] Wang L, Wang X, Chen A, et al. Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model. Healthcare (Basel). 2020; 8: 247.

[88] Marcovecchio ML, Chiarelli F. The effects of acute and chronic stress on diabetes control. Sci Signal. 2012;5:pt10

[89] Schofield J, Leelarathna L, Thabit H. COVID-19: Impact of and on diabetes. Diabetes Ther. 2020;11:1429-1435.

[90] Hussain A, Bhowmik B, do Vale Moreira NC. COVID-19 and diabetes: Knowledge in progress. Duabetes Res Clin Pract. 2020; 162: 108142.

[91] Joensen LE, Madsen KP, Holm L, et al. Diabetes and COVID-19: Psychosocial consequences of the COVID-19 pandemic in people with diabetes in Denmark-what characterizes people with high levels of COVID-19-related worries?. Diabet Med. 2020; 37: 1146-1154.

[92] Sankar P, Ahmed WN, Koshv VM, et al. Effects of COVID-19 lockdown on type 2 diabetes, lifestyle and psychosocial health: A hospital-based cross-sectional survey from South India. Diabetes Metab Syndr. 2020; 14: 1815-1819.

[93] Ruissen MM, Regeer H, Landstra CP, et al. Increased stress, weight gain and less exercise in relation to glycemic control in people with type 1 and type 2 diabetes during the COVID-19 pandemic. BMJ Open Diab Res Care 2021; 9: e002035.

[94] Kautzky-Willer A. Does diabetes mellitus mitigate the gender gap in COVID-19 mortality? European Journal of Endocrinology 2021; 185:C13–C17.

[95] Yoshida Y, Gillet SA, Brown MI, Zu Y, et al. Clinical characteristics and outcomes in women and men hospitalized for coronavirus disease 2019 in New Orleans. Biol Sex Differ 2021;12:20.

[96] Dasappa H, Fathima FN, Prabhakar R, et al. Prevalence of diabetes and pre-diabetes and assessments of their risk factors in urban slums of Bangalore. J Family Med Prim Care. 2015; 4: 399-404.

[97] Hilliard ME, Yi-Frazier JP, Hessler D, et al. Stress and A1c among people with diabetes across the lifespan. Curr Diab Rep. 2016; 16: 67.

[98] Peric S, Stulnig TM. Diabetes and COVID-19. Disease—Management—People. Wien Klin Wochenschr. 2020; May 20: 1–6.

[99] Tanji Y, Sawada S, Watanabe T, et al. Impact of COVID-19 pandemic on glycemic control among outpatients with type 2 diabetes in Japan: A hospital-based survey from a country without lockdown. Diabetes research and clinical practice 2021; 176: 108840.

[100] Faulenbach M. Uthoff H, Schwegler K, et al. Effect of psychological stress on glucose control in patients with Type 2 diabetes. Diabetes Med 2012; 29: 128-31.

[101] Vasanth R, Ganesh A, Shanker R. Impact of stress on type 2 diabetes mellitus management. Psychiatr Danub. 2017;27 (Suppl 3):416-421.

[102] Singh AK, Gupta R, Ghosh A, Misra A. Diabetes in COVID-19: Prevalence, pathophysiology, prognosis and practical considerations. Diabetes Metab Syndr. 2020; 14: 303-310.

[103] Wong H, Singh J, Go RM, et al. The effects of mental stress on non-insulin-dependent diabetes: Determining the relationship between catecholamine and adrenergic signals from stress, anxiety, and depression on the physiological changes in the pancreatic hormone secretion. Cureus. 2019; 11: e5474.

[104] Salari N, Hosseinian-Far A, Jalali R, et al. Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: A systematic review and meta-analysis. Global Health. 2020; 16: 57.

[105] Lim S, Bae JH, Kwon H-S, et al. COVID-19 and diabetes mellitus: from pathophysiology to clinical management. Nat Rev Endocrinol. 2021;17:11-30.

[106] Erener S. Diabetes, infection risk and COVID-19. Mol Metab. 2020; 39: 101044.

[107] Pima Indians Diabetes dataset. Available from: http://archive.ics.uci.edu/ml/machine learning-databases/pima-indians-diabetes/pima-indians-diabetes data. Accessed: 1st May 2008.

[108] Yusoff MSB. Psychometric properties of the depression anxiety stress scale in a sample of medical degree applicants. Int Med J. 2013; 20: 295-300

[109] Patil R, Gothankar J. Risk factors for type 2 diabetes mellitus: An urban perspective. Indian J Med Sci 2019; 71: 16-21.

[110] Chatrati SP, Hossain G, Goyal A, et al. Smart home health monitoring system for predicting type 2 diabetes and hypertension. J. King Saud Univ. Comput. Inf. Sci., Jan. 2020.

LIST OF PUBLICATIONS AND PATENT

LIST OF PUBLICATIONS

Journal

1. Rohini Patil ,Dr.Kamal Shah ,"Assessment of Risk of Type 2 Diabetes Mellitus With Stress as a Risk Factor Using Classification Algorithms" published in International Journal of Recent Technology and Engineering(IJRTE) Vol-8,Issue-4 November 2019.
2. Rohini Patil, Dr Kamal Shah, Dr,Deepak Bhosle,"Impact of COVID-19 related Stress on Glycaemic control in Patients with Type-2 Diabetes Mellitus" accepted in Journal of Health Science and Surveillance System(Shiraz University of Medical Science,Iran) [Scopus Indexed]
3. Rohini Patil, Dr Kamal Shah," Impact of Lifestyle Parameters for Type 2 Diabetes Prediction: A Machine Learning paradigm "published in IJERT journal 2022

Conferences

1. Rohini Patil ,Dr.Kamal Shah ,"Performance Evaluation of different stress factors for Type-2 Diabetes Mellitus using Logistic Regression Model", published in the International conference on Intelligent Computing and Networking (IC-ICN-2020)Feb-2020
2. Rohini Patil, Dr Kamal Shah," Performance Evaluation of Machine Learning Classifiers for Prediction of Type 2 Diabetes using Stress Related Parameters" presented in International Conference on Data Science, Computation and Security (IDSCS 2022) [Springer, LNNS indexed in ISI Proceedings, Scopus Indexed]

Book chapter

1. Rohini Patil ,Dr.Kamal Shah ," Machine learning in healthcare: Applications, current status and future prospectus", Handbook of Research on Machine Learning" published in Apple Academic Press, CRC Press, Taylor & Francis Group (2022) (ISBN- 9781774638682)

Patent

1. Rohini Patil, Dr. Kamal Shah, “TYPE 2 DIABETES PREDICTION USING IMPROVED ENSEMBLE LEARNING DIMENSIONALITY REDUCTION.” Indian patent application number 202221005490 dated March 4, 2022.

Copyright

1. Rohini Patil, Dr. Kamal Shah, “Mathematical model for Diabe_Lifestyle” Diary Number 8673/2022-CO/L.

ACKNOLWEDGEMENT

ACKNOWLEDGEMENT

I would like to take the opportunity to express my sincere thanks to my guide **Dr. Kamal Shah,** Vice-Principal, Dean's(R&D)Thakur College of Engineering and Technology, Mumbai for her invaluable and constant support and regular guidance throughout my research work. Without her expert guidance and support, this would not have been possible. I am grateful to her for continuous feedback, which helped me effectively track and schedule my research work. Time, ideas, and encouragement given by her were really appreciable. I would also like to thank **Dr. B.K. Mishra**, Principal, Thakur College of Engineering and Technology, Mumbai for encouragement and providing the necessary and quick support throughout the research period. I would like to take the opportunity to express my sincere thanks to **Dr. R. R. Sedamkar,** Head, Research Centre, HOD-PhD programs, Thakur College of Engineering and Technology, Mumbai for his invaluable support and guidance.

I would also like to thank **Dr. Vaishali Khairnar**, Head of Department, IT, Terna Engineering College, Navi Mumbai and **Dr. Madhuri Rao**, Professor, Thadomal Shahani Engineering College, Mumbai who provided me valuable input with respect to research direction. I also wish to thank **Dr. Deepak Bhosle**, Professor and Head, Department of Pharmacology and Consultant Diabetologist, Aurangabad for his guidance in drafting the questionnaire and in collection of dataset for research. I express my gratitude to my husband, **Dr. Anant Patil** for his help and support in provide medical inputs. This thesis would not have been completed without the constant support of my parents, beloved daughter **Anushka** and son **Aayush**. I am very grateful to them for all their support over these years.

I thank management of my college for providing learning opportunity and support. I also thank my colleagues at Terna College of Engineering and Technology, Navi Mumbai for their constant encouragement and direct or indirect support.

Mrs. Rohini Anant Patil

APPENDIX-I

DATASET VALIDATION LETTER

DEOGIRI DIABETES CARE CENTRE

Dr. Deepak S. Bhosle
MBBS, MD, D. Diab.
Consultant Diabetologist
Regd.No. 87375

Dr. Vandana D. Bhosle
MBBS, DPB (Mumbai)
Consultant Pathologist
Regd.No. 2252

FACILITIES :- • DIABETES DIAGNOSIS & TREATMENT • ADVICE ABOUT DIET & EXERCISE • DIABETIC FOOT CLINIC • GESTATIONAL DIABETES MANAGEMENT. • INSULIN PUMP CLINIC • CGMS / AGP • PATHOLOGY LABORATORY • OBESITY MANAGEMENT

Date: 07/11/2020

TO WHOMSOEVER IT MAY CONCERN

Mrs. Rohini Anant Patil, pursuing her Ph.D. in information technology at Thakur College of Engineering and Technology, Kandivali, Mumbai is currently working on research project for "Estimation of risk of type 2 diabetes by developing a predictive model".

For the above said research project , she has designed a questionnaire in consultation with her guide and a healthcare professional (Physicians, Diabetologist and Endocrinologist) to collect relevant information from general population and people living with diabetes.

The questionnaire used for this project contains relevant questions covering different risk factors for the development of type 2 diabetes including demographic parameters, diet related factors, family history, lifestyle related factors and stress related factors.

Yours Sincerely

Dr. Deepak S. Bhosle
Deogiri Diabetes Care Centre
Siddi Arcade, Kalad Corner,
Besides HDFC Bank
Aurangabad, Maharashtra

Address : Siddhi Arcade, Office No. 101, 1st Floor, Kalda Corner, Besides HDFC Bank, Shriram Chowk, Aurangabad.

APPENDIX-II

ETHICS COMMITEE APPROVAL

Mahatma Gandhi Mission's
Ethics Committee for Research on Human Subjects
MGM campus, N-6, CIDCO, Aurangabad - 431003
Ph. No.: 0240-6601100, 6601174, Fax No.: +91-0240-2487727
Email: mgmecrhs@gmail.com
Office: Dept of Pharmacology & Member Secretary MGM-ECRHS,
MGM Medical College, Aurangabad

MGM-ECRHS/2020/

Date: 21st Dec, 2020

MGM - ECRHS APPROVAL LETTER

To,
The PI: Mrs. Rohini Patil
Research Scholar ,
Thakur College of Engineering & Technology ,
Mumbai ,

Dear Dr,

The meeting of Ethics Committee for Research on Human Subjects (ECRHS) was held on 21st Dec., 2020,at 10 AM. in the Boardroom with Dr. Manvendra Kachole as Chairperson. At the said meeting, the ECRHS approved the proposal for the clinical study entitled as **"Impact of COVID-19 related stress on glycaemic control in patients with type 2 diabetes mellitus: A Retrospective Study"** The ECRHS **reviewed and approved** the following documents submitted for the above mentioned clinical study.

List of Documents:

1. Study Protocol
2. Informed consent form
3. Case Record form
4. Appendices/Questionnaire
5. Adverse event reporting form / Product monograph
6. Current CV of Principal Investigator
7. Principal Investigator's undertaking

It is policy of ECRHS that, it should be informed about any serious adverse event occurring during the course of the study within seven days of the occurrence of the adverse event. The ECRHS expects to be informed about any changes in the protocol and informed Consent Document and asks to be provided a copy of the final report. The status of the study (completed/ongoing/terminated) should be reported to the ECRHS annually. 12 Members attended the meeting held on 21st Dec., 2020 fulfilling the quorum. It is hereby confirmed that neither you nor any of the study team members have participated in the voting/decision making procedures of the committee.

21/12/2020

Dr. Jyoti Bobde
Member Secretary
MGM-ECRHS
Department of Pharmacology,

MGM-ECRHS AURANGABAD

Member Secretary
MGM-ECRHS
MGM's Medical College,Aurangabad

APPENDIX-III

DATASET SCREENSHOT

APPENDIX-III A **LS_DIABETES DATASET**

1	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
2	sweets	How frequently d	Do you eat refined sugar?	Milk consumption (Milk	milk quantity	Are you Pregnant?	Any History of diabetes in family?	Do you suffer from anxiety?	What causes you stress in your work profile? Rank	poor family income	Pressure of working	Frequent travel	Monotonous work	poor and unplanned work	lack of career developme	powerlessness	financi securit
3	Yes	Once a week	Yes	Yes	200-400 ml/day	NA	Yes	Yes	Average	Average	High	Low	Average	Low	Low	Low	Low
4	Yes	Once a week	Yes	Yes	100-200 ml /day	No	Yes	No	High	Average	High	Average	High	Average	Average	Low	Low
5	Yes	Rarely	Yes	Yes	100-200 ml /day	NA	Yes	Yes	Very low	Very low	Very low	Very low	Very low	Very low	Very low	Very low	Very lo
6	Yes	Once a week	Yes	Yes	100-200 ml /day	No	Yes	No	Low	Very High	Average	Very low	High	Average	Average	Very low	Averag
7	Yes	Once a week	Yes	Yes	100-200 ml /day	No	Yes	No	High	Average	Low	Very low	Very low	Very low	Very High	Average	Averag
8	Yes	Once a week	Yes	Yes	200-400 ml/day	No	Yes	No	Average	Average	Very High	High	Average	High	High	Average	Very Hi
9	Yes	Once in 15-30 days	Yes	Yes	100-200 ml /day	NA	Yes	Yes	Low	Very High	Low	Low	High	Very High	High	High	Very Hi
10	Yes	Once in 15-30 days	No	Yes	200-400 ml/day	NA	No	Yes	High	High	Average	Average	Average	Very High	High	Low	Averag
11	Yes	Rarely	Yes	Yes	100-200 ml /day	NA	No	No	High	High	High	Very High	High	High	High	High	High
12	Yes	Once in 15-30 days	No	Yes	200-400 ml/day	No	Yes	No	Average	Average	Average	Average	Average	Average	Average	Average	Averag
13	Yes	Once a day	Yes	Yes	100-200 ml /day	No	Yes	No	Average	Average	Very High	Average	Average	Average	Average	Average	Averag
14	Yes	Once a day	Yes	Yes	200-400 ml/day	NA	No	No	Very High	Average	Average	Average	Average	Average	Average	Average	High
15	Yes	Once in 15-30 days	Yes	Yes	100-200 ml /day	No	Yes	No	Average	High	Average	Average	High	High	High	Average	Averag
16	Yes	Once a day	Yes	Yes	200-400 ml/day	No	Yes	No	Average	Very low	Very low	Average	Low	Low	Low	Low	Very lo
17	Yes	Once in 15-30 days	No	Yes	100-200 ml /day	NA	Yes	No	Average	Very low	Low	Very low	Low	High	Average	Very low	Very lo
18	Yes	Once a week	Yes	Yes	200-400 ml/day	NA	Yes	No	Average	Very low	Low	High	High	Very low	Very low	Very low	Very lo
19	Yes	Once a week	Yes	Yes	200-400 ml/day	NA	Yes	Yes	High	Average	High	Average	Average	High	Low	Average	Low

2	sr.no	Gender	Age (yea	Weight (kg)	Height (cm)	Profession	smoke	Exercise	cereal	salad	vegetables	grains
3	1	Male	42	61	165	Office Executive (sitting job profile)	No	30 mins- 1 hour	Once in a Day	1 Katori with lunch a	Twice in a day	Polished Rice;Wheat
4	2	Female	30	49	165	Other	No	I don't do any exerc	Twice in a day	Rarely (once in a we	Twice in a day	Polished Rice;Wheat;M
5	3	Male	52	60	159	Involves heavy exercise	Yes	30 mins- 1 hour	Twice in a day	Rarely (once in a we	Twice in a day	Polished Rice
6	4	Male	46	61	172	Other	No	Less than 30 mins	Three times in a da	Once in 2-3 days	Twice in a day	Wheat
7	5	Female	45	65	155	Involves heavy exercise	No	I don't do any exerc	Twice in a day	1/2 Katori for lunch	Twice in a day	Brown Rice
8	6	Female	26	56	123	other	No	Less than 30 mins	Twice in a day	1 Katori with lunch	Twice in a day	Polished Rice;Wheat;Jo
9	7	Male	51	69	167	Office Executive (sitting job profile)	No	Less than 30 mins	Three times in a da	1 Katori with lunch	Twice in a day	Jowar/Bajra/Raggi/Othe
10	8	Male	22	72	170	other	No	I don't do any exerc	Twice in a day	Once in 2-3 days	Twice in a day	Polished Rice;Wheat
11	9	Female	22	51	163	other	No	Dance	Once in a Day	Rarely (once in a we	Once in a Day	Wheat
12	10	Male	32	100	193	Office Executive (sitting job profile)	Yes	I don't do any exerc	Once in a Day	Once in 2-3 days	Once in a Day	Polished Rice
13	11	Female	43	52	163	Housewife	No	I don't do any exerc	Once in a Day	Rarely (once in a we	Twice in a day	Polished Rice;Wheat;Jo
14	12	Male	53	65	168	other	No	30 mins- 1 hour	Once in a Day	1 Katori with lunch	Twice in a day	Polished Rice;Wheat;Jo
15	13	Female	32	62	160	Office Executive (sitting job profile)	No	Less than 30 mins	Twice in a day	Once in 2-3 days	Once in a Day	Wheat
16	14	Female	35	63	161	Office Executive (sitting job profile)	No	I don't do any exerc	Once in a Day	Rarely (once in a we	Twice in a day	Polished Rice;Wheat;Jo
17	15	Male	39	64	167	Office Executive (sitting job profile)	No	I don't do any exerc	None	Rarely (once in a we	Twice in a day	Wheat
18	16	Female	27	75	157	other	No	30 mins- 1 hour	Once in a Day	None	Once in a Day	Polished Rice;Wheat
19	17	Male	48	58	162	Office Executive (sitting job profile)	No	I don't do any exerc	Once in a Day	1 Katori with lunch a	Twice in a day	Polished Rice;Wheat
20	18	Female	46	75	155	Office Executive (sitting job profile)	No	Less than 30 mins	Once in a Day	1 Katori with lunch a	Twice in a day	Polished Rice;Wheat
21	19	Female	70	54	142	Housewife	No	30 mins- 1 hour	Once in a Day	1 Katori with lunch a	Twice in a day	Polished Rice;Wheat

1	28	29	30	31	32	33	34	35	36
2	financial security	Unable to satisfy all stakeholders	High BP	Blood Press	Blood Pressu	Do you have rais	Range of Raised blo	Range of Raised blood	Value of HbA1c (glycated haemoglobin)?
3	Low	Low	No	60-80	90-120	No	Less than 100	less than 140	Below 5.7%
4	Low	Low	No	60-80	90-120	No	100-125	140-200	5.7 to 6.4%
5	Very low	Very low	No	80-90	90-120	Yes	Above 126	greater than 200	6.5% or higher
6	Average	Average	No	80-90	90-120	No	Less than 100	less than 140	Below 5.7%
7	Average	Low	No	80-90	90-120	No	Less than 100	less than 140	Below 5.7%
8	Very High	Very low	No	80-90	90-120	No	Less than 100	less than 140	Below 5.7%
9	Very High	Low	No	80-90	90-120	No	100-125	less than 140	5.7 to 6.4%
10	Average	Low	No	80-90	90-120	No	Less than 100	less than 140	5.7 to 6.4%
11	High	High	No	80-90	90-120	No	Less than 100	less than 140	Below 5.7%
12	Average	Average	No	80-90	90-120	No	100-125	less than 140	Below 5.7%
13	Average	Average	No	80-90	90-120	No	Less than 100	less than 140	Below 5.7%
14	High	Average	No	60-80	90-120	No	Less than 100	less than 140	Below 5.7%
15	Average	Average	No	80-90	90-120	No	Less than 100	less than 140	Below 5.7%
16	Very low	Very low	No	80-90	90-120	No	100-125	less than 140	Below 5.7%
17	Very low	Average	No	80-90	90-120	No	Less than 100	less than 140	Below 5.7%
18	Very low	Very low	No	60-80	90-120	No	Less than 100	less than 140	Below 5.7%
19	Low	High	Yes	90-100	120-140	Yes	Above 126	greater than 200	6.5% or higher

Survey on Healthcare | Final data | Stress Data | Food data | Apply model(non diabetes) | 244_ ...

APPENDIX-III B **COVI_STRESS DATASET**

2	Age	Ger	Weigh	Height	Height	BMI (if av	Residence (Rur	Family stat	Profession (S	Type 2 di	On trea	Current H	Current FE	Curren	Did CO	HbA1c %	FBG mg	PPG mg/dl	Did COVID-1	Was there m
3	55	MALE	75	164	1.64	27.9	URBAN	Married	BUISENESS	6	6	8.60	156.6	263.3	Yes	8	145.3	203.3	Yes	yes....Diet re
4	50	MALE	65	175	1.75	21.2	RURAL	Married	ELECTRICIAN	2	2	14	168	238	Yes	8.6	138	210	Yes	yes....Diet re
5	57	MALE	82	166	1.66	29.8	URBAN	Married	CLERK	5	5	12.7	196	421	Yes	14	161	386	Yes	yes....Diet re
6	43	FEMALE	96	150	1.5	42.7	URBAN	Married	FARMER	4	4	12.9	177	372	Yes	12.7	145	337	Yes	yes....Diet re
7	31	FEMALE	68	155	1.55	28.3	URBAN	Married	HOUSE WIFE	5	5	13	188	300	Yes	12.9	158	270	Yes	yes....Diet re
8	42	MALE	82	178	1.78	25.9	URBAN	Married	FARMER	4	4	14.1	150	324	Yes	13	220	292	Yes	yes....Diet re
9	45	MALE	93	180	1.8	28.7	URBAN	Married	ELECTRICIAN	3	3	11.7	133	421	Yes	14.1	194	390	Yes	yes....Diet re
10	50	FEMALE	90	160	1.6	35.2	URBAN	Married	HOUSE WIFE	2	2	12.6	182	472	Yes	11.7	198	437	Yes	yes....Diet re
11	47	MALE	80	172	1.72	27.0	URBAN	Married	GOVT.SERVANT	5	5	10.6	160	328	Yes	12.6	143	293	Yes	yes....Diet re
12	57	FEMALE	89	142	1.42	44.1	RURAL	Married	HOUSE WIFE	6	6	8.3	131	281	Yes	10.6	128	248	Yes	yes....Diet re
13	35	FEMALE	74	147	1.47	34.2	URBAN	Married	HOUSE WIFE	4	4	11.1	189	403	Yes	8.3	197	368	Yes	yes....Diet re
14	44	MALE	86	165	1.65	31.6	RURAL	Married	RETIRED	3	3	12	160	381	Yes	11.1	154	346	Yes	yes....Diet re
15	48	MALE	69	175	1.75	22.5	RURAL	Married	GOVT.SERVANT	2	2	12.2	231	408	Yes	12	190	373	Yes	yes....Diet re
16	53	MALE	91	170	1.7	31.5	RURAL	Married	FARMER	3	3	14.7	195	386	Yes	12.2	160	346	Yes	yes....Diet re
17	37	MALE	80	162	1.62	30.5	RURAL	Married	GOVT.SERVANT	3	3	13.6	156	291	Yes	14.7	126	251	Yes	yes....Diet re
18	30	MALE	71	163	1.63	26.7	RURAL	Married	FARMER	2	2	9.9	172	289	Yes	13.6	156	264	Yes	yes....Diet re
19	42	FEMALE	86	155	1.55	35.8	RURAL	Married	HOUSE WIFE	5	5	10.4	188	388	Yes	9.9	150	263	Yes	yes....Diet re
20	60	MALE	93	168	1.68	33.0	RURAL	Married	FARMER	7	7	13.8	230	387	Yes	10.4	200	342	Yes	yes....Diet re
21	41	FEMALE	112	153	1.53	47.8	RURAL	Married	HOUSE WIFE	6	6	14.2	242	411	Yes	13.8	207	359	Yes	yes....Diet re
22	54	FEMALE	81	150	1.5	36.0	RURAL	Married	HOUSE WIFE	5	5	10.9	169	363	Yes	14.2	137	313	Yes	yes....Diet re
23	68	MALE	78	169	1.69	27.3	RURAL	Married	RETIRED	4	4	13.9	190	329	Yes	10.9	159	294	Yes	yes....Diet re
24	47	FEMALE	65	147	1.47	30.1	RURAL	Married	HOUSE WIFE	3	3	11.4	182	251	Yes	13.9	174	216	Yes	yes....Diet re
25	61	FEMALE	84	156	1.56	34.5	URBAN	Married	HOUSE WIFE	2	2	10.7	199	368	Yes	11.4	169	333	Yes	yes....Diet re
26	70	FEMALE	72	150	1.5	32.0	URBAN	Married	HOUSE WIFE	9	9	9.8	240	402	Yes	10.7	210	367	Yes	yes....Diet re
27	36	FEMALE	95	152	1.52	41.1	RURAL	Married	HOUSE WIFE	2	2	8.6	182	342	Yes	9.8	152	302	Yes	yes....Diet re
28	68	FEMALE	74	147	1.47	34.2	URBAN	Married	HOUSE WIFE	2	2	12.1	211	267	Yes	8.6	176	234	Yes	yes....Diet re
29	44	FEMALE	56	144	1.44	27.0	RURAL	Married	HOUSE WIFE	3	3	10.5	196	324	Yes	12.1	161	288	Yes	yes....Diet re
30	53	FEMALE	96	140	1.4	49.0	RURAL	Married	HOUSE WIFE	6	6	9.4	172	276	Yes	10.5	137	241	Yes	yes....Diet re
31	52	FEMALE	74	150	1.5	32.9	RURAL	Married	FARMER	5	5	10.4	242	414	Yes	9.4	207	369	Yes	yes....Diet re

2	On trea	Current H	Current FE	Curren	Did CO	HbA1c %	FBG mg	PPG mg/dl	Did COVID-1	Was there major ch	Was phy	Was tre	Were 1.Qua	I found it hard to wind down		I tended to over-react to situations		I found myself getting agitated		I was intol of anything kept me l getting or
3	6	8.60	156.6	263.3	Yes	8	145.3	203.3	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	sometimes	1	often	2	always
4	2	14	168	238	Yes	8.6	138	210	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	sometimes	1	often	2	always
5	5	12.7	196	421	Yes	14	161	386	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	sometimes	1	often	2	always
6	4	12.9	177	372	Yes	12.7	145	337	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
7	5	13	188	300	Yes	12.9	158	270	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
8	4	14.1	150	324	Yes	13	220	292	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	sometimes	1	often	2	always
9	3	11.7	133	421	Yes	14.1	194	390	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	sometimes	1	often	2	always
10	2	12.6	182	472	Yes	11.7	198	437	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
11	5	10.6	160	328	Yes	12.6	143	293	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	sometimes	1	often	2	often
12	6	8.3	131	281	Yes	10.6	128	248	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
13	4	11.1	189	403	Yes	8.3	197	368	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
14	3	12	160	381	Yes	11.1	154	346	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	sometimes	1	often	2	often
15	2	12.2	231	408	Yes	12	190	373	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	sometimes	1	often	2	often
16	3	14.7	195	386	Yes	12.2	160	346	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	sometimes	1	often	2	often
17	3	13.6	156	291	Yes	14.7	126	251	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	sometimes	1	often	2	often
18	2	9.9	172	289	Yes	13.6	156	264	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	sometimes	1	often	2	often
19	5	10.4	188	388	Yes	9.9	150	263	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
20	7	13.8	230	387	Yes	10.4	200	342	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	sometimes	1	often	2	often
21	6	14.2	242	411	Yes	13.8	207	359	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
22	5	10.9	169	363	Yes	14.2	137	313	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
23	4	13.9	190	329	Yes	10.9	159	294	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	sometimes	1	often	2	often
24	3	11.4	182	251	Yes	13.9	174	216	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
25	2	10.7	199	368	Yes	11.4	169	333	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
26	9	9.8	240	402	Yes	10.7	210	367	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
27	2	8.6	182	342	Yes	9.8	152	302	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
28	2	12.1	211	267	Yes	8.6	176	234	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
29	3	10.5	196	324	Yes	12.1	161	288	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
30	6	9.4	172	276	Yes	10.5	137	241	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	never
31	5	10.4	242	414	Yes	9.4	207	369	Yes	yes....Diet reduced	yes	yes	Hospitalized	Never	0	Never	0	sometimes	1	sometimes

APPENDIX-III C **PIMA INDIAN DIABETES DATASET**

1	Pregnanci	Glucose	BloodPres	SkinThick	Insulin	BMI	DiabetesP	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1

APPENDIX-III D **DIABETES_2019 DATASET**

1	Age	Gender	Family_Di	highBP	Physically	BMI	Smoking	Alcohol	Sleep	SoundSle	RegularM	JunkFood	Stress	BPLevel	Pregancie	Pdiabetes	UriationFr	Diabetic
2	50-59	Male	no	yes	one hr or	39	no	no	8	6	no	occasiona	sometime	high	0	0	not much	no
3	50-59	Male	no	yes	less than	28	no	no	8	6	yes	very ofter	sometime	normal	0	0	not much	no
4	40-49	Male	no	no	one hr or	24	no	no	6	6	no	occasiona	sometime	normal	0	0	not much	no
5	50-59	Male	no	no	one hr or	23	no	no	8	6	no	occasiona	sometime	normal	0	0	not much	no
6	40-49	Male	no	no	less than	27	no	no	8	8	no	occasiona	sometime	normal	0	0	not much	no
7	40-49	Male	no	yes	none	21	no	yes	10	10	no	occasiona	sometime	high	0	0	not much	yes
8	less than 4	Male	no	no	one hr or	24	no	no	8	8	no	occasiona	sometime	normal	0	0	not much	no
9	less than 4	Male	no	no	less than	20	no	no	7	7	yes	occasiona	sometime	low	0	0	not much	no
10	40-49	Male	yes	no	one hr or	23	no	no	7	7	no	occasiona	sometime	normal	0	0	not much	no
11	less than 4	Male	no	no	more thar	20	no	no	8	8	o	occasiona	sometime	normal	0	0	not much	no
12	less than 4	Male	no	no	none	20	no	no	7	7	no	occasiona	not at all	normal	0	0	not much	no
13	40-49	Male	no	no	less than	26	yes	no	8	7	no	occasiona	sometime	normal	0	0	not much	no
14	less than 4	Female	no	no	less than	21	no	no	6	6	no	occasiona	sometime	normal	1	0	not much	no
15	less than 4	Female	no	no	one hr or	22	no	no	8	7	no	occasiona	very ofter	normal	2	0	not much	no
16	less than 4	Male	no	no	one hr or	15	no	no	7	7	no	occasiona	very ofter	normal	0	0	not much	no
17	40-49	Female	yes	no	none	34	no	no	6	6	no	occasiona	sometime	normal	3	0	not much	no
18	50-59	Male	yes	no	more thar	22	no	no	7	6	no	occasiona	sometime	normal	0	0	not much	no
19	less than 4	Female	no	no	more thar	24	no	no	6	6	no	occasiona	sometime	normal	2	0	quite ofte	no
20	40-49	Male	yes	no	one hr or	24	no	no	11	11	no	occasiona	sometime	normal	0	0	not much	no
21	less than 4	Male	yes	no	one hr or	22	no	no	8	6	no	often	not at all	normal	0	0	not much	no

APPENDIX-IV

PLAGIARISM REPORT

PLAGIARISM REPORT

turnitin

Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Rohini Patil
Assignment title:	D1
Submission title:	Thesis
File name:	Rohini_Thesis_300722.docx
File size:	1.24M
Page count:	129
Word count:	27,622
Character count:	156,054
Submission date:	31-Jul-2022 10:59AM (UTC+0530)
Submission ID:	1877076609

Copyright 2022 Turnitin. All rights reserved.

Thesis

ORIGINALITY REPORT

9%	5%	7%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Rohini Patil, Kamal Shah. "Chapter 8 Performance Evaluation of Machine Learning Classifiers for Prediction of Type 2 Diabetes Using Stress-Related Parameters", Springer	1%

APPENDIX-V

SYNOPSIS REPORT AND COMPLIANCES

Synopsis

On

DESIGN OF PREDICTIVE MODEL USING VARIOUS MACHINE LEARNING ALGORITHMS FOR IMPROVING DIAGNOSIS OF TYPE 2 DIABETES

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY (TECHNOLOGY)
In
INFORMATION TECHNOLOGY

By
Rohini Patil

(Registration No. 12/18-12-2019)

Under the supervision of

Prof. Kamal Shah

Thakur College of Engineering & Technology

UNIVERSITY OF MUMBAI

DEPARTMENT OF INFORMATION TECHNOLOGY

A - Block, Thakur Educational Campus, Shyamnarayan Thakur Marg,Thakur Village, Kandivali (East), Mumbai - 400 101
Tel.: 022-6730 8000 / 8106 / 8107 Telefax: 022-2846 1890 • Email: tcet@thakureducation.org

SYNOPSIS OF THE THESIS TO BE SUBMITTED TO THE UNIVERSITY OF MUMBAI FOR THE DEGREE OF DOCTORATE OF PHILOSOPHY IN THE SUBJECT OF TECHNOLGY, INFORMATION TECHNOLOGY

Name of Student	Rohini Anant Patil
Topic of Research	Design of Predictive Model Using Various Machine Learning Algorithms for Improving Diagnosis of Type 2 Diabetes.
Degree and Subject	Ph.D. in Technology (Information Technology)
Registration No./Date	12/18-12-2019
Name of Research Supervisor / Guide	Dr. Kamal Shah
Date of superannuation /retirement	Jan- 2035
Name of Research Centre	Thakur College of Engineering & Technology, A-Block, Thakur Educational Campus Shyamnarayan Thakur Marg, Thakur Village, Kandivali (East), Mumbai – 400101, Maharashtra
Signature of student	
Signature of Research Guide/ Supervisor	
Date of submission of synopsis	5/5/22

Dr. B.K. Mishra

Principal

Seal of the College

PRINCIPAL
Thakur College of Engg. & Tech.
A-Block, Thakur Educational Campus
Shyamnarayan Marg, Thakur Village
Kandivali (E), Mumbai-400 101

THAKUR COLLEGE OF ENGINEERING & TECHNOLOGY

Certificate

This is to certify that the work embedded in this synopsis report on the research title "**Design of Predictive Model Using Various Machine Learning Algorithms for Improving Diagnosis of Type 2 Diabetes**" is a bonafide work of Ms **Rohini A.Patil** (Ph.D. Registration No. 12/18-12-2019) for partial fulfilment of the requirement for the degree of Doctor of Philosophy in Technology in the subject of Information Technology under the guidance of **Dr. Kamal Shah**, at Thakur College of Engineering and Technology Research Centre of University of Mumbai.

Dr. Kamal Shah
Guide

Dr.R.R.Sedamkar,
Head, Research Centre,
HOD, Ph.D Program

Dr. B.K. Mishra
Principal

PRINCIPAL
Thakur College of Engineering & Technology

Date 5/5/22
Place: Mumbai

Declaration by the Candidate

I wish to state that the work embodied in the synopsis titled "**Design of Predictive Model Using Various Machine Learning Algorithms for Improving Diagnosis of Type 2 Diabetes**" forms my contribution to the work carried out under the guidance of Dr. Kamal Shah at the Thakur College of Engineering and Technology. I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented, fabricated or falsified any idea/data/fact/source in my submission.

Ms. Rohini Anant Patil,

Ph.D. Registration No. 12/18-12-2019

ABSTRACT

Diabetic mellitus (DM) is one of the most common non-communicable diseases globally. Due to increasing prevalence, the burden of diabetes is growing day by day. Preventive care in early stage is essential to avoid complications. Early diagnosis and better care with reasonable cost are the key aspects for improving patient satisfaction. Enormous data and increased complexities have led to rising interest in the use of machine learning (ML) in healthcare. ML uses several approaches to automatically learn and improvise the prediction without explicit programming. It develops on existing statistical methods and finds patterns in the data. Predictive modelling is capable of making predictions based on certain properties from a training dataset.

In order to overcome the projected burden of diabetes, a risk predictive model can be valuable for public health. The objective is to design a predictive model for estimation of diabetes in healthy people with diverse age groups based on different life style related factors i.e. stress, food habit, smoking, profession and exercise with use of series of pre-processing procedures on various ML algorithms such as Gradient boosting classifier (GBC), Logistic regression (LR), Random forest (RF), k-nearest neighbor (KNN), Support Vector Machine (SVM), Decision tree (DT) and etc.

This research contributions consists of 4 phases. Phase one work contributed to analyze the effect of stress for prediction of type 2 diabetes. In order to achieve this, a dataset of 374 records was collected from web-based questionnaire and validated by a consulting doctor. The proposed methodology was divided into 2 phases. In the first phase, we applied feature selection techniques namely filter method-Select-K best method, Feature importance technique, Information gain, Correlation technique and Hybrid method as recursive feature elimination (RFE) and with the help of these 6 different models, K fold cross validation accuracy was calculated. In the second phase, hyper parameter tuning using grid search method was used and accuracy of the model was estimated. Comparative assessment of accuracy, precision, recall, F1-measure was performed among 6 machine learning models. GBC and SVM showed good accuracy, precision and recall value.

In phase 2, the impact of COVID-19 related stress on blood glucose level in type-2 diabetes patients was examined using statistical methods. A dataset of 199 diabetes patients was collected from MGM Hospital, Aurangabad. DASS-12 stress sub-scale was used to estimate severity of stress. Fasting blood glucose (FBG) and post-prandial blood glucose

(PPBG) before admission and at the time of discharge from COVID-19 treatment was compared. There was rise in glucose level in type 2 diabetes patients discharged after COVID-19 treatment.

In the phase 3, using 6 machine learning models, the effect of lifestyle parameter for the prediction of type 2 diabetes was analyzed. The proposed model was implemented by a combination of 5 feature selection methods and 6 different classification methods. The experimental results were carried out using lifestyle based parameters. The highest feature selection accuracy was observed in the GBC with RFE method. Based on the comparative analysis, RF and GBC showed highest Receiver Operating Characteristics (ROC) value. The used framework showed good result in comparison with the existing methods.

In the phase 4, a novel- Improved Ensemble Learning with Dimensionality Reduction Model (IELDR) was designed. Auto encoder based feature extraction method along with ensemble learning algorithm was used. The experiments were carried out using LS_diabetes dataset. The IELDR model showed better results in terms accuracy, sensitivity, ROC and Mathew correlation when compared with standard PIMA and Diabetes_2019 dataset. This proposed IELDR system predicts risk for developing of type 2 diabetes in healthy person based upon the person's current lifestyle pattern, the system calculates the risk score and offers recommendation.

Program: Ph.D (Technology)

Annual Progress Seminar (APS)/
Six Monthly Research Progress Seminar (RPS)/
Pre-Synopsis Presentation Seminar (RAC) Reports

Evaluation Sheet

Date of conduct of APS/RPS/RAC: 29-04-2022 to 29-04-2022
Mode of Examination: Offline in College/Online Through Digital Platform
Time of APS/RPS/RAC: Second Half: 3:00pm to 3:30pm

Sr. No.	Particulars/Details	Observations of Expert/Examiner/Supervisor (Sr. No 1-10 to be filled by Research Scholar)
1	**Name of Research Scholar**:	Rohini Patil
2	**Department**	Information Technology
3	**Date of Admission with Type of Report:**	19/03/2018
4	**Registration / UID NO with date:**	12 /18-12-2019
5	**Name of the Supervisor:**	Dr.Kamal Shah
6	**Title of Research:**	Design of Predictive Model Using Various Machine Learning Algorithms for Improving Diagnosis of Type 2 Diabetes.
7	**Type and period of Report**	Pre-synopsis ,April-2022
8	**Objectives of Research (3 to 4) and Expected outcomes**	**RO1:** To analyze the effect of **stress** related parameters on prediction of type 2 diabetes. **RO2:** To analyze the impact of COVID-19 related stress on blood glucose in patients with diabetes of 3 different age groups.

		RO3: To analyze the effect of lifestyle factors on prediction of type 2 diabetes. **RO4:** Design a novel predictive algorithm which can predict risk of type-2 diabetes mellitus up to 80% accuracy and provide recommendation based on prediction.
9	**A brief write-up of progress with highlights of Research contributions and Claims in 8-10 bullet points or Maximum of 200 words:**	• Developed "LS_diabetes" dataset of 375 instances • Developed machine learning model (using Logistic regression, Random forest, Decision tree, Support vector machine, K-nearest neighbor and gradient boosting classifier along with 5 different feature selection methods and optimization technique for evaluating stress as a risk factor for the development of type 2 diabetes. • Performance comparison of 6 model was done on various measures. • Collected Hospitalized "COVI-Stress" dataset of 199 instances. • Evaluated impact of COVID-19 related stress on blood glucose level using "COVI-Stress" dataset. • DASS-12 scale was used to evaluate the impact of stress on blood glucose level in patients with type 2 diabetes after discharged from hospital. • Significance difference found in fasting and postprandial glucose level in type-2 diabetes patients discharged after COVID-19 treatment. • Significant difference found in type-2 diabetes with moderate stress. • Overall significant difference found in age. Age wise impact of stress observed in people with old age. • Significant difference found in rural and urban as well as in gender. • Developed machine learning model for evaluating lifestyle as a risk factor for the development of type 2 diabetes using 6 model and feature selection methods. • Performance comparison of 6 model was done. • Compared result with other researchers and with available dataset. • Designed Novel framework for prediction of risk of type 2 diabetes. • Proposed IELDR model on LS diabetes dataset compared with PIMA dataset and Diabetes_2019 dataset. • Proposed IELDR model results validated with other researcher (ensemble algorithm)on both

		PIMA and diabetes_2019 dataset cases and shows good results. • The proposed system provides recommendation.
10	**Publication Details till date in IEEE format**	1] Rohini Patil ,Dr.Kamal Shah ,"Assessment of Risk of Type 2 Diabetes Mellitus With Stress as a Risk Factor Using Classification Algorithms" published in International Journal of Recent Technology and Engineering(IJRTE) Vol-8,Issue-4 November 2019. 2] Rohini Patil ,Dr.Kamal Shah ,"Performance Evaluation of different stress factors for Type-2 Diabetes Mellitus using Logistic Regression Model", published in the International conference on Intelligent Computing and Networking (IC-ICN-2020)Feb-2020 3]Rohini Patil ,Dr.Kamal Shah ," Machine learning in healthcare: Applications, current status and future prospectus", Handbook of Research on Machine Learning" published in Apple Academic Press, CRC Press, Taylor & Francis Group (2022) (ISBN-9781774638682) 4] Rohini Patil, Dr Kamal Shah, Dr,Deepak Bhosle,"Impact of COVID-19 related Stress on Glycaemic control in Patients with Type-2 Diabetes Mellitus" accepted in Journal of Health Science and Surveillance System(Shiraz University of Medical Science,Iran) [Scopus Indexed] 5] Rohini Patil, Dr Kamal Shah," Performance Evaluation of Machine Learning Classifiers for Prediction of Type 2 Diabetes using Stress Related Parameters" presented in International Conference on Data Science, Computation and Security (IDSCS 2022) [Springer, LNNS indexed in ISI Proceedings, Scopus Indexed] 6] Rohini Patil, Dr Kamal Shah," Impact of Lifestyle Parameters for Type 2 Diabetes Prediction: A Machine Learning paradigm "accepted in IJERT journal 2022 **Patent:** Rohini Patil, Dr. Kamal Shah, "Type 2 diabetes prediction using Improved Ensemble Learning Dimensionality Reduction." Indian patent application number 202221005490 dated March 4, 2022. **Copyright:**

		Rohini Patil, Dr. Kamal Shah, "Mathematical model for Diabe_Lifestyle" Diary Number 8673/2022-CO/L. (24th April 2022)
11	**Observations/Suggestions by Expert/ Supervisor during last RPS/RAC/APS**	RPS I: • To study about training dataset pattern i.e.70-30 or 80-20 (why) • Specify machine learning algorithm and design optimal algorithm using ML. • Proposed methodology needs to be redefined • Social factors not considered in detail • Questionnaire needs to be more in detail • A student has understanding of Research Problem and work in that direction • To find impact of COVID-19 related stress on blood glucose. RPS-II: • Are the dataset age wise will compared as sector/area wise? • What life style factors you are considering? • What threshold value considered for technical comparison of ML algorithm? • Which ML algorithm gives best results compared with others and why? • What are your parameters/factors which you will consider for Novel Algorithm? • Need to devise Novel algorithm to test on standard as well as non-standard dataset • Work on Novel algorithm • Find the effect of stress on COVID-19 diabetes dataset RPS-III: • Work more towards research. • Paper to be published in scopus journal. • List the KPI to check the model performance in Novel • Proposed Algorithm and suggest unique name for your algorithm which will highlight your unique contribution. RPS-IV: • Good Work done can go for pre-synopsis.

		• Analysis is also done. Technology and medical terms must go hand-in-hand while presentation. • If good paper publication is done on all RO, can go for pre-synopsis. • Enough work is done can be considered for pre-synopsis
12	**Observation/Suggestions & Recommendations of Expert/Examiner/Supervisor**	• Good Work done. • Include publications/patent in specific RO slides • Includes all objectives in one slide • PPT must focus more on technical. • Recommended for thesis submission
13	**Name & Digital Signature with date of Expert/Examiner/Supervisor:**	DR.Vaishali Kairnar

Note: On completion of Examination in online mode, the Digitally signed/ Signed scanned copy of report in PDF format should be sent over E-mail by examiner to HoD (Ph.D Programs).

Program: Ph.D (Technology)

Annual Progress Seminar (APS)/
Six Monthly Research Progress Seminar (RPS)/
Pre-Synopsis Presentation Seminar (RAC) Reports

Evaluation Sheet

Date of conduct of APS/RPS/RAC: 29-04-2022 to 29-04-2022
Mode of Examination: Offline in College/Online Through Digital Platform
Time of APS/RPS/RAC: Second Half: 3:00pm to 3:30pm

Sr. No.	Particulars/Details	Observations of Expert/Examiner/Supervisor (Sr. No 1-10 to be filled by Research Scholar)
1	**Name of Research Scholar**:	Rohini Patil
2	**Department**	Information Technology
3	**Date of Admission with Type of Report:**	19/03/2018
4	**Registration / UID NO with date:**	12 /18-12-2019
5	**Name of the Supervisor:**	Dr.Kamal Shah
6	**Title of Research:**	Design of Predictive Model Using Various Machine Learning Algorithms for Improving Diagnosis of Type 2 Diabetes.
7	**Type and period of Report**	Pre-synopsis ,April-2022
8	**Objectives of Research (3 to 4) and Expected outcomes**	**RO1:** To analyze the effect of **stress** related parameters on prediction of type 2 diabetes. **RO2:** To analyze the impact of COVID-19 related stress on blood glucose in patients with diabetes of 3 different age groups. **RO3:** To analyze the effect of lifestyle factors on prediction of type 2 diabetes. **RO4:** Design a novel predictive algorithm which can predict risk of type-2 diabetes mellitus up to 80%

		accuracy and provide recommendation based on prediction.
9	**A brief write-up of progress with highlights of Research contributions and Claims in 8-10 bullet points or Maximum of 200 words:**	• Developed "LS_diabetes" dataset of 375 instances • Developed machine learning model (using Logistic regression, Random forest, Decision tree, Support vector machine, K-nearest neighbor and gradient boosting classifier along with 5 different feature selection methods and optimization technique for evaluating stress as a risk factor for the development of type 2 diabetes. • Performance comparison of 6 model was done on various measures. • Collected Hospitalized "COVI-Stress" dataset of 199 instances. • Evaluated impact of COVID-19 related stress on blood glucose level using "COVI-Stress" dataset. • DASS-12 scale was used to evaluate the impact of stress on blood glucose level in patients with type 2 diabetes after discharged from hospital. • Significance difference found in fasting and postprandial glucose level in type-2 diabetes patients discharged after COVID-19 treatment. • Significant difference found in type-2 diabetes with moderate stress. • Overall significant difference found in age. Age wise impact of stress observed in people with old age. • Significant difference found in rural and urban as well as in gender. • Developed machine learning model for evaluating lifestyle as a risk factor for the development of type 2 diabetes using 6 model and feature selection methods. • Performance comparison of 6 model was done. • Compared result with other researchers and with available dataset. • Designed Novel framework for prediction of risk of type 2 diabetes. • Proposed IELDR model on LS diabetes dataset compared with PIMA dataset and Diabetes_2019 dataset. • Proposed IELDR model results validated with other researcher (ensemble algorithm)on both PIMA and diabetes_2019 dataset cases and shows good results. • The proposed system provides recommendation.

10	**Publication Details till date in IEEE format**	1] Rohini Patil ,Dr.Kamal Shah ,"Assessment of Risk of Type 2 Diabetes Mellitus With Stress as a Risk Factor Using Classification Algorithms" published in International Journal of Recent Technology and Engineering(IJRTE) Vol-8,Issue-4 November 2019. 2] Rohini Patil ,Dr.Kamal Shah ,"Performance Evaluation of different stress factors for Type-2 Diabetes Mellitus using Logistic Regression Model", published in the International conference on Intelligent Computing and Networking (IC-ICN-2020)Feb-2020 3]Rohini Patil ,Dr.Kamal Shah ," Machine learning in healthcare: Applications, current status and future prospectus", Handbook of Research on Machine Learning" published in Apple Academic Press, CRC Press, Taylor & Francis Group (2022) (ISBN-9781774638682) 4] Rohini Patil, Dr Kamal Shah, Dr,Deepak Bhosle,"Impact of COVID-19 related Stress on Glycaemic control in Patients with Type-2 Diabetes Mellitus" accepted in Journal of Health Science and Surveillance System(Shiraz University of Medical Science,Iran) [Scopus Indexed] 5] Rohini Patil, Dr Kamal Shah," Performance Evaluation of Machine Learning Classifiers for Prediction of Type 2 Diabetes using Stress Related Parameters" presented in International Conference on Data Science, Computation and Security (IDSCS 2022) [Springer, LNNS indexed in ISI Proceedings, Scopus Indexed] 6] Rohini Patil, Dr Kamal Shah," Impact of Lifestyle Parameters for Type 2 Diabetes Prediction: A Machine Learning paradigm "accepted in IJERT journal 2022 **Patent:** Rohini Patil, Dr. Kamal Shah, "Type 2 diabetes prediction using Improved Ensemble Learning Dimensionality Reduction." Indian patent application number 202221005490 dated March 4, 2022. **Copyright:** Rohini Patil, Dr. Kamal Shah, "Mathematical model for Diabe_Lifestyle" Diary Number 8673/2022-CO/L. (24th April 2022)

11	**Observations/Suggestions by Expert/ Supervisor during last RPS/RAC/APS**	RPS I: • To study about training dataset pattern i.e.70-30 or 80-20 (why) • Specify machine learning algorithm and design optimal algorithm using ML. • Proposed methodology needs to be redefined • Social factors not considered in detail • Questionnaire needs to be more in detail • A student has understanding of Research Problem and work in that direction • To find impact of COVID-19 related stress on blood glucose. RPS-II: • Are the dataset age wise will compared as sector/area wise? • What life style factors you are considering? • What threshold value considered for technical comparison of ML algorithm? • Which ML algorithm gives best results compared with others and why? • What are your parameters/factors which you will consider for Novel Algorithm? • Need to devise Novel algorithm to test on standard as well as non-standard dataset • Work on Novel algorithm • Find the effect of stress on COVID-19 diabetes dataset RPS-III: • Work more towards research. • Paper to be published in scopus journal. • List the KPI to check the model performance in Novel • Proposed Algorithm and suggest unique name for your algorithm which will highlight your unique contribution. RPS-IV: • Good Work done can go for pre-synopsis. • Analysis is also done. Technology and medical terms must go hand-in-hand while presentation. • If good paper publication is done on all RO, can go for pre-synopsis.

		• Enough work is done can be considered for pre-synopsis
12	**Observation/Suggestions & Recommendations of Expert/Examiner/Supervisor**	Presentation must focus more on technical aspects rather than diabetes. Recommended for thesis submission.
13	**Name & Digital Signature with date of Expert/Examiner/Supervisor:**	Dr.Madhuri Rao

Note: On completion of Examination in online mode, the Digitally signed/ Signed scanned copy of report in PDF format should be sent over E-mail by examiner to HoD (Ph.D Programs).

Program: Ph.D (Technology)

Annual Progress Seminar (APS)/
Six Monthly Research Progress Seminar (RPS)/
Pre-Synopsis Presentation Seminar (RAC) Reports

Evaluation Sheet

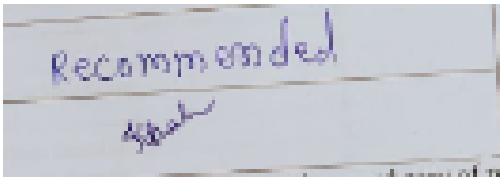
Date of conduct of APS/RPS/RAC: 29-04-2022 to 29-04-2022
Mode of Examination: Offline in College/Online Through Digital Platform
Time of APS/RPS/RAC: Second Half: 3:00pm to 3:30pm

Sr. No.	Particulars/Details	Observations of Expert/Examiner/Supervisor (Sr. No 1-10 to be filled by Research Scholar)
1	**Name of Research Scholar**:	Rohini Patil
2	**Department**	Information Technology
3	**Date of Admission with Type of Report:**	19/03/2018
4	**Registration / UID NO with date:**	12 /18-12-2019
5	**Name of the Supervisor:**	Dr.Kamal Shah
6	**Title of Research:**	Design of Predictive Model Using Various Machine Learning Algorithms for Improving Diagnosis of Type 2 Diabetes.
7	**Type and period of Report**	Pre-synopsis ,April-2022
8	**Objectives of Research (3 to 4) and Expected outcomes**	**RO1:** To analyze the effect of **stress** related parameters on prediction of type 2 diabetes. **RO2:** To analyze the impact of COVID-19 related stress on blood glucose in patients with diabetes of 3 different age groups. **RO3:** To analyze the effect of lifestyle factors on prediction of type 2 diabetes. **RO4:** Design a novel predictive algorithm which can predict risk of type-2 diabetes mellitus up to 80%

		accuracy and provide recommendation based on prediction.
9	**A brief write-up of progress with highlights of Research contributions and Claims in 8-10 bullet points or Maximum of 200 words:**	• Developed "LS_diabetes" dataset of 375 instances • Developed machine learning model (using Logistic regression, Random forest, Decision tree, Support vector machine, K-nearest neighbor and gradient boosting classifier along with 5 different feature selection methods and optimization technique for evaluating stress as a risk factor for the development of type 2 diabetes. • Performance comparison of 6 model was done on various measures. • Collected Hospitalized "COVI-Stress" dataset of 199 instances. • Evaluated impact of COVID-19 related stress on blood glucose level using "COVI-Stress" dataset. • DASS-12 scale was used to evaluate the impact of stress on blood glucose level in patients with type 2 diabetes after discharged from hospital. • Significance difference found in fasting and postprandial glucose level in type-2 diabetes patients discharged after COVID-19 treatment. • Significant difference found in type-2 diabetes with moderate stress. • Overall significant difference found in age. Age wise impact of stress observed in people with old age. • Significant difference found in rural and urban as well as in gender. • Developed machine learning model for evaluating lifestyle as a risk factor for the development of type 2 diabetes using 6 model and feature selection methods. • Performance comparison of 6 model was done. • Compared result with other researchers and with available dataset. • Designed Novel framework for prediction of risk of type 2 diabetes. • Proposed IELDR model on LS diabetes dataset compared with PIMA dataset and Diabetes_2019 dataset. • Proposed IELDR model results validated with other researcher (ensemble algorithm)on both PIMA and diabetes_2019 dataset cases and shows good results. • The proposed system provides recommendation.

10	**Publication Details till date in IEEE format**	1] Rohini Patil ,Dr.Kamal Shah ,"Assessment of Risk of Type 2 Diabetes Mellitus With Stress as a Risk Factor Using Classification Algorithms" published in International Journal of Recent Technology and Engineering(IJRTE) Vol-8,Issue-4 November 2019. 2] Rohini Patil ,Dr.Kamal Shah ,"Performance Evaluation of different stress factors for Type-2 Diabetes Mellitus using Logistic Regression Model", published in the International conference on Intelligent Computing and Networking (IC-ICN-2020)Feb-2020 3]Rohini Patil ,Dr.Kamal Shah ," Machine learning in healthcare: Applications, current status and future prospectus", Handbook of Research on Machine Learning" published in Apple Academic Press, CRC Press, Taylor & Francis Group (2022) (ISBN-9781774638682) 4] Rohini Patil, Dr Kamal Shah, Dr,Deepak Bhosle,"Impact of COVID-19 related Stress on Glycaemic control in Patients with Type-2 Diabetes Mellitus" accepted in Journal of Health Science and Surveillance System(Shiraz University of Medical Science,Iran) [Scopus Indexed] 5] Rohini Patil, Dr Kamal Shah," Performance Evaluation of Machine Learning Classifiers for Prediction of Type 2 Diabetes using Stress Related Parameters" presented in International Conference on Data Science, Computation and Security (IDSCS 2022) [Springer, LNNS indexed in ISI Proceedings, Scopus Indexed] 6] Rohini Patil, Dr Kamal Shah," Impact of Lifestyle Parameters for Type 2 Diabetes Prediction: A Machine Learning paradigm "accepted in IJERT journal 2022 **Patent:** Rohini Patil, Dr. Kamal Shah, "Type 2 diabetes prediction using Improved Ensemble Learning Dimensionality Reduction." Indian patent application number 202221005490 dated March 4, 2022. **Copyright:** Rohini Patil, Dr. Kamal Shah, "Mathematical model for Diabe_Lifestyle" Diary Number 8673/2022-CO/L. (24th April 2022)

11	**Observations/Suggestions by Expert/ Supervisor during last RPS/RAC/APS**	RPS I: • To study about training dataset pattern i.e.70-30 or 80-20 (why) • Specify machine learning algorithm and design optimal algorithm using ML. • Proposed methodology needs to be redefined • Social factors not considered in detail • Questionnaire needs to be more in detail • A student has understanding of Research Problem and work in that direction • To find impact of COVID-19 related stress on blood glucose. RPS-II: • Are the dataset age wise will compared as sector/area wise? • What life style factors you are considering? • What threshold value considered for technical comparison of ML algorithm? • Which ML algorithm gives best results compared with others and why? • What are your parameters/factors which you will consider for Novel Algorithm? • Need to devise Novel algorithm to test on standard as well as non-standard dataset • Work on Novel algorithm • Find the effect of stress on COVID-19 diabetes dataset RPS-III: • Work more towards research. • Paper to be published in scopus journal. • List the KPI to check the model performance in Novel • Proposed Algorithm and suggest unique name for your algorithm which will highlight your unique contribution. RPS-IV: • Good Work done can go for pre-synopsis. • Analysis is also done. Technology and medical terms must go hand-in-hand while presentation. • If good paper publication is done on all RO, can go for pre-synopsis.

		• Enough work is done can be considered for pre-synopsis
12	**Observation/Suggestions & Recommendations of Expert/Examiner/Supervisor**	Recommended for thesis submission
13	**Name & Digital Signature with date of Expert/Examiner/Supervisor:**	

Note: On completion of Examination in online mode, the Digitally signed/ Signed scanned copy of report in PDF format should be sent over E-mail by examiner to HoD (Ph.D Programs).

List of Contents

Abstract v
List of Tables ix
List of Figures xi
Abbreviations and Symbols xiii

Chapter 1. Introduction 1

1.1 Background 1
1.2 Diabetes Mellitus 1
1.3 Machine Learning: An Overview 3
1.4 Motivation 3

Chapter 2. Literature Review 5

2.1 Introduction 5
2.2 Survey on Machine Learning Algorithm in Healthcare 5
2.3 Survey on Feature Selection Techniques 8
2.4 Type-2 Diabetes Prediction Using Machine Learning 8
2.4.1 Type-2 Diabetes Prediction for Stress Parameter 8
2.4.2 Type-2 Diabetes Prediction for Lifestyle Parameter 8
2.5 Impact of COVID-19 Related Stress on Blood Glucose Level for Type-2 Diabetes patients 9
2.6 Research Gap and Findings 16

Chapter 3. Problem Definition, Research Objectives, and Scope 17

3.1 Problem definition 17
3.2 Research Objectives 17
3.3 Scope 17

Chapter 4. Impact of Various parameters on Type-2 diabetes prediction 18

4.1 Type-2 Diabetes Data Collection Protocols 18
4.1.1 Development of Machine Learning Model Using LS_Diabetes Dataset 18
4.1.2 Development of Model Using CCWI_Stress Dataset 19
4.1.3 Development of Machine Learning Model to Test Using PIMA Indian Diabetes Dataset 20
4.1.4 Development of Machine Learning Model to Test Using 2019_Diabetes Dataset 20
4.2 Proposed Architectures and Models 20
4.3 Impact of Stress Parameter on Type-2 Diabetes Prediction 22

4.3.1 Description of Stress Data 22
4.3.2 Data Preprocessing and Feature Selection Using Preferred Techniques 22
4.3.3 Diabetes Prediction for Stress Parameter Using Preferred Classifier 23
4.4 Impact of COVID-19 Related Stress on Blood Glucose Level for Type-2 Diabetic Patients 24
4.4.1 Description of COVT_Stress Data 24
4.4.2 Data cleaning 24
4.4.3 Preferred Statistical Methods 24
4.5 Impact of Lifestyle Parameters on Type-2 Diabetes Prediction 25
4.5.1 Data collection 25
4.5.2 Data Pre-processing and Feature Selection Using Preferred Techniques 25
4.5.3 Diabetes Prediction Using Preferred Classifier 26
Chapter 5. Design of Novel Model for Type 2 Diabetes Prediction- Improved Ensemble Learning with Dimensionality Reduction (IELDR) 27
5.1 Proposed Architecture Overview 27
5.2 Type-2 Diabetes Prediction System Using Proposed IELDR 27
5.2.1 Data Collection 27
5.2.2 Pre-Processing 27
5.2.3 Diabetes Prediction Using IELDR 27
Chapter 6. Result and Performance Analysis 30
6.1 Performance Analysis of Type-2 Diabetes Prediction for Stress Parameters 30
6.2 Statistical Analysis for Impact of COVID-19 Related Stress on Blood Glucose Level for Type-2 Diabetics Patients 33
6.3 Performance Analysis of Type-2 Diabetes Prediction for Lifestyle Related Parameters for LS_diabetes dataset 36
6.4 Comparative Analysis of Type-2 Diabetes Prediction for LS_diabetes Dataset and 2019_Diabetes Dataset 39
6.5 Performance Analysis of Type-2 Diabetes Prediction Using IELDR Model for LS_Diabetes Dataset 40
6.6 Comparative Analysis of Type 2 Diabetes Prediction Using IELDR Model with PIMA and 2019_Diabetes Dataset 41
Chapter 7. Conclusion and Future Scope 45
7.1 Conclusion 45
7.2 Future Scope 46
References 47
List of Publications 50
Details of IPR 50
Patent 50

APPENDIX-I 51
APPENDIX-II 52

List of Tables

Table No	Table Content	Page no
Table 1.1	Ranges of laboratory parameters for different levels of diagnosis	02
Table 2.1	Depression, Anxiety &Stress Scale(DASS)	09
Table 2.2	Literature survey summary	10
Table 6.1	Comparative analysis of feature selection methods based on the number of feature	30
Table 6.2	Comparative analysis of classifiers along with different feature selection methods	31
Table 6.3	Comparative analysis of different classifiers on various measures.	32
Table 6.4	Demographics characteristics of study participants (n=199)	33
Table 6.5	Blood glucose parameters comparison	34
Table 6.6	Intragroup comparison of glycaemic parameters based on the severity of stress	34
Table 6.7	P values of intergroup comparison	35
Table 6.8	Group wise comparison	36
Table 6.9	Age group wise p values comparison	36
Table 6.10	Comparative analysis of various feature selection methods	37
Table 6.11	Comparative analysis of model based on feature selection	37
Table 6.12	Performance comparison of the generated prediction models	38
Table 6.13	Comparison of existing method Vs proposed method	39
Table 6.14	Comparison of IELDR model with different classifier model on various measure	40
Table 6.15	Comparative analysis with various dataset	42
Table 6.16	Proposed IELDR model VS best model on PIMA dataset	42
Table 6.17	Proposed IELDR methodology VS methodology used on Diabetes_2019 dataset	43

List of Figures

Figure No.	Figure Content	Page no
Figure 1.1:	Global Projection	1
Figure 1.2:	Rising Number of People with Diabetes in India	2
Figure 2.1:	Random Forest Model	7
Figure 2.2:	Support vector machine	7
Figure 4.1:	LS_diabetes dataset parameters	19
Figure 4.2:	Proposed system flow	21
Figure 4.3:	Research framework	21
Figure 4.4:	Proposed D_Stress model	23
Figure 4.5:	COVI- Stress Methodology	25
Figure 4.6:	Proposed Diabe_lifestyle framework	26
Figure 5.1:	Improved Ensemble Learning Dimensionality Reduction Model	28
Figure 6.1:	Comparison of classification algorithm -Accuracy	30
Figure 6.2:	Accuracy Comparison of classification algorithm	32
Figure 6.3:	ROC curve analysis	32
Figure 6.4:	Distribution of patients based on severity of stress as assessed by DASS-12 score	34
Figure 6.5 :	Intra group analysis	35
Figure 6.6 :	Accuracy Comparison with and without feature selection method	37
Figure 6.7 :	Accuracy Comparison of K-fold, feature selection and hyper parameter optimization	38
Figure 6.8 :	Receiver operating characteristic curve analysis	39
Figure 6.9 :	Accuracy comparison of proposed Dibe_lifestyle method with existing method	40
Figure 6.10 :	Top five best model	41
Figure 6.11 :	Proposed model Vs Ensemble model for case-I	43
Figure 6.12 :	Proposed methodology Vs adopted methodology for case-II	44

Abbreviations and Symbols

BMI	Body Mass Index
DM	Diabetes mellitus
FBG	Fasting blood glucose
PPG	Postprandial blood glucose
HbA1c	Hemoglobin A1c
T2D	Type 2 diabetes
IDF	International Diabetes Federation
WHO	World Health Organization
ML	Machine learning
AI	Artificial Intelligence
NPV	Negative predictive value
PPV	Positive predictive value
ROC	Receiver operating characteristics curve
AUC	Area under curve
TP	True Positive
FP	False positive
TN	True Negative
FN	False negative
SVM	Support vector machine

RF	Random forest
GBC	Gradient Boosting Classifier
LR	Logistic Regression
KNN	K-Nearest Neighbor
DT	Decision Tree
SARS-COV-2	Severe acute respiratory syndrome coronavirus 2
DASS	Depression anxiety stress scale
COVID-19	Corona virus disease
RFE	Recursive feature elimination
*	Multiplication
Σ	Summation
+	Addition
x	Feature Vector
α	constant
β	constant
$\mathcal{F}$	Feature
w'	weight
b'	bias
X_i	Input variable
Y_i	Output variable
x^*	Set of hyper parameters
$g(x)$	Objective score

CHAPTER-1

INTRODUCTION

1.1 Background

Health is not merely the condition of being free from illness, injury or pain. It is overall state of wellness of a person on all levels. As per World Health Organization (WHO) health is "a state of complete physical, mental, and social well-being and not merely the absence of disease" [1].

Physical health refers to physical fitness, which is the outcome of exercise, proper diet and nutrition, cleaning and adequate rest. Mental health realizes individual potential [1]. Lifestyle diseases are driven by several factors including globalization of unhealthy lifestyles, rapid urbanization and population ageing. Such diseases are preventable and their incidence can be minimized with modifications in dietary pattern and lifestyle. Examples of lifestyle related diseases include type 2 diabetes, depression, stroke, metabolic syndrome, heart disease, osteoporosis, and obesity. They may take years to develop and once encountered they are difficult to cure [1, 2].

1.2 Diabetes Mellitus

Diabetes mellitus is one of the most challenging of the chronic diseases from a psychosocial and behavioural perspective. Diabetes is chronic, metabolic disease characterized by raised levels of blood sugar. The global prevalence of diabetes of 9.3% in 2019 is expected to increase to 10.2% in 2030 and 10.9% by 2045[2].

According to global report of International Diabetes Federation (IDF) published in 2017[2], a total of 82 million adults are living with diabetes. South East Asia is the 2nd highest region among all other regions. Overall global projections of diabetes mellitus are shown in Figure 1.1.

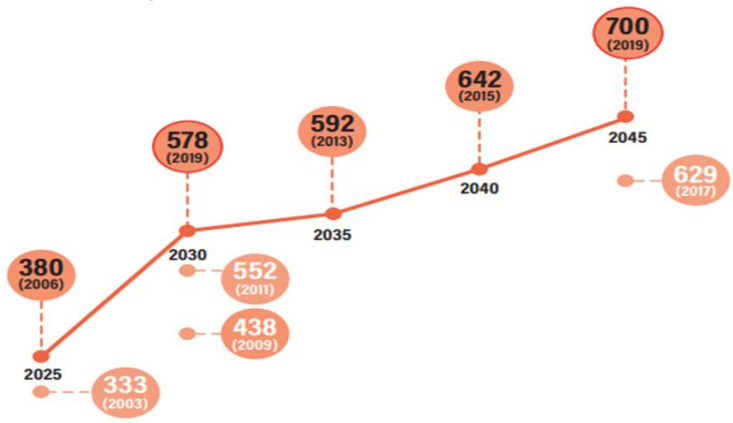


Figure 1.1: Global Projection [2]

In Southeast Asia region, out of 88 million people with diabetes, India contributes 77 million people which is expected to increase to 134.2 in 2045, as shown in Figure (1.2)[2].

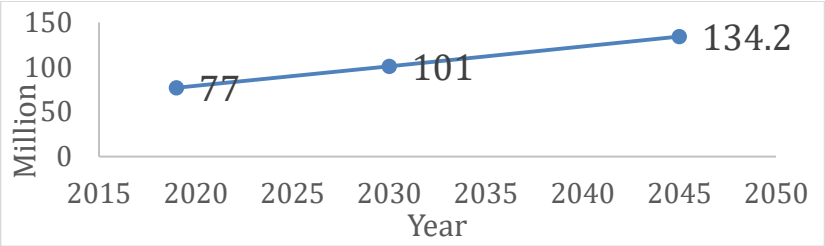


Figure 1.2: Rising Number of People with Diabetes in India [2]

There are 3 types of diabetes mellitus as discussed as below,

- **Type 1 diabetes:** In this type, body does not properly metabolize food due to insufficient or lack of insulin resulting in high level of blood glucose. These patients need insulin injections [1, 2].
- **Type 2 diabetes:** Due to lack of effectiveness of insulin, type 2 diabetes occurs [2]. By adhering to healthy life-style, it can be prevented.
- **Gestational diabetes:** The diabetes which occurs during pregnancy is the gestational diabetes [2].

Pre-diabetes is a state of increase in blood glucose levels [2]. It is a risk factor for the development of diabetes in future. At this earlier stage, preventive measures are necessary to prolong the occurrence of diabetes mellitus.

Diabetes mellitus is a condition associated with short as well as long-term complications such as, kidney failure, stroke, heart attack, vision loss and damage of nerve, physical and cognitive disability, depression. Therefore, early diagnosis and interventions are required for living well with diabetes [2]. Different laboratory investigations in diabetes include HbA1C, FBG and PPG. Ranges foof these parameters for diabetes, prediabetes and normal people are shown in Table 1.1.

Table1.1: Ranges of laboratory parameters for different levels of diagnosis

Due to modern methods of living, stress is contributing huge burden on healthcare. Sleep deprivation, poor eating habits and sedentary lifestyle are significantly contributing to growth of lifestyle diseases.

1.3 Machine Learning: An Overview

Status	Fasting blood glucose level(mg/dl)	Blood glucose level after oral glucose tolerance test (mg/dl)	HbA1c
Diabetes	≥126	≥200	≥6.5%
Pre diabetes	100 to125	140 to 199	5.7 to 6.4%
Normal	≤99	≤139	~5%

Today, extraction of useful data for the diagnosis as well as treatment of diseases is increasingly becoming necessary. ML is a subdivision of data science used for building analytical models. ML is a challenging area in healthcare research. Using ML computers can find insights and patterns that would be missed by the human eye without being programmed to do so. Predictive analysis, can be applied in medical field for the diagnosis of diseases and predicting disease outcomes. Due to outstanding performance on large data and noisy data ML is popular for predictive analytics.

Use of ML can help in healthcare in following ways

1] Prediction of onset and progression of chronic diseases

2] Reduction in mortality

3] To reduce readmissions in hospitals

4] For prevention of hospital acquired infections

5] Prediction of patient propensity to pay

With increasing advances in machine learning and better understanding of disease process today, helps to diagnose the disease at an early phase and reduce the burden of disease and its complications.Stress coping and affective regulations are important in management of diabetes.

1.4 Motivation

Many chronic diseases like diabetes, cancer and cardiovascular diseases are affecting the health of several people across the world. These illnesses are the main contributors to disability and death. In term of cost, prevalence and physical as well as psychological burden on individuals living with the illness of diabetes is one of the largest healthcare worldwide problem. According to the IDF, now India is the second largest country with diabetes cases. India contributes about 49% of world's burden and the current prevalence of 77 million is projected to reach 135 million by 2045 with first largest country with diabetes cases, a major challenge [2].

Diabetes is not just high in numbers but is also leading causes of death in world. The disease can adversely affect eye, nerves, arteries, kidney and heart [2]. Every person is

concerned about his/her health. Unlike other chronic diseases, self-care plays a significant role in the diabetes management.

Lifestyle measures are integral part of care for delaying or preventing the complications due to diabetes. Lifestyle measure refers to the personal habits, attitudes, profession, economic level, etc., that together constitute the mode of living of an individual or group. Another important aspect of lifestyle is stress. Situational response of the body is known the stress response.

The research work is designed to analyze the influence of stress, lifestyle and examine the utility of models for the risk of development of diabetes based on different stress and lifestyle related features.

CHAPTER- 2

LITERATURE REVIEW

2.1 Introduction:

ML is defined as , "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [3]. ML is a division of artificial intelligence (AI) which allows to learn from prior examples, identify patterns from large, complex data sets that can be used to formulate hypotheses [4].

2.2 Survey on Machine Learning Algorithm in Healthcare

Contreas I et al. study focuses on all AI techniques for diabetic prediction. Study shows an AI is powerful tool applicable for prediction of diabetes and prevention of complications [5].

Sneha N.,Tarun G. proposed method for selecting the attributes which will be used in early detection of Diabetes and showed Random forest model and decision tree model has specificity of 98.00% and 98.20% [6]

Jayanthi N. et.al gives survey of prediction model based on traditional and hybrid approach. Showed that in comparison with traditional model, hybrid model to be more accurate [7].

Rajappa T. et.al gives awareness study in rural population based on diet, exercise, and lifestyle. They shows 74% were aware about avoidance of food item. 54% were familiar with the food proportion, 29% was lifestyle modifications and practice followed by 15% [8].

Kelly, Lloyd, et al. shows through physiological and behavior nature, stress affect diabetes, identified as stressful working conditions has been as increased risk for diabetes in people [9-10].

How stress affect diabetes shown by Martinez et al. and found that risk for diabetes people is increased due to stressful working conditions, depression, personality traits or mental health problems. [11, 12, 13]

Machine Learning Algorithms:

ML algorithms are,

- Supervised learning: In this learning, the training consists of input vectors and target vectors. For each instance target vector is represented as, y (p).The model is in the form [(x (p), y (p)) n = 1, . . ., P].

- Unsupervised learning: A learning model is trained on unlabeled data, to predict outcomes .It is pattern based learning.
- Reinforcement learning: It is a reward based learning depends on behaviour in a certain condition.

1] Logistic Regression(LR):

It is a classification problem, classifies the data in the form of binary means if the person has diabetes or not. To develop a logit model-based classification system for determining group membership.LR is based on the Maximum Likelihood estimation technique[3,4,13] .The logit model shown in the following equation,

$$Z = \frac{e^{y}}{1+e^{y}} \quad (1)$$

$$\log\left(\frac{z}{1-z}\right) = y \quad (2)$$

$$\log\left(\frac{z}{1-z}\right) = \text{logit}(z) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \quad (3)$$

Where, $\beta_0 \rightarrow$ Intercept and $\beta_1, \beta_2 \ldots \beta_n$ are the coefficient

Dagliati A. et al. published a paper in 2018 on methods to predict diabetes complications. They developed prediction model for retinopathy, neuropathy, nephropathy at 3,5& 7 yr. time span from first visit using LR with stepwise feature selection algorithm having an accuracy of 83.8%. [14]

Harris et al. shows how stress is a risk factor for diabetes for female centric data. Through traditional risk factors showed indirect and direct using lifestyle related factors [15].

Kumar P.et.al. designed predictive model for the occurrence of anxiety, depression and stress using 8 machine learning algorithm on DASS42 tool. Hybrid model showed good accuracy [16]

Olivera AR, et.al. Developed a model using 4 step and identified result of all ML algorithm and showed RF gives less accuracy while ANN, LR gives best result [17]

2] Decision Trees (DT):

In DT, each node, transition path, and leaf represents a feature, decision, and outcome respectively and decisions are provided in the form of if-then-else statements [4,8]. Through Entropy and Gini Index information gain is calculated which is used for the splitting of a node.

Ayush A.et. al. published a paper in 2015, prediction based on personal indicators. The author used CART model accuracy of 75% based on lifestyle and identified as blood pressure is a significant factor for the development [18].

3] Random Forest:

It is an ensemble classifier formed by training multiple DT simultaneously and hence called "forest". For decision, each RF takes a dataset sample and a subset of features. Each tree performs classification and then selects the classification those are having the more 'votes' or the average votes. Below figure(2.1) shows Random forest model.

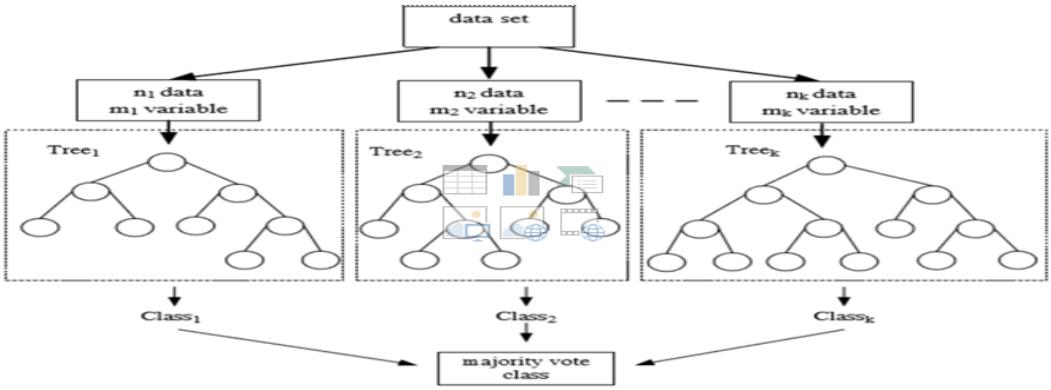


Figure 2.1: Random forest model [4]

Maniruzzaman Md et al. published a paper on classification of methods and develops a system which replaces outliers or missing values by median [19].

4] Support Vector Machine (SVM):

SVM model applicable for both regression as well as classification problems. SVM identifies maximum-margin in such a way that separate hyper plane and minimizes the classification errors shown in figure (2.2)[4, 13].

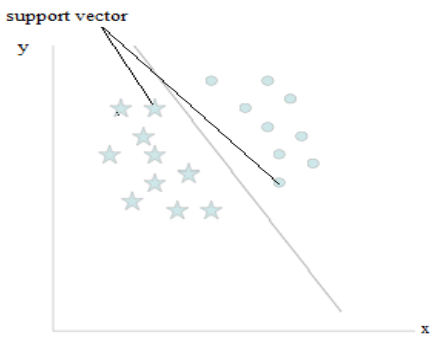


Figure 2.2: Support vector machine [13]

Kaur H.Kumari V. developed predictive model using Linear SVM, RBF SVM, ANN, K-NN, and MDR on PIMA Dataset and shows Linear SVM -88%with Auc-0.90 KNN-88%with Auc-0.92 [20].

5] Gradient Boosting Classifier:

Gradient boosting classifier (GBC) train different weak classifiers and then combining these weak classifiers to form a strong classifier. It works on same training set. It is less prone to over fitting.

Lai H et al. developed model using LR, Gradient Boosting algorithm, RF, DT on different features Age, sex, FBG, BMI, HDL, LDL and shows GBM and LR model performs better i.e.AROC value is 84.07% and 84% the dataset is Canadian centric[21].

6] K- nearest neighbor:

K- nearest neighbor **(**K-NN) is a non-parametric algorithm, it does not make any assumption on underlying data. This algorithm stores the available data and based on similarity it classifies new data point. It is also called a lazy learner. During the training phase it just stores and during testing after getting new data, it calculate the distance using from the query point to the points in the training phase. For distance calculation Hamming distance, Euclidean's distance, Manhattan distance or Minkowski distance method used. After calculation of distance, '*k*' nearest neighbors are selected and the resultant class is calculated based on major votes.

Kaur H.Kumari V. developed predictive model several ML algorithms on PIMA dataset and shows KNN algorithm achieves an accuracy of 88% along with Auc value -0.92 [20].

2.3 Survey on Feature Selection Techniques

Feature Selection was be the most prominent step. It reduces overfitting, increases performance by selection of right subset of features and minimizes execution time. The

feature selection is discussed as,Let the input vector, $X = \{x_1, x_2, \ldots, x_p\}$ with class label is, $Y = \{0,1\}$ with k features such as, $H = \{ f_1, f_2, \ldots, f_k\}$is given as,

$$F(X,Y) = \{(Xj, Yj) |\ Xj \in R\,p\,, Yj \in \{0,1\}\}\} \quad (1)$$

Where $Xj = \{x_1, x_2, \ldots, x_p\}$ are independent variables and $Yj \in \{0,1\}$ is class labels in the dataset.

Wrapper algorithm measures the usefulness of feature, while filter method measures the relevance of a feature.

2.4 Type-2 Diabetes Prediction Using Machine Learning

2.4.1 Type-2 Diabetes Prediction for Stress Parameter

Perceived stress can contribute to the risk of type 2 diabetes. Over time chronic stress can be a risk factor in the development of type 2 diabetes [9-13]. A 12 year longitudinal study in women showed moderate/high stress levels were associated with a 2.3-fold higher risk of diabetes three years later [15].

2.4.2 Type-2 Diabetes Prediction for Lifestyle Parameter

Tigga and Garg designed a risk predictive model using 952 records with 18 features related to lifestyle, health and family background. The proposed model used 6 different classifiers, LR, SVM, KNN, RF, DT and Naïve Bayesian (NB) and compared with PIMA diabetes dataset. The result showed RF achieved an accuracy of 94.1% [31].

2.5 Impact of COVID-19 Related Stress on Blood Glucose Level for Type-2 Diabetes patients

Apart from traditional stress factors, during last one year corona virus disease (COVID-19), a viral disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) has emerged as a significant stress on people, families and community. Diabetes is expected to have poor outcomes after COVID-19 infection [22]. Diabetes has emerged as a significant risk factor for the development of complications in patients with COVID-19 and interaction between COVID-19 and diabetes is knowledge still in progress [23].

Table 2.1: Depression, Anxiety &Stress Scale(DASS)

Version scale	Sub scale	Questions per sub scale	Total no of questions
DASS-42	Depression Anxiety Stress	14 14 14	42

DASS-21	Depression Anxiety Stress	7 7 7	21
DASS-12	Depression Anxiety Stress	4 4 4	12

DASS was originally developed for people having an age of 17 or older. To assess the severity of depression, anxiety and stress. DASS-12 stress sub-scale was used to estimate severity of stress as shown in Table 2.1 [24].

Literature Survey Summary:

Literature Survey is divided into four sections as described below.

- Section1: Survey on diabetes mellitus, research challenges across the world and in India [1..2]
- Section2: Survey on Artificial intelligence, machine learning and prediction model and risk factors associated to diabetes [3…8,14,17..21,25…27,31..39]
- Section3: Survey on stress as risk factor. [9….13,15..16,22..24,28…30]
- Feature extraction using auto encoder[41]

Detailed description of literature survey according to above section has been explained in table 2.2 below.

Table 2.2: Literature survey summary

Sr. No	Journal details	Author	Current Work	Findings
[1]	Accessed on 21st Feb -19	https://www.who.int/en/newsroom/fact-sheets/detail/diabetes		
[2]	Assessed on 22nd Feb 2019	https://www.idf.org/our-network/regions-members/south-east-asia/.../94-india.html		

[3]	First Edition, McGraw-Hill, 1997	Tom Mitchell.		
[4]	Pearson Education India	Saikat D.,et.al.		
[5]	J Med Internet Res 2018	Contreas I, Vehi J.	The study focuses on allAI techniques for diabetic prediction	AI is powerful tools for prediction and prevention of diabetes complications.
[6]	J Big Data (2019)	Sneha N.Gangil T.	Developed a model using optimal feature selection algorithm	showed the DT and the RF has specificity of 98.20% and 98.00%, Naïve Bayesian gives an accuracy of 82.30%
[7]	Journal of Big Data 2017	Jayanthi N, Babu BV, Rao NS.	Discussed various predictive models, steps to develop, applications in diabetes.	Observation shows hybrid model becomes more accurate than traditional
[8]	Int J of Med Sci ence and Public Health 2018	Rajappa T, et. Al	In rural population assess awareness of diet pattern, exercise and lifestyle changes among diabetes patients.	74% - aware of to avoid food. 54% - familiar with to eat food in proportion. Lifestyle modification is 29% and practice was followed by 15%
[9]	Annu. Rev. Public Health 2015	Kelly SJ, Ismai M	Shows relationship between T2D and psychosocial factors	For T2D stress-related factors plays an important role
[10]	Diabetes Spectrum 2005	Lloyd C, et.al	stress affect diabetes, both through physiological mechanisms and via behaviour	For diabetes stress is important , it has effect on glycaemic index which affect lifestyle.
[11]	IEEE Latin America Transacti ons 2018	Martinez A, et.al.	Developed a predictive model of job stress Algo- ZeroR, I Bayes, Simple Logistics, SVM, kNN, AdaBoost algo and	model having accuracy of 0.947,with F-measure of 0.939

			Random Tree.Developed an mobile app-test stress	
[12]	IEEE 2018	Reddy U,et.al.	Analysed pattern of stress in working adults and determine the stress levels using machine learning techniques	Accuracy of different models compared, Boosting-75.13% . Using Decision Trees identified important features of stress in the workplace.
[13]	IJRTE-2019	Patil, R., & Shah, K.	Developed stress prediction model using 3 different classifiers	Showed SVM gives an accuracy with 80.17%
[14]	SAGE journal 2018	Dagliati A, et.al	Algo- LR To predict retinopathy, neuropathy, nepropathy at 3,5& 7 yr time span from first visit already developed complication were excluded	Duration of diabetes & BMI are risk factors for retinopathy & neuropathy hypertension for retinopathy &nephropathy
[15]	PloS ONE 2017	Harris M.L,et.al	Identifies the association between stress and diabetes , and estimation done based on indirect and direct effects	Identified stress is risk factor, along with hypertension, BMI, and physical activity and its effect
[16]	Procedia Computer Science 171 (2020)	Kumar P, et.al	Developed model for prediction of the occurrence of depression ,anxiety and stress using eight machine learning algorithm on DASS42 tool.	Showed hybrid method improved the accuracy.(Kstar +RF)
[17]	Sao Paulo Medical Journal 2017	Olivera AR, et.al.	Algo- LR, ANN, NB, KNN, RF Uses Wrapper strategy for generation of subset of variables forward selection strategy is used. Developed a model using 4 step	Identified result of all ML algo with discretization and without discretization RF gives less accuracy while ANN, LR gives best result
[18]	1st International con.on NGCT:IEEE;2015	Aysh Anand, Divya Shakti	Developed a predictive model based on daily lifestyle activities in the form of questionnaire used CART model of accuracy 75%	Identified Blood Pressure as a significant factor

[19]	Journal of Medical Systems (springer) 2018	Maniruzzaman Md, et.al.	Used 6 feature selection technique- RF,LR,PCA,AOV,FDR) Algo- (NB,LDA, QDA, GPC, ANN,SVM, Adaboost ,LR, DT, RF)replaces outliers or missing values by median	Median used to replace missing values or outliers gives high risk stratification. RF-based classifiers with median gives accuracy of 92.1.1% for K10 and 99.99~100% for JK protocols
[20]	Applied Computing and Informatics2018	Kaur H.Kumari V.	Algo-Linear SVM,RBF SVM,ANN,K-NN,MDR on PIMA Dataset, Feature selection-wrapper,perorm hyper tuning parameters	Model shows Linear SVM - 88%with Auc-0.90 KNN- 88%with Auc-0.92 wrapper method is best than manual selection
[21]	BMC Endocrine Disorders 2019	Lai H, Huang H, et al.	Algo-LR,Gradient Boosting algorithm,RF,DT features-Age, sex, FBG, BMI,HDL,LDL	This study is Canadian centric GBM and LR model performs better i.e.AROC value is 84.07% and 84%
[22]	Ther. 2020;	Schofield J, Leelarathna L, Thabit H	Explored the association between severity of illness and blood glucose and impact of the pandemic on diabetes	
[23]	Duabetes Res Clin Pract. 2020	Hussain A, Bhowmik B, do Vale Moreira NC	Review on characteristics of coronavirus (SARS-CoV-2) and (COVID-19) in people with diabetes and management	
[24]	Int Med J. 2013	Yusoff MSB.	Study was done on total of 839 applicants using DASS-21	The DASS-12 and DASS-9 demonstrated better psychometric values than the DASS-21
[25]	Procedia Computer Science 2015	Nai-aruna N, Moungmaia R.	Developed web application by using real data set for prediction of risk Algo-DT, ANN, LR,RF and NB along with bagging and boosting	Random forest gives best performance random variable selection (used Family history)

[26]	ACM COMPUTE '16, October 21-23, 2016	Chandrakar O, Saini JR.	To observe certain issues in Diabetic risk score (DRS) and address the same. Algo-K-means clustering, discretization collected information of past 10 years	For all population shows 61.5% accuracy
[27]	Int. J. Environ. Res. Public Health 2021,	Deberneh, H.M.; Kim I.	Based on current data developed a ML model to predict occurrence of diabetes in the following year.	Ensemble techniques gives good performance.
[28]	J. of Ambient Intelligence & Humanized Computing	Sanchez W., Martinez A.,et.al.	Highlighted the importance of stress in job related functions and developed a model for recognizing a stress using 6 classifiers	showed good accuracy of RF
[29]	Procedia Computer Science-PerCAA 2019	Ahuja R. Banga A	worked on a model for identification of stress among 206 university students using 4 classifiers	showed SVM gives an 85.71% accuracy
[30]	Springer International Publishing Switzerland	Elena Smets, Pierluigi Casale, Ulf Großekathöfer and Bishal Lamichhane, Walter De Raedt	stress based on physiological responses were measured during laboratory stress test. Used 6 different classifers	Bayesian networks and support vector machines gives results with 84.6 % and 82.7 % respectively.
[31]	Procedia Computer Science-(ICCIDS 2019)	Tigga N.,Garg S	Designed a risk predictive model using 952 instances with 18 features. Used 6 classifiers and compared result to PIMA diabetes dataset.	showed RF achieved an accuracy of 94.1%

[32]	IEEE Access-2020	Hasan MD. K.,Alam MD.A.,Das D.,Hossain E.,Hasan M	Proposed framework by applying preprocessing steps and feature selection using classifiers DT, KNN, AdaBoost ,RF, XGBoost ,NB , and Multilayer Perceptron(MLP). developed weighted ensemble model on Pima Dataset	showed ensemble classifier is the best classifier with highest AUC value as 0.950
[33]	SN Applied Sciences, Springer-2019	Mahabub A	Used 11 classifiers on PIMA dataset by applying hyper parameter tuning and cross validation. Proposed method developed ensemble voting classifier using 3 best classifiers	framework gives an accuracy of almost 86%
[34]	Sensors 2020	Haq A.,Li J.P.,Khan J.,Memon M.H.,Nazir S.,Ahmad S.,Khan G.A.,Ali	Proposed a diagnosis system on the clinical diabetes data set. For feature selection RF, filter based DT and Ada Boost used. The results are compared with wrapper feature selection	showed that the proposed feature selection achieved optimal accuracy
[35]	Int J of Scientific & Technology research 2020	Jahan N.,Islam A.,Mamun A.A	Designed a system of 555*9 data size using MLP, DT and IBK algorithm for different levels of risk of as low, medium, and high on weka tool. proposed an Android application for awareness among people	The result showed that IBK gives best result for 12 fold cross validation with an accuracy 98.73%
[36]	Heliyon 2021	Lama L.,Wilhelmsson O.,Norlander E.,Gustafsson L.,Lager A.et.al	performed research in middle aged people using machine learning and highlighted the importance of stress along with BMI, diet and tobacco consumption	Increasing /decreasing effects on the risk were determined.

[37]	International Journal of Cognitive Computing in Engineering 2021	Kumari S.,Kumar D.,Mittal M	used ensemble soft voting approach for prediction of diabetes using PIMA Indian diabetes dataset. In this study they used ensemble of three algorithms RF,LR and Naïve Byes.	algorithm provided an accuracy of 79.04%
[38]	SN Applied Sciences 2019	Birjais R.,Mourya A.K.,Chauhan R.,Kaur H.	GB, LR and NBC are useful for prediction and diagnosis of diabetes on the PIMA dataset	GB gives accuracy of 86% and accuracy of LR and NBC was 79% and 77% respectively
[39]	International Journal of Cognitive Computing in Engineering,2021	Ahmeda N. , Ahammeda R. ,Islama, M.,Uddina A.,et.al.	Used label encoding, normalization and feature selection techniques. Proposed methodology compared on two different datasets .developed an web app	Showed model provide an accuracy of 2.71% to 13.13% .
[41]	,Clinical Epidemiology and Global Health-2019	Kannadasan K, Edla D., Kuppili V	Proposed a Deep Neural Network framework using stacked autoencoders. framework experimented on PIMA dataset	Showed accuracy of 86.26%

2.6 Research Gap and Findings

- Most of the studies work on PIMA Indian diabetes dataset
- PIMA dataset is female centric only
- Limited study on Stress as a risk factor and that to be using DASS scale only
- Limited studies on lifestyle parameter at the ground level.

CHAPTER-3

PROBLEM DEFINITION, RESEARCH OBJECTIVES AND SCOPE

3.1 Problem Definition

Diabetes mellitus is a group of metabolic diseases characterized by hyperglycaemia resulting from defects in insulin secretion, insulin action, or both. The risk factors include age, family history, unhealthy eating habit, obesity, exercise, smoking and stress related parameters.

High prevalence and associated complications of type 2 diabetes underline the importance of early detection of the disease and need of prompt measures to control it. The goal of early detection can be achieved by regular screening of blood glucose level and glycosylated hemoglobin level. However, it requires high motivation of the person for screening at regular interval.

To design predictive model for early detection of type 2 diabetes mellitus based on stress and lifestyle pattern by using different machine learning algorithms.

3.2 Research objectives

For early stage diagnosis of type 2 diabetes mellitus, the research work aims to design novel predictive model in a predictive time frame based on different risk factors i.e. stress, unhealthy food, smoking, profession and exercise.

The research objectives are

- To analyze the effect of stress related parameters on prediction of type 2 diabetes.
- To analyze the impact of COVID-19 related stress on blood glucose in patients with diabetes of 3 different age groups.
- To analyze the effect of lifestyle factors on prediction of type 2 diabetes.
- Design a novel predictive algorithm which can predict risk of type-2 diabetes mellitus up to 80% accuracy and provide recommendation based on prediction.

3.3 Scope

Considering high burden of diabetes and risk of associated complications, there is a need to develop diabetes prediction model in India. Use of prediction model may help in early diagnosis and intervention which in turn can reduce the healthcare and economic burden associated with disease. The available literature suggests use of standardized PIMA dataset, which is female centric. There is need for a study using comprehensive dataset consisting of parameters from both the genders. The limited available literature from India shows effect of stress on type-2 diabetes and its associated complications.

CHAPTER- 4

IMPACT OF VARIOUS PARAMETERS ON TYPE 2 DIABETES PREDICTION

4.1 Type2 Diabetes Data Collection Process

Datasets: Self-collected questionnaire, PIMA Indian diabetes dataset, Diabetes _2019 dataset

Self-collected (questionnaire based) dataset for the study:

The dataset was developed by web-based questionnaires. A dataset comprising of 374 people with of 35 features the questions were related to demographic information, dietary pattern, life style pathological and stress related factors.

To estimate the effect of COVID-19 related stress on blood glucose of type 2 diabetes patients, collected COVI_Stress dataset comprising of 199 hospitalized diabetes patients with 27 different questions.

4.1.1 Development of Machine Learning Model Using LS_Diabetes Dataset:

- **LS_Diabetes Dataset:**

Data will be collected using a questionnaire-based survey. The questionnaire will be designed and developed in consultation with a qualified healthcare professional an Diabetologist Dr. Bhosle. The developed questionnaire will be validated and pre-tested before using on a large sample data set. The questionnaire used in the dataset used for research administered web based as well as personally through face to face interview. It has 374 instances collected from 35 features. The features are a combination of the demographic, stress related, lifestyle related and other diagnostic results. The dataset consist of both numerical and categorical values from the questionnaire answers. Collected data will be compiled for analysis shown below in figure (4.1)

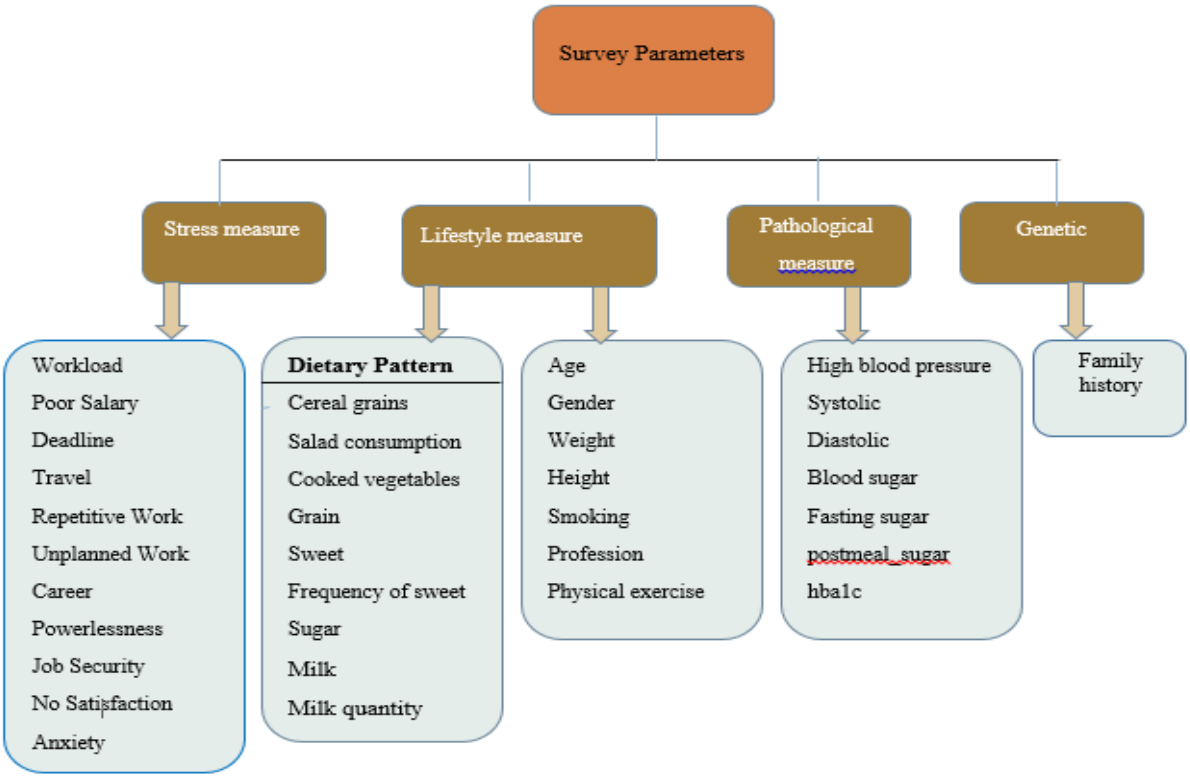


Figure 4.1: LS_diabetes dataset parameters

4.1.2 Development of Model Using COVI_Stress Dataset:

- **COVI-Stress dataset:**

Due to COVID-19 human life adversely affected including physical health, psychological and emotional health, education, travel and economical aspects. To find the impact of stress on blood sugar level during COVID-19 Period, Collected dataset of Patients having type 2 diabetes on anti-diabetes medications. Patients with both gender,

Patients with available investigations related to glycemic parameters (current and previous 3 months), History of current or past history of asymptomatic or symptomatic COVID-19 infection. The records of patients in diabetic groups will be screened for following parameters.

- Age
- Gender(Male/Female)
- Weight
- Height
- BMI
- Residence (Rural/Urban)
- Family status (Single/married/widow)
- Profession (Service/Business/Retired/Student/Housewife/Others)
- Type 2 diabetes since (years)

- On treatment since (years)
- Current HbA1%
- Current FBG mg/dl
- Current PPG mg/dl
- Did COVID-19 affect glycemic control (Yes/No)
- HbA1c % before COVID-19
- FBG mg/dl before COVID-19
- PPG mg/dl before COVID-19
- Did COVID-19 increase psychological stress on the patient? Yes/No
- Was there major change in diet during COVID-19 pandemic? (Yes/No)
- Was physical activity decreased during COVID-19 pandemic (yes/no)
- Was treatment for diabetes taken regularly during COVID-19 pandemic? (Yes/No)

4.1.3 Development of Machine Learning Model to Test Using PIMA Indian Diabetes Dataset

- **PIMA Dataset:** Standard Dataset:

Pima dataset is a UCI Repository dataset, consist of 768 records with female centric data. It contains both 268 diabetic instances and 500 non-diabetic instances. Dataset comprises of numeric-valued 8 attributes. Data contain both medical examination data as well as personal health data.

Age (age) , Body mass index (bmi), Diastolic blood pressure (pres),Number of times pregnant (preg) , 2-h serum insulin (insu),Plasma glucose concentration at 2 h in an oral glucose tolerance test (plas) , Skin fold thickness (skin) , Pedigree function (pedi), Class variable (class)[40].

4.1.4 Development of Machine Learning Model to Test Using 2019_Diabetes Dataset

- **Diabetes_2019 Dataset:**

This dataset was collected for research. Dataset comprises of total 952 participants having aged 18 and above, among them372 are females and 580 are males . The questionnaire was self-prepared based on the constraints that could lead to diabetes. It contains both 267 diabetic instances and 685 non-diabetic instances. Dataset comprises of numeric-valued 18 attributes out of that 17 independent predictor variables and one outcome or variable.

Age, Gender, BMI, Family history, High blood pressure, Blood pressure level, Walk/run/physically active, Smoking, Alcohol consumption, Sleep hours, Sound sleep hours, Daily intake of medicine, Consumption of junk food, Stress, Number of pregnancies, Gestation diabetes, Urination frequency, Diabetic[31]

4.2 Proposed Architectures and Models

The research will cover extensive work on two levels; collection of data and development of predictive models on the collected data. Various classification methods such as gradient boosting classifier (GBC), decision tree (DT), support vector machine (SVM), logistic regression (LR), random forest (RF), K-nearest neighbor(KNN) will be applied on the collected data set.

After collection of data, preprocessing of data will be done. Then data will be divided into training data and testing data. By applying suitable model both accuracy and cross validation accuracy of different models will be calculated. Comparative analysis will be made using different classifiers. By applying different feature selection methods and optimization using Grid search performance of the model can be measured. Below figure (4.2) shows proposed system which is divided into four different phases as, Data collection, Predictive model directing, Model construction and lastly Model validation. The detailed research framework is shown in below figure (4.3).

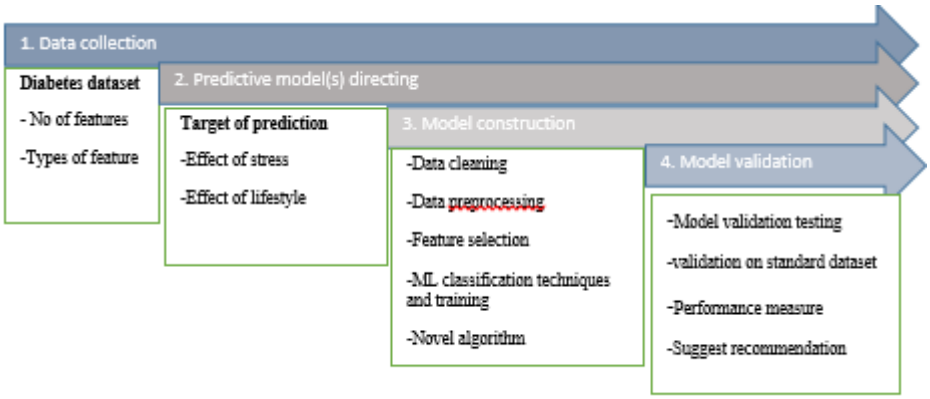


Figure 4.2: Proposed system flow

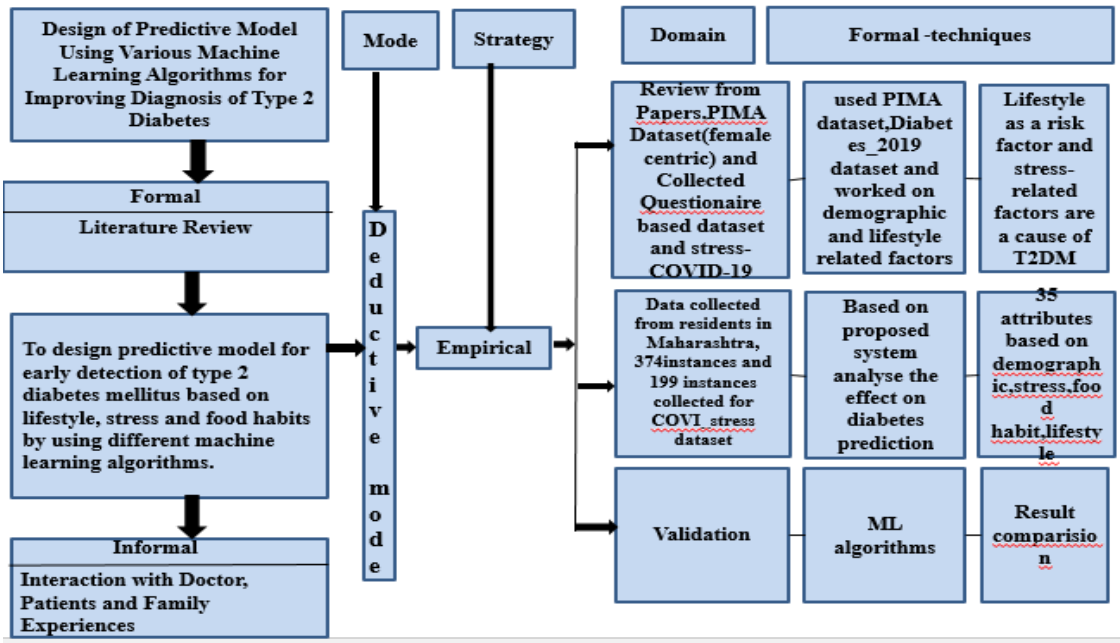


Figure 4.3: Research framework

The Proposed Methodology works on, LS_diabetes dataset which contain both population i.e. Diabetic and non-diabetic and COVI_stress dataset ,to find the impact of COVID related stress in diabetes people.

Model performance can be measured using precision, sensitivity, accuracy, specificity, F1-score,mcc and area under curve (AUC). Sensitivity is called as recall rate. Specificity defines the rate of real negative samples can be detected correctly. These measures are derived using following equations.

$$Accuracy = \frac{Tp + Tn}{\mathrm{Tn + Tp + Fp + Fn}} \quad (1)$$

$$Precision = \frac{Tp}{Tp+Fp} \quad (2)$$

$$Recall/Sensitivity = \frac{Tp}{Tp+Fn} \quad (3)$$

$$F1score = 2 \times \frac{Precision*Recall}{Precision+Recall} \quad (4)$$

$$Specificity = \frac{Tn}{Tn+Fp} \quad (5)$$

MCC=$(Tp + \mathrm{Fp}) - (Tn + \mathrm{Fn})\sqrt{(}\, Tp + \mathrm{Fp})(Tp + \mathrm{Fn}) * (Tn + \mathrm{Fp})(Tn + \mathrm{Fn})$
(6)

Where,

- True positive (Tp) the instances are true (T) while they are (T)
- True negative (Tn) the instances are false (F) while they are (F).
- False negative (Fn) the instances are (F) while they are (T).
- False positive (Fp) the instances are s (T) while they are (F).

4.3 Impact of Stress Parameter on Type-2 Diabetes Prediction

4.3.1 Description of Stress Data

In this research, data collected from general adult population with more than 18 years of age using a pre-developed prevalidated questionnaire. The questionnaire consisted of items related to demographic and stress features, gender, age, weight, height, BMI, anxiety, workload, poor salary, deadline, unplanned work, travel, repetitive work, carrer, job security, powerlessness and no satisfaction with class label. Among these features four variables are numerical variables (age, weight, height, BMI) and rest all other variables are categorical variables. A total of 16 features are to be considered as stress parameters. The dataset consist of total 374 records of both diabetes and non-diabetes population.

4.3.2 Data Preprocessing and Feature Selection Using Preferred Techniques

To analyze the final scores the dataset was preprocessed. As most of the attributes are categorical features so, label encoding is adopted for conversion purpose. The stress features were set to values from 1 to 5 as low, medium, average, high, very high. To improve the model performance used K-fold cross-validation. As selecting important features is the key for success of early diagnosis of diabetes, machine learning approach along with different feature selection techniques are used. The proposed methodology used 5 different feature selection techniques.

4.3.3 Diabetes Prediction for Stress Parameter Using Preferred Classifier

Due to sedentary lifestyle early detection of diabetes is required hence developed a predictive model using six machine learning classifiers as LR, SVM, DT,RF, GBC and KNN. The flow of proposed D_Stress methodology is shown in figure (4.4).

In phase I, preprocessing of data was done. All these 6 models were trained on 80% and 20% of data was used for testing purpose. Accuracy of different machine learning models was calculated for training and testing dataset separately. In order to improve the assessment of model performance 10 fold cross validation technique applied on all the 6 models.

In phase II , study used feature selection techniques namely filter method-Select-K best method, Feature importance technique, Information gain, Correlation technique and hybrid method recursive feature elimination (RFE). The number of features providing highest accuracy for each model using all 5 feature selection methods was noted along with highest accuracy.

In phase III, the model optimization was done using grid search method for all selected classifiers. The optimal hyper parameters were ranked based on their different performance measure.

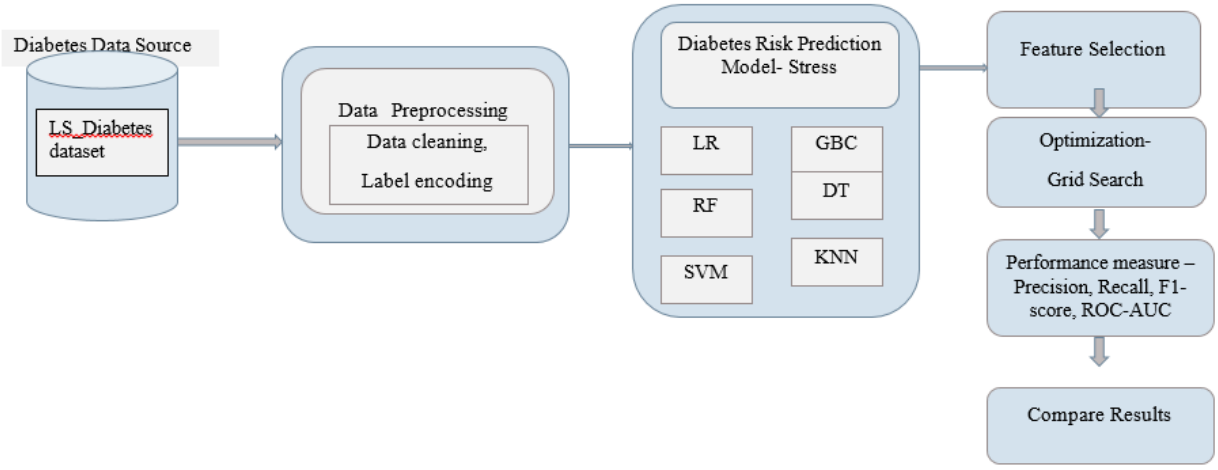


Figure 4.4: Proposed D_Stress model

Proposed D_Stress algorithm is mentioned below.

Algorithm:

```
procedure Label Encoder (LS_Diabetes dataset)
    return ( LS_Diabetes dataset)
procedure Split_ Data (LS_Diabetes dataset)
    Train_data,Test_data= split( attributes,label_var)
    return Train_data,Test_data
procedureL1_Crossvalidation(LS_Diabetesdataset,clf,K=10)
    clf=(LR | RF | KNN |SVM | GBC |DT )
    return
procedure Feature Selector(LS_Diabetes dataset ,F)
    F= (Select k best | Feature importance | correlation |
    Information gain |Recursive feature elimination)
    return F
procedure L2_Crossvalidation ( clf, F)
    return
procedure grid search estimator(clf,F,parameter grid)
    return estimator
procedure L3_Crossvalidation(clf,F,estimator)
    return
```

4.4 Impact of COVID-19 Related Stress on Blood Glucose Level for Type2 Diabetes Patients

Study explored the effect of perceived stress on blood glucose level in diabetes patients who were discharged from hospital after their COVID-19 treatment.

4.4.1 Description of COVI_Stress Data

To accomplish this objective, a total of 199 patients with available laboratory values for fasting blood glucose (FBG) level before admission and post-prandial blood glucose (PPBG) level after discharge were used. Newly diagnosed cases of type 2 diabetes, patients with type 1 diabetes, and patients with gestational diabetes were excluded. A total of 199 patients with female predominant population were included.

Demographic details [gender, age, weight, bmi] and duration of type 2 diabetes was noted. DASS-12 stress sub-scale was used to estimate severity of stress (8).

4.4.2 Data Cleaning

Data compilation was done in master chart i.e. in MS-EXCEL sheet. Representation of categorical data in the form of number and percentages while continuous data are presented as mean, standard deviation.

4.4.3 Preferred Statistical Methods

To estimate severity of stress DASS-12 stress sub-scale was used. There are four items in the stress sub-scale.What causes you stress in your profile? Never (0), Sometimes (1), Often (2), Almost Always (3)

1) I found it hard to wind down
2) I tended to over-react to situations
3) I found myself getting agitated
4) I was intolerant of anything that kept me from getting on with what I was doing

Based on the stress scores patients were categorized (0 to 4) as normal, 5-moderate, 6-severe and >=7extremely severe groups.

Difference in glycaemic parameters (FBG, and PPBG) before and at the time of discharge was estimated. Gender wise, residence wise and age wise (≤ 34 years, 35-49 years and ≥50 years) comparison was done for estimating the difference in severity of stress.

Paired t-test was used to identify the difference in glycaemic parameters before and after COVID-19 related admission. Unpaired t test was used to compare difference in glycaemic parameters between different groups. For categorical variables chi square test was used. Less than 0.05 value was considered to be statistically significant p value.

The methodology (COVI-Stress) used is shown in figure.(4.5)

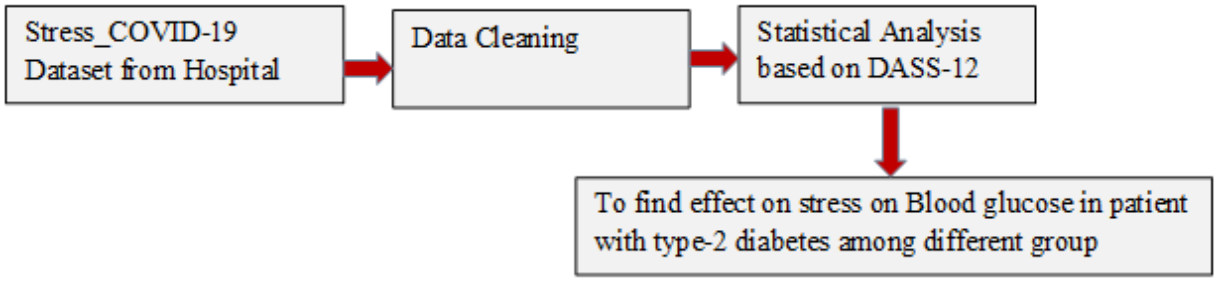


Figure 4.5: COVI- Stress methodology

4.5 Impact of Lifestyle Parameters on Type-2 Diabetes Prediction

4.5.1 Data Collection

To analyze the effect of lifestyle related parameters the dataset used for this study is of 374 instances of both non-diabetes and diabetes population with 18 features. The dataset contains food habit related features, demographic and lifestyle features from the dataset. Along with class label Cereal grains, Salad consumption, Cooked vegetables, Grain, Sweet, Frequency of sweet, Sugar, Milk, Milk quantity, Smoking, Profession, Physical exercise age, weight, height, gender are applied to analyse the effect.

4.5.2 Data Preprocessing and Feature Selection Using Preferred Techniques

From the used features numerical variables are age, weight, height, BMI remaining are categorical variables. BMI is calculated by considering the value of weight and height. To develop a predictive model for early diagnosis of diabetes machine learning approach is used. After collection of data, data pre-processing steps such as data cleaning, label

encoding technique were applied. In this study 5 feature selection techniques were applied. Applying Recursive feature elimination, Select K-best method, Feature importance, Correlation, and Information gain technique important features were extracted.

4.5.3 Diabetes Prediction Using Preferred Classifier

This study uses six different machine learning classifiers for analyzing the effect. The used classifiers are, LR, SVM, DT, RF, GBC and KNN. Training data is 80% and 20% of data is considered for testing. By applying different classifier model has trained and tested. By applying 10-fold cross-validation technique the model is developed. Optimization done using hyper parameter tuning. The below figure (4.6) shows proposed Diabe_lifestyle framework.

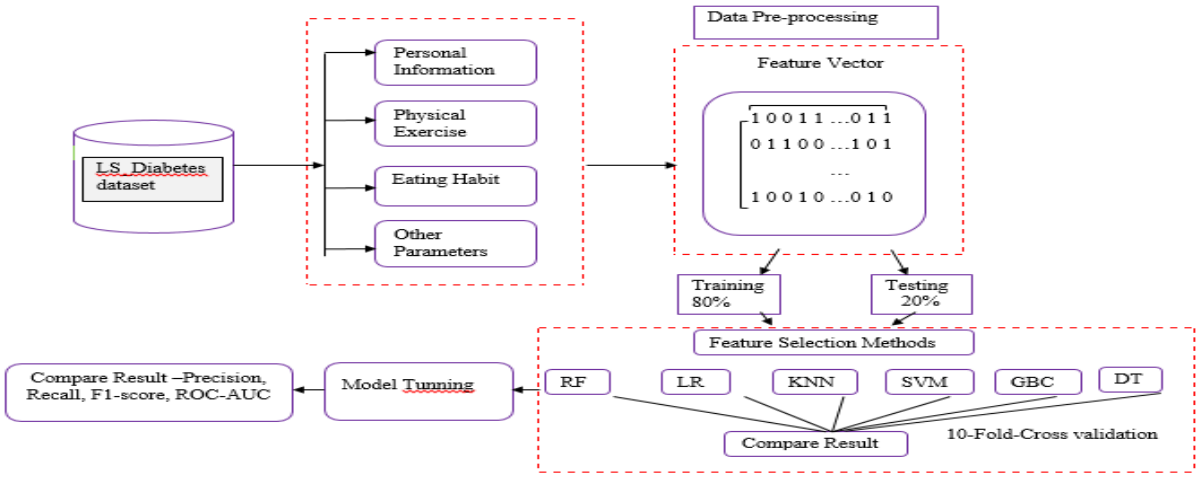


Figure 4.6: Proposed Diabe_lifestyle framework

CHAPTER 5

DESIGN OF NOVEL MODEL FOR TYPE 2 DIABETES PREDICTION – IMPROVED ENSEMBLE LEARNING WITH DIMENSIONALITY REDUCTION (IELDR)

5.1 Proposed Architecture Overview

To achieve this objective proposed auto encoder based two level stacking super learner model. Here we are training the base learner on the complete dataset as well as different blocks of dataset using k-fold. Proposed model does training the base model on different blocks of data and storing a prediction and then combine the predictions from the base learner into a single vector and pass this vector as a input to meta learner to do the final prediction using k-fold cross validation. Existing stacking model trained the models on complete dataset and combining the predictions using Meta learner.

5.2 Type-2 Diabetes Prediction System Using Proposed IELDR

5.2.1 Data Collection

The proposed novel model uses a total of 374 records of data collected using a questionnaire-based survey. The questionnaire used in the dataset used for research administered web based as well as personally through face to face interview. The dataset contains a total of 35 features. These features are a combination of the demographic, stress related, lifestyle related and other diagnostic results.

5.2.2 Pre-Processing

The dataset consist of both categorical and numerical values from the questionnaire answers. The collected dataset is above age group of 18 years .In our collected dataset 1 record was of below 18 years so removed from dataset. As most of the features are categorical and machine learning works only on numerical data, hence label encoding is used to convert categorical features into numerical .The proposed model used auto encoder based feature extraction method. The auto encoder consists of encoder and decoder. The encoder learns how to interpret the input and compress it to an internal representation defined by the bottleneck layer. The decoder takes the output of the encoder and attempts to recreate the input. Once the auto encoder is trained, the decoder is discarded and with the help of encoder to compress data.

5.2.3 Diabetes Prediction Using IELDR

The below figure (5.1) shows the proposed Improved Ensemble Learning Dimensionality Reduction Model (IELDR) The algorithm is mentioned as below.

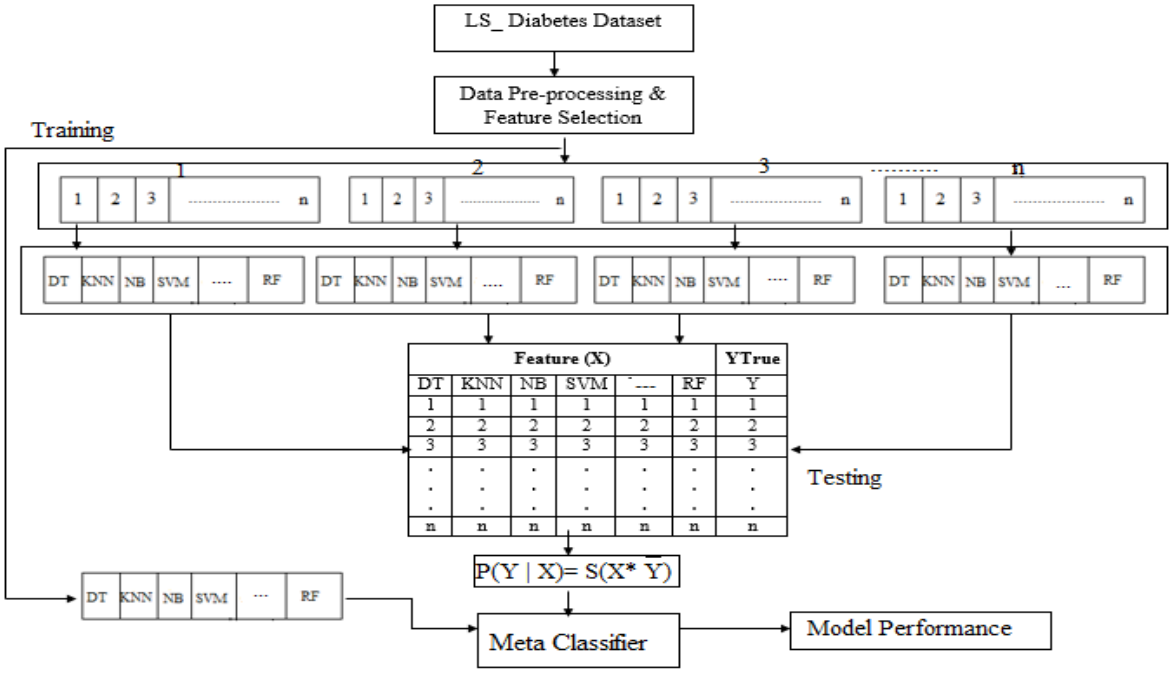


Figure 5.1: Improved Ensemble Learning with Dimensionality Reduction Model

Mathematical model for proposed two levels stacking super learner model with auto encoder for feature extraction

Input: Training Dataset: D_{train} = {features(f_i), Class(y_i)}

Output: Stacking ensemble classifier with super learner *S*

IELDR model is divided into two phases:

3) Feature Extraction
4) Learning Ensemble model.

3. Feature Extraction using autoencoder

- Build the autoencoder using the encoder network and the decoder model.

$$\alpha: \mathcal{D} \rightarrow \mathcal{F} \quad (1)$$

$$\beta: \mathcal{F} \rightarrow \mathcal{D} \quad (2)$$

$$\alpha, \beta = \underset{\alpha,\beta}{argmin} \parallel \mathcal{D} - (\alpha \circ \beta)\mathcal{D} \parallel^2 \quad (3)$$

- The encoder model function α in eq. (1), maps the actual data D to the latent feature space F. The decoder model function β in eq. (2), maps the latent feature space F to the output D. In this case, the output obtained from the decoder is similar to the input function with a reduced feature set as denoted in Eq. (3).
- The encoder model represented by eq. (4) is defined with the sigmoid activation function.

$$z = \sigma(wf + b) \quad (4)$$

where ***z*** is the latent feature dimension.

- Similarly, the decoder model represented by eq. (5) is defined with different weights, biases, and sigmoid activation function:

$$\mathfrak{f}' = \sigma'(w'z + b') \qquad (5)$$

- The error function $\mathbb{E}$ represented by eq. (6) is defined with encoder and decoder network functions. This error function trains the autoencoder using the standard backpropagation method.

$$\mathbb{E}(\mathfrak{f},\mathfrak{f}') = \| \mathfrak{f} - \mathfrak{f}' \|^2 = \| \mathfrak{f} - \sigma'(w'(\sigma(w\mathfrak{f} + b)) + b') \|^2 \qquad (6)$$

- The purpose of using an autoencoder for feature selection is to use the information to encode the data efficiently to regenerate the same data with important features on the decoder side.

4. Two-Levels Stacking Super Learner Model

- **Step1:** Select a 10-fold split of the training dataset D_{train}
- **Step 2:** Identify the appropriate list of base Models *M*
- **Step3:** Train the first-level *M* base learners on the training dataset

  for m ← 1 to M do

  i. Perform 5-fold cross-validation on each base learner b_m on D_{train}

  ii. Collect the cross-validated predicted values from each of the M base algorithms.

  end for

- **Step 4:** Build the new dataset with the *R* cross-validated predicted score obtained from each M base algorithms and construct a new dataset with the shape as R x M.

  for m← 1 to M do

  Build the new dataset as D_{new} = ({features(f_new_i), Class(y_i)}

  where f_new_i = { $b_1(f_i)$, $b_2(f_i)$, $b_3(f_i)$, ……., $b_m(f_i)$)

  end for

- **Step 5:** Build the second-level meta-model based on out-of-fold predicted values and evaluate the meta-model based on a holdout dataset to make predictions using following equation:

$$y=g(f) = B(f)= b' (b_1(f_i), b_2(f_i), b_3(f_i), \ldots\ldots, b_m(f_i)) \qquad (7)$$

- **Step 6: Prediction on new data.**

First, generate predictions from the base models and provide the obtained predictions to the super learner *S* to generate the ensemble prediction on testing or new data samples.

CHAPTER 6
RESULT AND PERFORMANCE ANALYSIS

6.1 Performance Analysis of Type-2 Diabetes Prediction for Stress Parameters

The dataset contains a total of 374 records with 16 features are independent variables, the data is divided into 80-20. Out of these 186 were male and 88 had diabetes. Training accuracy of RF was 100% whereas for LR, GBC, DT, SVM and KNN was 81%, 98%, 82%, 83% and 82% respectively. The Testing accuracy of LR, RF, GBC, DT, SVM and KNN was 84%, 82.67%, 85.33%, 80%, 81.33% and 81.33% respectively. The respective SVM shows good result in comparison with others by achieving an accuracy of 80.28%, shown in figure (6.1).

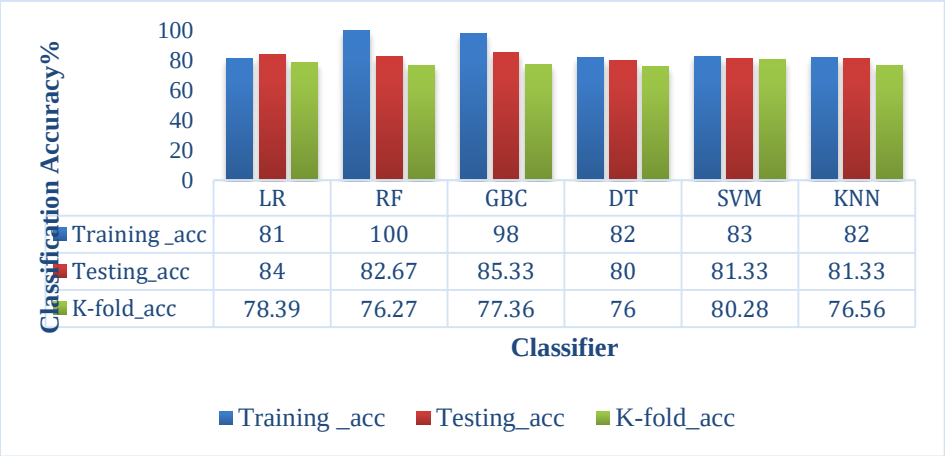


Figure 6.1: Comparison of classification algorithm –Accuracy

By applying various feature selection techniques such as filter method-Select-K best method, Feature importance technique, Information gain, Correlation technique and hybrid method RFE shows SVM gives good accuracy of 82.14% using select K best method. Analysis of Different feature selection technique is given in Table 6.1.

Table 6.1: Comparative analysis of feature selection methods based on number of feature.

Method Used	No of Selected features	Accuracy (%)
LR	16	78.39
LR+ K-Best	9	80.79
LR+Feature importance	12	80.8
LR+Correlation	6	81.07
LR+ Information gain	7	81.07
LR+RFE	7	**81.34**
RF	16	76.27
RF+ K-Best	5	**80.28**
RF +Feature importance	3	77.9

RF +Correlation	6	78.68
RF+ Information gain	7	77.35
RF+RFE	15	78.95
GBC	16	77.36
GBC+ K-Best	5	78.68
GBC +Feature importance	6	**81.08**
GBC+ Correlation	6	80.01
GBC+ Information gain	5	79.77
GBC+RFE	6	80.55
DT	16	76
DT+ K-Best	6	**77.08**
DT +Feature importance	8	75.75
DT +Correlation	10	75.22
DT+ Information gain	7	75.23
DT+RFE	8	77.06
SVM	16	80.28
SVM + K-Best	11	**82.14**
SVM +Feature importance	10	81.35
SVM + Correlation	8	81.6
SVM + Information gain	8	81.33
SVM +RFE	9	81.35
KNN	16	76.56
KNN+ K-Best	11	80
KNN +Feature importance	8	78.14
KNN+ Correlation	10	**80.26**
KNN+ Information gain	7	80

Table 6.2: Comparative analysis of Classifiers along with different feature selection methods

Feature Selector	LR	RF	GBC	DT	SVM	KNN
chi-squared statistical test	80.79	80.28	78.68	77.08	82.14	80
Feature importance	80.8	77.9	81.08	75.75	81.35	78.14
Correlation	81.07	78.68	80.01	75.22	81.6	80.26
Information Gain	81.07	77.35	79.77	75.23	81.33	80
RFE	81.34	78.95	80.55	77.06	81.35	-

Hyper parameter tuning can be done using grid search method shows among 6 different classifiers SVM and GBC shows an highest accuracy of 86.67%. Accuracy comparison of with feature selection, without feature selection and hyper parameter tuning accuracy comparison is shown in below figure (6.2).

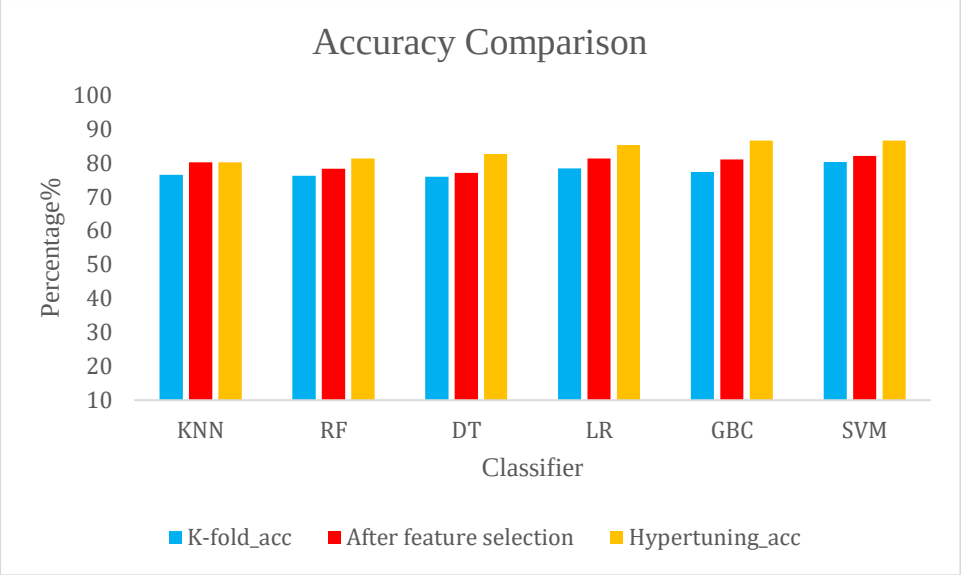


Figure 6.2: Accuracy Comparison of classification algorithm

Performance evaluation of these classification algorithm can be done by different measure such as, accuracy, positive predicted value can also be called as precision, recall and F-measure. Comparison of ROC curve analysis of various classification algorithm shown, figure (6.3). Comparative analysis of Different Classifiers on Various Measures is shown in Table 6.3.

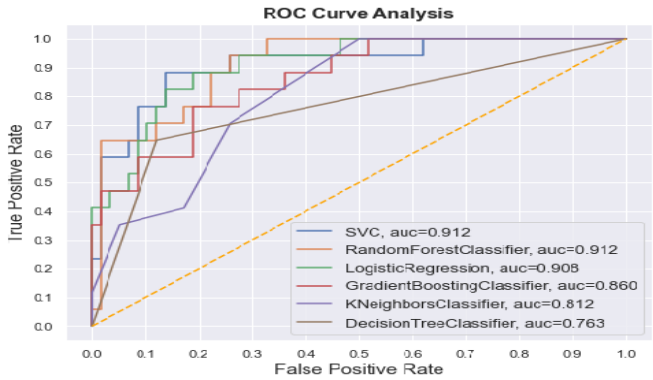


Figure 6.3: ROC curve analysis

Table 6.3: Comparative analysis of different classifiers on various measures.

Classifier	Accuracy	Precision	Recall	F1-score	Roc_Auc Score
LR	85.33	0.85	0.85	0.84	90.87
RF	81.33	0.81	0.81	0.77	91.17
GBC	86.67	0.87	0.87	0.85	86
DT	82.67	0.83	0.83	0.83	76.31
SVM	86.67	0.87	0.87	0.85	91.18
KNN	80.26	0.8	0.81	0.79	81.24

The result shows that different stress factors which have considered to be contributing for the development of diabetes, so there is a need to develop a predictive model at an early stage.

6.2 Statistical Analysis for Impact of COVID-19 Related Stress on Blood Glucose Level for Type-2 Diabetes Patients

In this study, a total of 199 patients with mean (SD) age of 54 (12.8) years were included. The study included 122 (61.30%) females and 77 (38.7%) males. The mean (SD) weight and BMI of patients was 78.1 (14.6) kg and 30.9 (7.5) kg/m^2 respectively. A total of 116 (58.3%) patients were from urban area whereas 83 (41.7%) were from rural area **(**Table 6.4**)**. A total of 114 (57.3%) patients were housewives.

Table 6.4: Demographics characteristics of study participants (n=199)

Parameter	Result
Mean (SD) age in years Age range in years	54 (12.8) 30-92
Gender n (%) Male Female	 77 (38.7%) 122(61.3%)
Mean (SD) weight in kg Range of weight in kg	78.1 (14.6) 30-114
Mean (SD) BMI in Kg/m^2 Range of BMI in Kg/m^2	30.9 (7.5) 10.6-51.5
Residence n (%) Rural Urban	 83 (41.7%) 116 (58.3%)
Profession n (%) Business Service Housewife Other	 7 (3.5%) 10 (5.03%) 114 (57.3%) 68 (34.2%)
Mean (SD) duration of diabetes in years Range of duration of diabetes in years	5 (3.6) 0.08-25

A total of 73 (36.7%) had perceived stress. A total of 54 (27.1%) patients had moderate stress. Severe or extremely severe stress was observed in 19 (9.6%) patients (Figure 6.4).

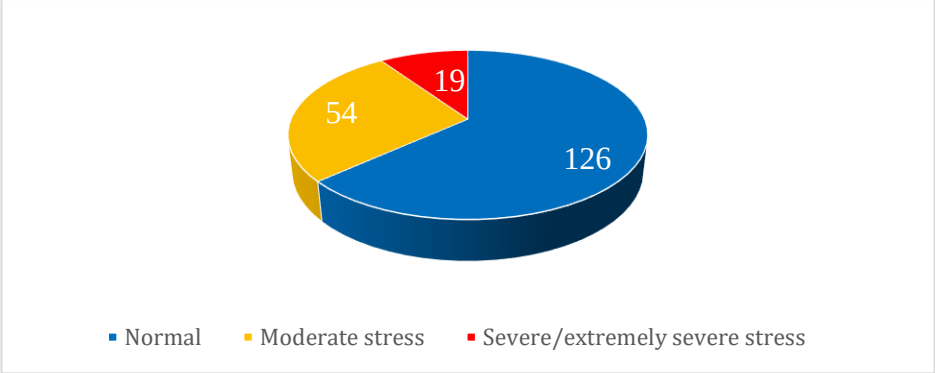


Figure 6.4: Distribution of patients based on severity of stress as assessed by DASS-12 score

In a group of moderate stress out of 54 patients, 23 (42.6%) were females and 31 (57.4%) were male. In a group of patients with severe/extremely severe stress out of 19, a total of 11 (57.9%) were females and 08 (42.1%) were male.
(Table 6.5).shows the significance difference found between FBG and PPG value of before and at the time of discharge.

Table 6.5: Blood glucose parameters comparison

	Before admission	After discharge	P value
Mean (SD) fasting blood glucose mg/dl	168.4 (30.6)	195.9 (28.8)	**<0.001**
Mean (SD) post prandial blood glucose mg/dl	312 (62.3)	351.6 (61.9)	**<0.001**

Comparison of glycemic parameters based on the severity of stress is presented in **(**Table 6.6).

Table 6.6: Intragroup comparison of glycemic parameters based on the severity of stress

	No stress (n=126)	Moderate stress (n=54)	Severe/Extremely severe stress (n=19)
Mean (SD) age in years	52.6 (13)	57.3 (12.5)	54.2 (10.6)
Mean (SD) weight in Kg	80 (13.9)	74 (15.1)	77.2 (16.5)
Mean (SD) BMI in kg/m^2	32.6 (7.3)	27.8 (6.7)	28.3 (7.5)
Gender n (%) Male	38 (30.16%)	23 (42.59%)	8 (42.11%)

Female	88 (69.84%)	31 (57.41%)	11 (57.89%)
Mean (SD) fasting blood glucose (mg/dl) Before admission After admission	164.7 (30) 195.1 (27.8)	171.3 (24.5) 198.3 (30.8)	185 (43.2) 194.2 (30.8)
P value	<0.001	**<0.001**	0.43
Mean (SD) post-prandial blood glucose (mg/dl) Before admission After admission	309.2 (62.3) 347.3 (64.7)	315.8 (59) 359 (52.6)	319.7 (72.8) 358.7 (68.1)
P value	<0.001	**<0.001**	0.06

There was significant difference in mean FBG before and after admission in patients without stress and those with moderate stress. However, there was no significant difference in mean FBG in patients with severe/extremely severe stress (p=0.43). Similar observations were seen for PPBG (without stress p<0.001; moderate stress p<0.001; severe/extremely severe stress p=0.06) Intra group analysis for FBG and PPBG shown in (Figure 6.5)

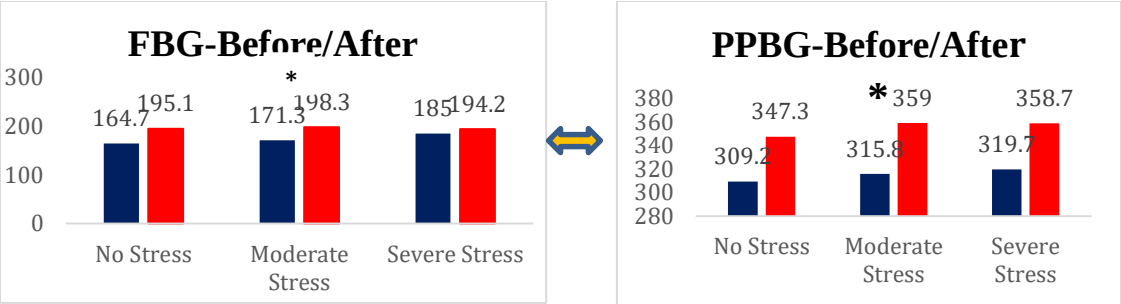


Figure 6.5:Intra group analysis

Intergroup analysis showed no difference in the FBG or PPBG in patients without stress versus moderate stress, moderate versus severe/extremely severe stress, no stress versus severe/extremely severe stress before admission or after discharge shown in (Table 6.7).

Table 6.7: P values of intergroup comparison

	Without stress versus moderate stress	**Moderate versus severe/extremely severe stress**	**Without stress versus severe/extremely severe stress**
Fasting blood glucose before admission	0.15	0.09	0.01
Fasting blood glucose at the time of discharge	0.497	0.68	0.98
Post-prandial blood glucose before admission	0.51	0.82	0.50
Post-prandial blood glucose at the time of discharge	0.24	0.99	0.48

Overall study population showed rise in fasting and postprandial glucose level in type-2 diabetes patients discharged after COVID-19 treatment. The rise was significant in type-2 diabetes without stress and those with moderate stress.

In patients with stress, compared age wise, gender wise as well as residence wise and found that there was significant difference in the FBG and PPBG based on their age, gender and residence i.e. urban versus rural population shown in (Table6.8)

Table 6.8: Group wise comparison

Groups	**P-Value**
Age-Stress	**< 0.05**
Gender-Stress	**< 0.05**
Urban-Stress	**< 0.001**
Rural-Stress	**< 0.001**

Overall age group wise comparison showed significant difference in severity of stress ($p<0.05$) shown in (Table 6.9).

Table 6.9: Age group wise p values comparison

Age Group	p-value
Old age(>=50) --- Stress	P<0.05
Middle age (35-49)--- Stress	p>0.05
Young age (<=34)--- Stress	p>0.05

6.3 Performance Analysis of Type-2 Diabetes Prediction for Lifestyle Related Parameters for LS_diabetes dataset

By considering 80-20 data, LR,SVM, GBC,RF,DT and KNN model has trained . By selecting a subset of the most relevant features in the dataset feature selection techniques applied and showed RFE for GBC gives highest accuracy of 82.92%.Number of feature selection based on model and their accuracy is shown in Table (6.10) and feature selection method comparison is shown Table (6.11)

Table 6.10: Comparative analysis of various feature selection methods

Classifier	No of Features					
	Recursive Feature Elimination	Select K-Best	Feature Importance	Correlation	Information Gain	Feature selection accuracy
RF	9	11	13	11	12	80.75
GBC	10	11	11	12	12	82.92
SVM	10	8	12	10	11	82.39
KNN	--	10	14	13	12	80.24
LR	14	8	5	11	6	80.48
DT	10	8	13	11	12	81.57

Table 6.11: Comparative analysis of model based on feature selection

Classifier	RFE	Select K-Best	Feature Importance	Correlation	Information Gain
RF	80.75	80.24	80.53	78.62	80.51
GBC	82.92	79.16	79.96	78.62	80.23
SVM	80.23	81.31	81.85	81.31	82.39
KNN	-	80.24	80.24	80.24	80.24
LR	80.24	79.95	80.48	80.48	80.48
DT	81.57	76.22	78.07	77.04	76.74

The below figure (6.6) shows accuracy comparison based on using feature selection and before applying feature selection.

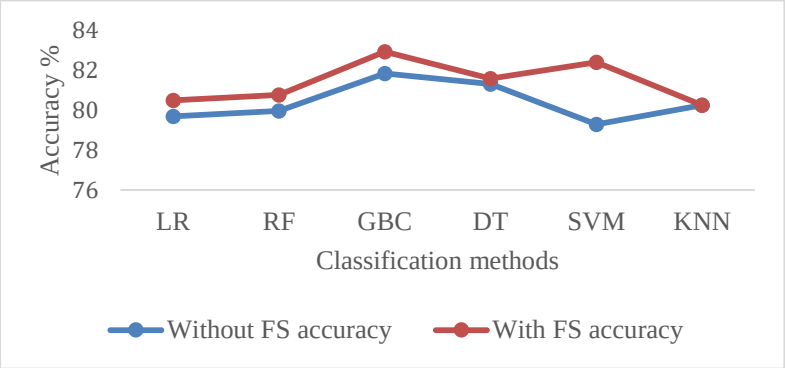


Figure 6.6: Accuracy Comparison with and without feature selection method

Optimization of machine learning model can be done using Grid search optimization technique. Below figure (6.7) shows comparison of result of K-fold accuracy, feature selection accuracy and hyper parameter optimization accuracy. The result shows that LR,RF and KNN achieved an accuracy of 85.33%.

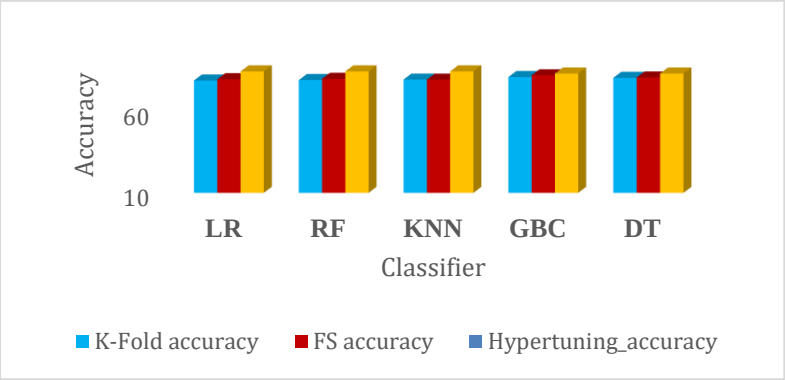


Figure 6.7: Accuracy Comparison of K-fold, feature selection and hyper parameter optimization

Performance measure used for analyzing the effect of lifestyle parameters shown in below table (6.12). Which shows RF, LR and KNN model gives an accuracy of 85.33%. Precision is nothing but measure of correctness .RF gives highest precision value of 0.88.Recall is the measure of exactness or completeness, here LR, KNN and RF model shows good recall value of 0.85.

Table 6.12: Performance comparison of the generated prediction models

Classifier	Precision	Recall	F1-score	Roc-Auc score	Accuracy
RF	0.88	0.85	0.82	92.8	85.33
GBC	0.84	0.84	0.81	92.49	84
LR	0.86	0.85	0.83	87.52	85.33
K-NN	0.85	0.85	0.84	85.55	85.33
DT	0.84	0.84	0.84	82.5	84
SVM	0.80	0.81	0.80	81.74	81.33

All model shows high predictive performance, with roc-auc value ranging from 0.928 to 0.817. By comparing with other models, the best performance showed by RF model, with an AUC value of 0.928.The outstanding performance of the Random forest model helps in the assessment of risk of developing type 2 diabetes. Figure (6.8) shows roc curve analysis.

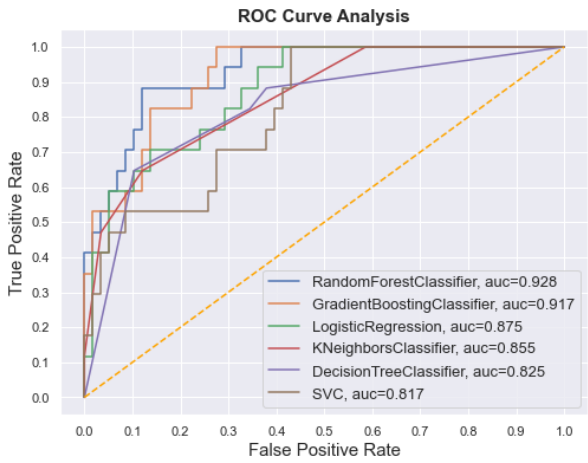


Figure 6.8: Receiver operating characteristic curve analysis

Using machine learning, this study developed risk assessment models for the risk of developing diabetes using lifestyle features.

6.4 Comparative Analysis of Type-2 Diabetes Prediction for LS_diabetes Dataset and 2019_Diabetes Dataset

Experimental Validation:

The proposed Diabe_Lifestyle method applied on Diabetes_2019 dataset and compared the result with the existing researcher and found that proposed method shows 0.16%, 1.9%,10.24%,6.17%,16.41% better accuracy in LR,RF,DT,SVM and KNN. Below table (6.13) and Figure (6.9) shows the comparative analysis.

Table 6.13: Comparison of existing method Vs proposed method

Classifier	Proposed Method	[31]
DT	94.24	84
RF	96	94.1
SVM	92.67	86.5
LR	85.86	85.7
KNN	93.71	77.3
GB	94.24	

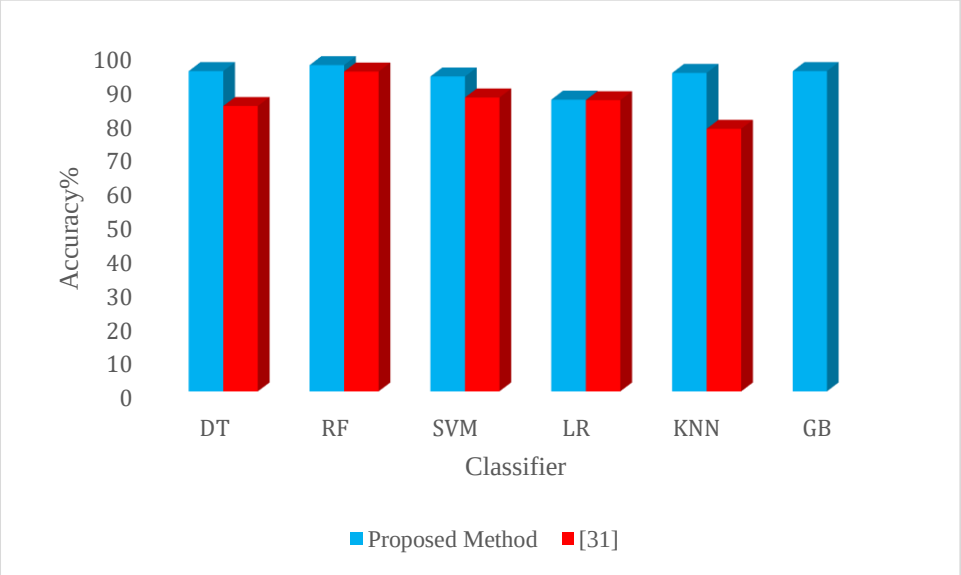


Figure 6.9: Accuracy comparison of proposed Diabe_lifestyle method with existing method

6.5 Performance Analysis of Type-2 Diabetes Prediction Using IELDR Model for LS_Diabetes Dataset

Proposed novel -IELDR model applied on LS diabetes dataset achieved an accuracy of 98.67% in comparison with various classifier shown in table (6.14) IELDR model gives

good result on various performance measure. Top best five model based comparison shown in figure (6.10)

Table 6.14: Comparison of IELDR model with different classifier model on various measure

Model	Accuracy	Precision	Sensitivity	Specificity	f1 score	ROC	Log_ Loss	Mathew_ corrcoef
IELDR	**98.67**	**0.9524**	**100**	**0.9818**	**0.9756**	**0.9909**	**0.46**	**0.9670**
MLP	90.67	0.8421	0.8	0.9455	0.8205	0.8727	3.22	0.7580
SVC	90.67	1.0000	0.65	1.0000	0.7879	0.8250	3.22	0.7594
LR	88.00	0.9231	0.6	0.9818	0.7273	0.7909	4.14	0.6797
LDA	86.67	0.8125	0.65	0.9455	0.7222	0.7977	4.61	0.6428
Adaboost	86.67	0.8125	0.65	0.9455	0.7222	0.7977	4.61	0.6428
SGD	85.33	0.8462	0.55	0.9636	0.6667	0.7568	5.07	0.6000
XGB	85.33	0.8462	0.55	0.9636	0.6667	0.7568	5.07	0.6000
RF	85.33	1.0000	0.45	1.0000	0.6207	0.7250	5.07	0.6124
KNN	85.33	1.0000	0.45	1.0000	0.6207	0.7250	5.07	0.6124
Extra tree classifier	85.33	1.0000	0.45	1.0000	0.6207	0.7250	5.07	0.6124
GBM	84	0.7857	0.55	0.9455	0.6471	0.7477	5.53	0.5623

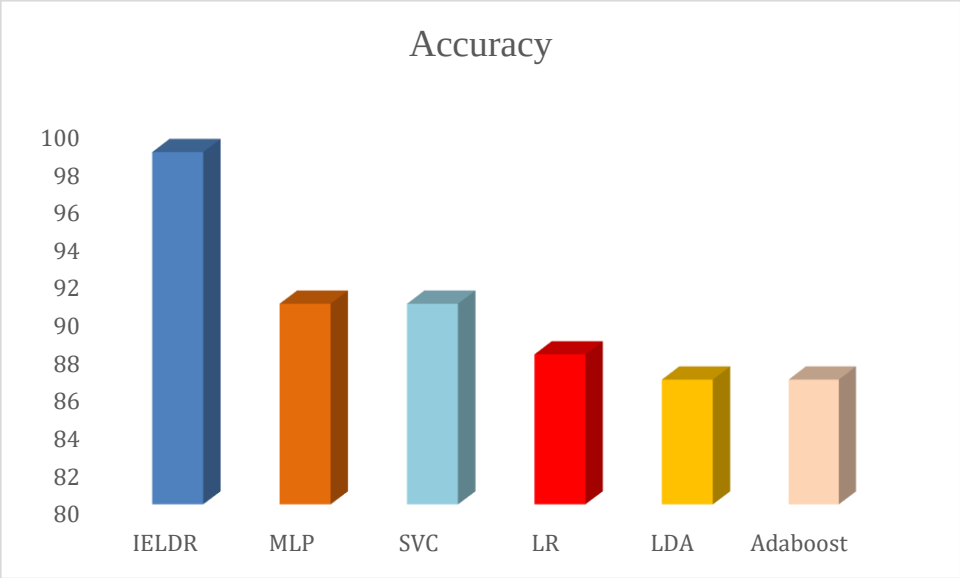


Figure 6.10: Top five best model

The system predicts risk for developing of type 2 diabetes in healthy person based upon the persons current lifestyle pattern, the system calculates the risk score and offers recommendation.

1] Predictive analysis shows your health score is good, Follow the healthy practices

2]

```
-----------------------
Diabetes Risk Analysis
-----------------------
You are Predicted On the way to developing diabetes ,if you don,t change lifestyle.Your
diabetes risk index is 44.54/50
```

3]

```
-------------------------
Diabetes Risk Analysis
-------------------------
You are predicted having diabetes so you are on higher risk,kindly consult with
doctors.Your diabetes risk index is 49.58/50
```

6.6 Comparative Analysis of Type-2 Diabetes Prediction Using IELDR Model with PIMA and 2019_Diabetes Dataset

Experimental Validation:

The proposed IELDR model evaluated on PIMA Indian diabetes dataset, diabetes_2019 dataset and on Ls_diabetes dataset and showed accuracy, sensitivity, ROC and mcc value gives good result in comparison with other dataset shown in table (6.15).

Table 6.15: Comparative analysis with various dataset

Evaluation Type	PIMA_Diabetes[40]	Diabetes_2019[31]	LS_Diabetes
Precision	82.09	96.89	95.24
Sensitivity	59.78	98.42	100
Specificity	92.59	98.42	98.18
F1 score	69.18	97.65	97.56
ROC	76.19	96.08	99.09
Mathew_corrcoef	0.57	0.929	0.97

Case I-

Comparison of proposed IELDR model and best model experimented on PIMA Indian diabetes dataset from state of the art is shown in table (6.16) and comparison on various measures showed in figure (6.11) highlight proposed IELDR method gives good result with respect to accuracy, precision and specificity.

Table 6.16: Proposed IELDR model VS best model on PIMA dataset

Case1-[PIMA[40]]				
Evaluation Method	**IELDR**	**[37]**	**[42]**	**[43]**
Accuracy	**80.71**	79.08	75	76.3
Precision	**82.09**	73.13	47	75.9
Specificity	**92.59**	83.44	76	

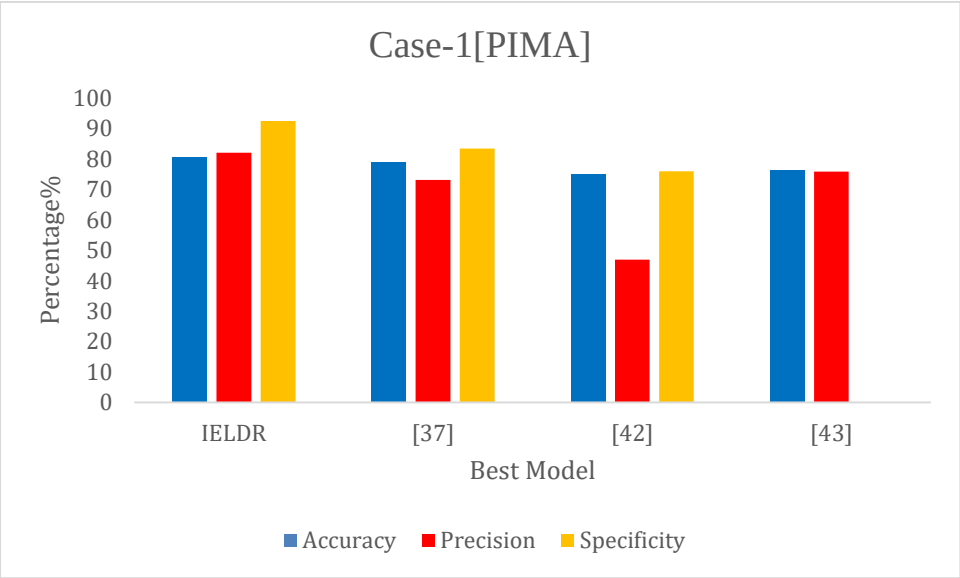


Figure 6.11: Proposed IELDR model Vs Best model for case-I

Case II-

Proposed IELDR based framework and the framework used by existing researchers on diabetes_2019 dataset have been compared and found to be IELDR framework gives good result shown in table (6.17) and figure (6.12).

Table 6.17: Proposed IELDR methodology VS methodology used on Diabetes_2019 dataset

Case2-[Diabetes_19][31]			
Evaluation Method	**IELDR model**	**[31]**	**[39]**
Accuracy	**96.85**	94.1	96.81
Precision	**96.89**	97.6	96
Sensitivity	**98.42**	94.3	92.3
Specificity	98.42	93.4	98.52
F1 score	**97.65**	95.9	94.11

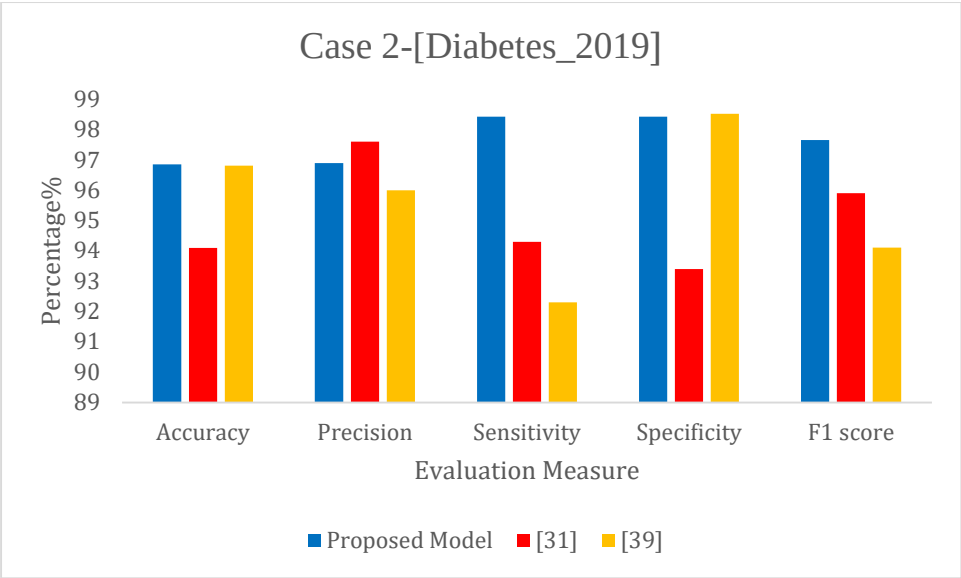


Figure 6.12: Proposed methodology Vs adopted methodology for case-II

The Proposed IELDR model gives good result in terms of accuracy, precision, sensitivity specificity and f1 score value with the existing researchers.

CHAPTER-7

CONCLUSION AND FUTURE SCOPE

7.1 Conclusion

Type 2 diabetes is a health concern worldwide including in India because of its complications, direct and indirect costs and psychological burden associated with it. Uncontrolled disease and delayed treatment can be associated with several complications affecting organs like kidneys, eyes, nerves and heart. The disease affects productive age group and elderly people posing significant economic burden on the family.

LS_Diabetes dataset was developed by collecting records from 374 people with 35 features. Using this dataset, proposed framework examined the performance of 6 different machine learning model (SVM,LR, RF,DT,GBC,KNN) by applying 5 different feature selection methods namely ilter method-Select-K best method, Feature importance technique, Information gain, Correlation technique and Hybrid method as RFE. Hyper parameter optimization was done using grid search cv. The proposed method was validated against from PIMA and Diabetes_2019 database from another Indian researcher. Using the LS-Diabetes dataset, we used stress related parameters for analysing effect of stress. Machine learning models GBC and SVM showed good accuracy, precision and recall value for prediction of type 2 diabetes.

The impact of COVID-19 related stress on blood glucose level in type-2 diabetes patients was examined using a dataset of 199 diabetes patients. Results of COVI-Stress dataset showed significant increase in blood glucose level in type-2 diabetes patients before and during discharge of COVID-19 treatment and the rise was significant in type-2 diabetes without stress and those with moderate stress.

Effect of lifestyle based parameters was also analyzed using LS_Diabetes dataset. The highest feature selection accuracy was observed in the GBC with RFE method. Based on the comparative analysis, RF and GBC showed highest ROC value.

A novel- Improved Ensemble Learning with Dimensionality Reduction Model (IELDR) was designed. Auto encoder based feature extraction method along with ensemble learning algorithm was used. This model showed better results in terms accuracy, sensitivity, ROC and Mathew correlation when compared with standard PIMA and Diabetes_2019 dataset. This proposed IELDR system can effectively predict the risk of type 2 diabetes development in healthy people based on the current lifestyle pattern. The system calculates the risk score and offers recommendation.

The findings of this study will help to the society considering that machine learning plays a vital role in prediction of type 2 diabetes. Overall, results of our research and validation with other dataset provides significant insights for prediction of type 2 diabetes development. Stress, lifestyle should be considered as strong risk factors for the development of diabetes mellitus and appropriate strategies should be designed to control these risk factors.

7.2 Future Scope

Studies with larger sample size of type-2 diabetes patients with stress may be needed to provide more insights regarding difference between those without stress and moderate to severe/extremely severe stress. Further large scale studies may be conducted covering wider population in India. Studies for estimation of duration required for development of type 2 diabetes based on the lifestyle of a person and prediction of organs at higher risk of development of complications can also be planned.

References

[1] http://www.who.int/en/news-room/fact-sheets/detail/diabetes accessed on 21st Feb 2019

[2] https://www.idf.org/our-network/regions-members/south-east-asia/.../94-india.html assessed on 22nd Feb 2019

[3] Tom Mitchell, Machine Learning. First Edition, McGraw- Hill, 1997

[4] Saikat Dutt,Amit Kumar Das , Machine Learning ,Pearson Education India.

[5] Contreas I, Vehi J. Artificial intelligence for diabetes management and decision support: Literature review. J Med Internet Res 2018. 20: e10775. doi: 10.2196/10775

[6] Sneha and Gangil. Analysis of diabetes mellitus for early prediction using optimal features selection: J Big Data (2019) 6:13 https://doi.org/10.1186/s40537-019-0175-6

[7] Jayanthi N, Babu BV, Rao NS. Survey on clinical prediction models for diabetes prediction. J Big Data 2017; 4:26

[8] Rajappa T, Ponniraivan K, Kalyan H, Selvaraju K, Karunanandham S. Assessment of degree of awareness about diet, physical exercise, and lifestyle modifications among diabetic patients. Int J Med Sci Public Health 2018;:481-48

[9] Kelly SJ, Ismai M. Stress and type 2 diabetes: A review of how stress contributes to the development of type 2 diabetes. Annu. Rev. Public Health 2015. 36:441–62

[10] Lloyd C, Smith J, Weinger K. Stress and diabetes: A review of the links. Diabetes Spectrum 2005;18:121-27

[11] Martinez A, Sanchez W, Benitez R, Gonzalez Y, Mejia M, Otiz J. A job stress predictive model evaluation through classifier's algorithms. IEEE Latin America Transactions, VOL. 16, NO. 1, JAN. 2018

[12] Reddy, U. S., Thota, A. V., & Dharun, A. (2018). Machine Learning Techniques for Stress Prediction in Working Employees. 2018 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2018, 1–4. https://doi.org/10.1109/ICCIC.2018.8782395

[13] Patil, R., & Shah, K. (2019). Assessment of Risk of Type 2 Diabetes Mellitus with Stress as a Risk Factor using Classification Algorithms. International Journal of Recent Technology and Engineering, 8(4), 11273–11277. https://doi.org/10.35940/ijrte.d9509.118419

[14] Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, De Cata P, Chiovato L, Bellazzi R. Machine learning methods to predict diabetes complications. Journal of Diabetes Science and Technology 2018;12: 295-302

[15] Harris ML,Oldmeadow C, Hure A, Luu J, Loxton D, Attia J. Stress increases the risk of type 2 diabetes onset in women: A 12-year longitudinal study using causal modelling. PLoS ONE 12(2):e0172126.doi:10.1371/journal. pone.0172126

[16] Kumar P, et.al. Assessment of Anxiety, Depression and Stress using Machine Learning Models: Procedia Computer Science 171 (2020)1989-1998

[17] Olivera AR, Roesler V, Iochpe C, Schmidt MI, Vigo A, Barreto SM, Duncan BB. Comparison of machinelearning algorithms to build a predictive model for detectin

gundiagnosed diabetes - ELSA-Brasil: accuracy study Sao Paulo Med J 2017;135:234-46

[18] Aysh Anand, Divya Shakti. Prediction of diabetes based on personal lifestyle indicators,In :1st international conference on next generation computing technology:IEEE;2015

[19] Maniruzzaman Md, Rahman Md J, Al-MehediHasan Md, Suri HS, Abedin Md M, Ei-Baz A, Suri JS. Accurate diabetes risk stratification using machine learning: Role of missing value and outliers. Journal of Medical Systems 2018; 42: 92 https://doi.org/10.1007/s10916-018-0940-

[20] Kaur H.Kumari V. Predictive Modelling and analytics for diabetes using a machine learning approach, Applied Computing and Informatics 2018

[21] Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders, 19(1), 1–9. https://doi.org/10.1186/s12902-019-0436-6

[22] Schofield J, Leelarathna L, Thabit H. COVID-19: Impact of and on diabetes. Diabetes Ther. 2020;11:1429-1435.

[23] Hussain A, Bhowmik B, do Vale Moreira NC. COVID-19 and diabetes: Knowledge in progress. Duabetes Res Clin Pract. 2020; Apr; 162: 108142.

[24] Yusoff MSB. Psychometric Properties of the Depression Anxiety Stress Scale in a Sample of Medical Degree Applicants. Int Med J. 2013; 20: 295-300.

[25] Nai-aruna N, Moungmaia R. Comparison of classifiers for the risk of diabetes prediction. 7th International Conference on Advances in Information Technology. Procedia Computer Science 2015;69:132-142

[26] Chandrakar O, Saini JR. Development of Indian weighted diabetic risk score (IWDRS) using machine learning techniques for type-2 diabetes. ACM COMPUTE '16, October 21-23, 2016. DOI: http://dx.doi.org/10.1145/2998476.2998497

[27] Deberneh, H.M.; Kim, I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. Int. J. Environ. Res. Public Health 2021, 18, 3317. https://doi.org/10.3390/ ijerph18063317.

[28] Sanchez W., Martinez A.,et.al. A predictive model for stress recognition in desk jobs. Journal of Ambient Intelligence and Humanized Computing.Dec-2018, https://doi.org/10.1007/s12652-018-1149-9

[29] Ahuja R. Banga A.: Mental Stress Detection in University Students using Machine Learning Algorithms. International Conference on Pervasive Computing Advances and Applications (PerCAA 2019), Procedia Computer Science vol.152,pp.349-353.ScienceDirect(2019).

[30] Elena Smets, Pierluigi Casale, Ulf Großekathöfer and Bishal Lamichhane, Walter De Raedt, “Comparison of Machine Learning Techniques for Psycho physiological Stress Detection,” Springer International Publishing Switzerland, pp.13–22, June 2016.

[31] Tigga N.,Garg S.: Prediction of Type 2 Diabetes using Machine Learning Classification Methods Classification, International Conference on Computational Intelligence and Data Science, Procedia Computer Science 2020;.167:706-16.

[32] Hasan MD. K.,Alam MD.A.,Das D.,Hossain E.,Hasan M.: Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers.IEEE Access 8(2020). https://doi.org/10.1109/ACCESS.2020.2989857.
[33] Mahabub A.: A robust voting approach for diabetes prediction using traditional machine learning techniques, SN Applied Sciences,Springer(2019); 1:1667.
[34] Haq A.,Li J.P.,Khan J.,Memon M.H.,Nazir S.,Ahmad S.,Khan G.A.,Ali A.,Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data, Sensors 2020;20: 2649
[35] Jahan N.,Islam A.,Mamun A.A.,: Machine Learning With Factor Scoring To Predict Diabetes Risk Level In Bangladesh., Int J of Scientific & Technology research 2020; 9(2):2863-67.
[36] Lama L.,Wilhelmsson O.,Norlander E.,Gustafsson L.,Lager A.et.al.: Machine learning for prediction of diabetes risk in middle-aged Swedish people.Heliyon 2021;7:1-6.
[37] Kumari S.,Kumar D.,Mittal M.:An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier International Journal of Cognitive Computing in Engineering 2021;2:40-6.
[38] Birjais R.,Mourya A.K.,Chauhan R.,Kaur H. Prediction and diagnosis of future diabetes risk: a machine learning approach,SN Applied Sciences 2019;1:1112.
[39] Ahmeda N. , Ahammeda R. ,Islama, M.,Uddina A.,et.al.,Machine learning based diabetes prediction and development of smart web application, International Journal of Cognitive Computing in Engineering,2021;229-41
[40] Pima Indians Diabetes dataset. Available from: http://archive.ics.uci.edu/ml/machine learning-databases/pima-indians-diabetes/pima-indians-diabetes data. Accessed: 1st of (2008, May).
[41] Kannadasan K, Edla D., Kuppili V.,Type 2 diabetes data classification using stacked autoencoders in deep neural networks ,Clinical Epidemiology and Global Health,(2019);530-35
[42] S. P. Chatrati, G. Hossain, A. Goyal, A. Bhan, S. Bhattacharya, D. Gaurav, and S. M. Tiwari, ‘‘Smart home health monitoring system for predicting type 2 diabetes and hypertension,’’ J. King Saud Univ. Comput. Inf. Sci., Jan. 2020.
[43] Sisodia D. Sisodia D.: Prediction of Diabetes using Classification Algorithms: Procedia Computer Science 132 (2018) 1578–1585
[44] Tanji Y, Sawada S, Watanabe T, Mita T, Kobayashi Y, Murakam T, et al. Impact of COVID-19 pandemic on glycemic control among outpatients with type 2 diabetes in Japan: A hospital-based survey from a country without lockdown. Diabetes research and clinical practice 2021; 176: 108840
[45] Patil R, Gothankar J. Risk factors for type 2 diabetes mellitus: An urban perspective. Indian J Med Sci 2019; 71(1): 16-21

LIST OF PUBLICATIONS

1] Rohini Patil ,Dr.Kamal Shah ,"Assessment of Risk of Type 2 Diabetes Mellitus With Stress as a Risk Factor Using Classification Algorithms" published in International Journal of Recent Technology and Engineering(IJRTE) Vol-8,Issue-4 November 2019.

2] Rohini Patil ,Dr.Kamal Shah ,"Performance Evaluation of different stress factors for Type-2 Diabetes Mellitus using Logistic Regression Model", published in the International conference on Intelligent Computing and Networking (IC-ICN-2020)Feb-2020

3]Rohini Patil ,Dr.Kamal Shah ," Machine learning in healthcare: Applications, current status and future prospectus", Handbook of Research on Machine Learning" published in Apple Academic Press, CRC Press, Taylor & Francis Group (2022) (ISBN-9781774638682)

4] Rohini Patil, Dr Kamal Shah, Dr,Deepak Bhosle,"Impact of COVID-19 related Stress on Glycaemic control in Patients with Type-2 Diabetes Mellitus" accepted in Journal of Health Science and Surveillance System(Shiraz University of Medical Science,Iran) [Scopus Indexed]

5] Rohini Patil, Dr Kamal Shah," Performance Evaluation of Machine Learning Classifiers for Prediction of Type 2 Diabetes using Stress Related Parameters" presented in International Conference on Data Science, Computation and Security (IDSCS 2022) [Springer, LNNS indexed in ISI Proceedings, Scopus Indexed]

6] Rohini Patil, Dr Kamal Shah," Impact of Lifestyle Parameters for Type 2 Diabetes Prediction: A Machine Learning paradigm "published in IJERT journal 2022

Details of IPR

1] Rohini Patil, Dr. Kamal Shah, "TYPE 2 DIABETES PREDICTION USING IMPROVED ENSEMBLE LEARNING DIMENSIONALITY REDUCTION." Indian patent application number 202221005490 dated March 4, 2022.

2] Rohini Patil, Dr. Kamal Shah, "Mathematical model for Diabe_Lifestyle" Diary Number 8673/2022-CO/L. (24th April 2022)

APPENDIX-I

DEOGIRI DIABETES CARE CENTRE

Dr. Deepak S. Bhosle
MBBS, MD, D. Diab.
Consultant Diabetologist
Regd.No. 87375

Dr. Vandana D. Bhosle
MBBS, DPB (Mumbai)
Consultant Pathologist
Regd.No. 2252

FACILITIES :- • DIABETES DIAGNOSIS & TREATMENT • ADVICE ABOUT DIET & EXERCISE • DIABETIC FOOT CLINIC • GESTATIONAL DIABETES MANAGEMENT • INSULIN PUMP CLINIC • CGMS / AGP • PATHOLOGY LABORATORY • OBESITY MANAGEMENT

Date: 07/11/2020

TO WHOMSOEVER IT MAY CONCERN

Mrs. Rohini Anant Patil, pursuing her Ph.D. in information technology at Thakur College of Engineering and Technology, Kandivali, Mumbai is currently working on research project for "Estimation of risk of type 2 diabetes by developing a predictive model".

For the above said research project , she has designed a questionnaire in consultation with her guide and a healthcare professional (Physicians, Diabetologist and Endocrinologist) to collect relevant information from general population and people living with diabetes.

The questionnaire used for this project contains relevant questions covering different risk factors for the development of type 2 diabetes including demographic parameters, diet related factors, family history, lifestyle related factors and stress related factors.

Yours Sincerely

Dr. Deepak S. Bhosle
Deogiri Diabetes Care Centre
Siddi Arcade, Kalad Corner,
Besides HDFC Bank
Aurangabad, Maharashtra

Address : Siddhi Arcade, Office No. 101, 1st Floor, Kalda Corner, Besides HDFC Bank, Shriram Chowk, Aurangabad.

APPENDIX-II

Ethics Committee for Research on Human Subjects
MGM campus, N-6, CIDCO, Aurangabad - 431003
Ph. No.: 0240-6601100, 6601174, Fax No.: +91-0240-2487727
Email: [illegible]
Office: Dept of Pharmacology & Member Secretary MGM-ECRHS,
MGM Medical College, Aurangabad

MGM-ECRHS/2020

Date: 21st Dec, 2020

MGM - ECRHS APPROVAL LETTER

To,
The PI: Mrs. Rohini Patil
Research Scholar,
Thakur College of Engineering & Technology,
Mumbai.

Dear Dr.,
The meeting of Ethics Committee for Research on Human Subjects (ECRHS) was held on 21st Dec., 2020,at 10 AM in the Boardroom with Dr. Manvendra Kachole as Chairperson. At the said meeting, the ECRHS approved the proposal for the clinical study entitled as "Impact of COVID-19 related stress on glycaemic control in patients with type 2 diabetes mellitus: A Retrospective Study". The ECRHS reviewed and approved the following documents submitted for the above mentioned clinical study.

List of Documents:

1. Study Protocol
2. Informed consent form
3. Case Record form
4. Appendices/Questionnaire
5. Adverse event reporting form / Product monograph
6. Current CV of Principal Investigator
7. Principal Investigator's undertaking

It is policy of ECRHS that, it should be informed about any serious adverse event occurring during the course of the study within seven days of the occurrence of the adverse event. The ECRHS expects to be informed about any changes in the protocol and informed Consent Document and asks to be provided a copy of the final report. The status of the study (completed/ongoing/terminated) should be reported to the ECRHS annually. 12 Members attended the meeting held on 21st Dec., 2020 fulfilling the quorum. It is hereby confirmed that neither you nor any of the study team members have participated in the voting/decision making procedures of the committee.

Dr. Ajeet Bobde
Member Secretary
MGM-ECRHS
Department of Pharmacology

Member Secretary
MGM-ECRHS
MGM's Medical College, Aurangabad

1	Age	Gender	Weight in	Height in	Residence (Rural/	Family status	Profession (Ser	Type 2 diabe	On treatm	Current HbA	Current F	Current P	Did COVID	HbA1c % bef	FBG mg/dl	PPG mg/dl be	Did COVID-1	Was there major ch	Was ph	Was tr	Were t.Qua
2	55	MALE	75	164	URBAN	Married	BUSINESS	6 yrs	6 yrs	8.60%	156.6	263.9	Yes	8	145.3	203.9	Yes	yes....Diet reduced	yes	yes	Hospitalised
3	50	MALE	65	175	RURAL	Married	ELECTRICIAN	2	2	14	168	238	Yes	8.6	138	210	Yes	yes....Diet reduced	yes	yes	Hospitalised
4	57	MALE	82	166	URBAN	Married	CLERK	5	5	12.7	196	421	Yes	14	161	386	Yes	yes....Diet reduced	yes	yes	Hospitalised
5	43	FEMALE	96	150	URBAN	Married	FARMER	4	4	12.9	177	312	Yes	12.7	145	337	Yes	yes....Diet reduced	yes	yes	Hospitalised
6	31	FEMALE	68	155	URBAN	Married	HOUSE WIFE	5	5	13	188	300	Yes	12.9	158	270	Yes	yes....Diet reduced	yes	yes	Hospitalised
7	42	MALE	82	178	URBAN	Married	FARMER	4	4	14.1	150	324	Yes	13	220	292	Yes	yes....Diet reduced	yes	yes	Hospitalised
8	45	MALE	93	180	URBAN	Married	ELECTRICIAN	3	3	11.7	133	421	Yes	14.1	194	390	Yes	yes....Diet reduced	yes	yes	Hospitalised
9	50	FEMALE	90	160	URBAN	Married	HOUSE WIFE	2	2	12.6	182	472	Yes	11.7	198	437	Yes	yes....Diet reduced	yes	yes	Hospitalised
10	47	MALE	80	172	URBAN	Married	GOVT.SERVANT	5	5	10.6	160	328	Yes	12.6	143	293	Yes	yes....Diet reduced	yes	yes	Hospitalised
11	57	FEMALE	69	142	RURAL	Married	HOUSE WIFE	6	6	8.3	131	261	Yes	10.6	128	248	Yes	yes....Diet reduced	yes	yes	Hospitalised
12	35	FEMALE	74	147	URBAN	Married	HOUSE WIFE	4	4	11.1	189	403	Yes	8.3	197	368	Yes	yes....Diet reduced	yes	yes	Hospitalised
13	44	MALE	86	165	RURAL	Married	RETIRED	3	3	12	160	361	Yes	11.1	154	346	Yes	yes....Diet reduced	yes	yes	Hospitalised
14	48	MALE	69	175	RURAL	Married	GOVT.SERVANT	2	2	12.2	231	408	Yes	12	190	373	Yes	yes....Diet reduced	yes	yes	Hospitalised
15	53	MALE	91	170	RURAL	Married	FARMER	3	3	14.7	195	386	Yes	12.2	160	346	Yes	yes....Diet reduced	yes	yes	Hospitalised
16	37	MALE	80	162	RURAL	Married	GOVT.SERVANT	3	3	13.6	156	291	Yes	14.7	126	251	Yes	yes....Diet reduced	yes	yes	Hospitalised
17	30	MALE	71	163	RURAL	Married	FARMER	2	2	9.9	179	239	Yes	13.6	156	264	Yes	yes....Diet reduced	yes	yes	Hospitalised
18	42	FEMALE	86	155	RURAL	Married	HOUSE WIFE	5	5	10.4	183	308	Yes	9.9	150	263	Yes	yes....Diet reduced	yes	yes	Hospitalised
19	60	MALE	93	168	RURAL	Married	FARMER	7	7	13.8	230	367	Yes	10.4	200	342	Yes	yes....Diet reduced	yes	yes	Hospitalised
20	41	FEMALE	112	153	RURAL	Married	HOUSE WIFE	6	6	14.2	242	411	Yes	13.8	207	359	Yes	yes....Diet reduced	yes	yes	Hospitalised
21	54	FEMALE	81	150	RURAL	Married	HOUSE WIFE	5	5	10.9	169	363	Yes	14.2	137	313	Yes	yes....Diet reduced	yes	yes	Hospitalised
22	68	MALE	78	169	RURAL	Married	RETIRED	4	4	13.3	190	329	Yes	10.9	159	294	Yes	yes....Diet reduced	yes	yes	Hospitalised
23	47	FEMALE	65	147	RURAL	Married	HOUSE WIFE	3	3	11.4	182	251	Yes	13.9	174	216	Yes	yes....Diet reduced	yes	yes	Hospitalised
24	61	FEMALE	84	156	URBAN	Married	HOUSE WIFE	2	2	10.7	199	368	Yes	11.4	169	333	Yes	yes....Diet reduced	yes	yes	Hospitalised
25	70	FEMALE	72	150	URBAN	Married	HOUSE WIFE	3	3	9.8	240	402	Yes	10.7	210	367	Yes	yes....Diet reduced	yes	yes	Hospitalised
26	36	FEMALE	95	152	RURAL	Married	HOUSE WIFE	2	2	8.6	182	342	Yes	9.8	152	302	Yes	yes....Diet reduced	yes	yes	Hospitalised
27	68	FEMALE	74	147	URBAN	Married	HOUSE WIFE	2	2	12.1	211	267	Yes	8.6	176	234	Yes	yes....Diet reduced	yes	yes	Hospitalised
28	44	FEMALE	56	144	RURAL	Married	HOUSE WIFE	3	3	10.5	196	324	Yes	12.1	161	288	Yes	yes....Diet reduced	yes	yes	Hospitalised
29	53	FEMALE	96	140	RURAL	Married	HOUSE WIFE	6	6	9.4	172	276	Yes	10.5	137	241	Yes	yes....Diet reduced	yes	yes	Hospitalised
30	52	FEMALE	74	150	RURAL	Married	FARMER	5	5	10.4	242	414	Yes	9.4	207	369	Yes	yes....Diet reduced	yes	yes	Hospitalised
31	40	FEMALE	82	155	RURAL	Married	FARMER	4	4	10.4	201	410	Yes	10.4	167	372	Yes	yes....Diet reduced	yes	yes	Hospitalised
32	62	FEMALE	71	160	RURAL	Married	FARMER	7	7	8.9	189	314	Yes	10.4	155	279	Yes	yes....Diet reduced	yes	yes	Hospitalised
33	38	FEMALE	75	142	RURAL	Married	FARMER	6	6	13.6	161	306	Yes	8.9	148	260	Yes	yes....Diet reduced	yes	yes	Hospitalised
34	51	FEMALE	84	147	RURAL	Married	HOUSE WIFE	3	3	12.1	200	332	Yes	13.6	165	296	Yes	yes....Diet reduced	yes	yes	Hospitalised
35	42	FEMALE	101	155	RURAL	Married	HOUSE WIFE	2	2	10.3	270	452	Yes	12.1	200	407	Yes	yes....Diet reduced	yes	yes	Hospitalised
36	55	FEMALE	96	153	RURAL	Married	HOUSE WIFE	5	5	9.2	248	403	Yes	10.3	219	365	Yes	yes....Diet reduced	yes	yes	Hospitalised
37	50	FEMALE	71	150	RURAL	Married	HOUSE WIFE	3	3	10.4	206	390	Yes	9.2	180	351	Yes	yes....Diet reduced	yes	yes	Hospitalised