Predictal – A One-Stop Medical Solution for Early Diagnosis of Multiple Diseases

Submitted in partial fulfillment of the requirements of the

degree of

BACHELOR OF ENGINEERING

IN

COMPUTER ENGINEERING

By

Group No: 50

1902021	**Tushar Budhwani**
1902032	**Esha Datwani**
1902040	**Anushree Dutt**
1902058	**Yukta Jain**

Guide:

DR. ARCHANA B. PATANKAR

(Professor, Department of Computer Engineering, TSEC)

Computer Engineering Department
Thadomal Shahani Engineering College
Bandra(w), Mumbai - 400 050
University of Mumbai

(AY 2022-23)

CERTIFICATE

This is to certify that the project entitled **“Predictal – A One-Stop Medical Solution for Early Diagnosis of Multiple Diseases”** is a bonafide work of

1902021	Tushar Budhwani
1902032	Esha Datwani
1902040	Anushree Dutt
1902058	Yukta Jain

Submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **“BACHELOR OF ENGINEERING”** in **“COMPUTER ENGINEERING”**.

Dr. Archana B. Patankar
Guide

Dr. Tanuja Sarode
Head of Department

Dr.G.T.Thampi
Principal

Project Report Approval for B.E

Project report entitled *“**Predictal – A One-Stop Medical Solution for Early Diagnosis of Multiple Diseases”*** by

1902021	Tushar Budhwani
1902032	Esha Datwani
1902040	Anushree Dutt
1902058	Yukta Jain

is approved for the degree of ***“BACHELOR OF ENGINEERING” in “COMPUTER ENGINEERING”***.

Examiners

1.________________________________

2.________________________________

Date:

Place:

Declaration

We declare that this written submission represents my ideas in my own words and where others 'ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have a adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will because 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.

1) _______________________

Tushar Budhwani, 1902021

2) _______________________

Esha Datwani, 1902032

3) _______________________

Anushree Dutt, 1902040

3) _______________________

Yukta Jain, 1902058

Date:

Abstract

There are multiple techniques in machine learning that can do predictive analysis on a large amount of data in a variety of industries. Predictive analysis in healthcare is a difficult endeavor, but it can eventually assist practitioners in making timely decisions regarding patients' health and treatment based on massive volume of data. In this world, a human being suffers from many diseases. Diseases can have a physical, but also a psychological impact on people. Mainly for four reason, diseases are formed: (i) infection, (ii) deficiency, (iii) heredity and (iv) body organ dysfunction. Diseases like breast cancer, diabetes, and heart-related diseases are causing many deaths globally but most of these deaths are due to lack of timely check-ups of the diseases.

The above problem occurs due to lack of infrastructure and a low ratio of doctors to the population. Supporting the same, according to the World Health Organization (WHO), the ratio of doctors to patients is 1:1000 whereas in India the doctor-to-population ratio is 1:1456 indicating a shortage of doctors. These diseases can be a serious threat to mankind if not diagnosed early. Therefore, early recognition and diagnosis of these diseases can save a lot of lives. To address these life threatening issues, we aim to develop a one-stop medical solution platform for early recognition and diagnosis of various diseases like Breast cancer, Diabetes, Heart diseases, Chronic kidney diseases, Prakinson's disease, liver diseases and prediction of disease from symptoms using the concept of Machine Learning.

TABLE OF CONTENTS

Topic **Page No.**

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Diagnosis is a branch of Artificial Intelligence (AI) focused with developing the algorithms and techniques capable of determining whether a system's behavior is correct. Medical diagnosis identifies the diseases or conditions that explain a person's symptoms and signs. Typically, diagnostic information is gathered from the patient's history and physical examination. It is frequently difficult due to the fact that many indications and symptoms are ambiguous and can only be diagnosed by trained health experts. Therefore, countries that lack enough health professionals for their populations, such as developing countries like Bangladesh and India, face difficulty providing proper diagnostic procedures for their maximum patients. Moreover, diagnosis procedures often require medical tests, which low- income people often find expensive and difficult to afford.

As humans are prone to error, it is not surprising that a patient may be over-diagnosed more often. If a person is over-diagnosed, a problem like unnecessary treatments will arise, harming individuals' health and economic waste. Various factors may influence the misdiagnosis, which includes:

- lack of proper symptoms, which often unnoticeable
- the condition of rare disease
- the disease is omitted mistakenly from the consideration

Machine Learning (ML) is used practically everywhere, from cutting-edge technology (such as mobile phones, computers, and robotics) to health care (i.e., disease diagnosis, safety). ML is gaining traction in various fields, including disease diagnosis in health care. Many researchers and practitioners illustrate the promise of Machine Learning-based Disease Diagnosis (MLBDD), which is inexpensive and time-efficient. Traditional diagnosisprocesses are costly, time-consuming, and often require human intervention. While the individual's ability restricts traditional diagnosis techniques, ML-based systems have no such limitations, and machines do not get exhausted as humans do. As a result, a method to diagnose disease with outnumbered patients' unexpected presence in health care may be developed. The emergence of Machine Learning (ML) algorithms in disease diagnosis domains illustrates the technology's utility in medical fields.

Extending this further, the world today is going through a dynamic patch of technology where the demand of intelligence and accuracy is increasing beyond a person could imagine. Today people are likely addicted to internet but they are not concerned about their physical health. People ignore the small problem and don't visit the hospital which turns into a serious disease in no time. Taking the advantage of this growing technology, we aim to develop a system that will help in early diagnosis of multiple diseases in accordance with symptoms put down by the patients without visiting the hospitals / physicians.

There is a demand to make such a system that will help end users to predict diseaseson the basis of symptoms given in it without visiting hospitals. By doing so, it will decrease the rush at OPD's of hospitals and bring down the workload on medical staff. Not only this, this system will reduce the costly treatment and panic moment at the end stages so that proper medication can be provided at the right time and we can lower down the death rate as well.

The analysis accuracy is increased by using Machine Learning algorithms. Altogether this system will help in easier health management. The intent is to deduce a satisfactory Machine Learning algorithm which is efficient and accurate for the prediction of that particular disease.

1.2 Problem Statement & Objectives

The principle aim of this fourth year group project is to design and develop a single solution/stop/platform which will allow users to check whether they have a chronic disease like heart disease, diabetes, breast cancer, liver disease, chronic kidney disease and Parkinson's disease, at the comfort of their homes with just few clicks and key presses. Basically, developing these prediction engines for each disease will allow users to not visit the doctor unless they have been diagnosed with that disease-for further treatment. The prediction engine requires a large dataset and efficient machine learning algorithms to predict the presence of the disease. Pre-processing the dataset to train the machine learning models, removing redundant, null, or invalid data for optimal performanceof the prediction engine. Doctors rely on common knowledge for treatment. When common knowledge is lacking, studies are summarized after some number of cases have been studied. But this process takes time, whereas if machine learning is used, the patterns can be identified earlier. For using machine learning, a huge amount of data is required. There is very limited amount of data available depending on the disease. Also, the number of samples having no diseases is very high compared to the number of samples having the disease.

This project is about performing a detailed analyses of each disease's data, cleaning and preprocessing that data for normalization, randomization, cleaning, removal of outliers and checking for null or empty values, if any. Further the cleaned data is divided into the test and the train set - which is fit into suitable classifiers like SVM, Random Forest, KNN, XGBoost, AdaBoost, etc. and after hyperparameter tuning to exploit the classifier for better accuracies for prediction on the testing set, the models are compared and best one is chosen to dumped and loaded on the user-end (the GUI).

To summarize, through "Predictal" we aim:

- at providing a single platform to its users for early diagnosis of multiple diseases.
- at providing a user-friendly interface where the user is prompted to enter certain/asked symptoms and our models predicts whether the patient has that concerned disease or not, andfinally
- to deduce a good Machine Learning algorithm which is efficient and accurate for

the prediction of a disease. This prediction is done after running various machine learningalgorithms on the datasets and choosing the best one so that we get the best accuracy.

1.3 Scope

"Predictal – A One-Stop Medical Solution for Early Diagnosis of Multiple Diseases" will permit the end-users to predict whether they have heart disease, diabetes, breast cancer, liver disease, chronic kidney disease and Parkinson's disease in minutes and at the comfort of their homes.

1.3.1 Growth of AI System

Artificial Intelligence is one of the hottest topics today. The revenue for cognitive and artificial intelligence systems is expected to hit $12.5 billion.

1.3.2 Availability of Doctors and Chatbots

Other than disease diagnosis, artificial intelligence can be used to streamline and optimize the clinical process. There is only one doctor for over 1600 patients in India .AI health assistants can help in covering large part of clinical and outpatient services freeing up doctor's time to attend more critical cases. Chatbots like "Your.MD" can assist patients by understanding patients' symptoms and suggest easy-to-understand medical information about their condition. Other assistants like "Ada" integrated with "Amazon Alexa" provides a detailed symptom assessment report and also provides an option to contact a real doctor. Suchassistants make use of Natural Language Processing and Deep Learning to understand the user and generate suggestions.

1.3.3 Internet of Things (IoT), Healthcare and Machine Learning

Increasing use of Internet of Things has promising benefits in healthcare. Dynamically collecting patient data using remote sensors can help in early detection of healthproblems and aid in preventive care.

Chapter 2

Review of Literature

2.1 Domain Explanation

Diagnosis is a branch of Artificial Intelligence (A1) focused with developing algorithms and techniques capable of determining whether a system’s behavior is correct. Medical diagnosis identifies the disease or conditions that explain a person’s symptoms and signs. Typically, diagnostic information is gathered from patient’s history and physical examination [1]. It is frequently difficult due to the fact that many indications and symptoms are ambiguous and can only be diagnosed by trained health experts. Therefore, countries that lack enough health professionals for their populations, such as developing countries like Bangladesh and India, face difficulty providing proper diagnostic procedures for their maximum patients [2]. Moreover, diagnosis procedures often require medical tests, which low-income people often find expensive and difficult to afford.

As humans are prone to error, it is not surprising that a patient may be over-diagnosed more often. If a person is over-diagnosed, a problem like unnecessary treatments will arise, harming individuals’ health and economic waste [3]. According to the National Academics of Science, Engineering, and Medicine report of 2015, the majority of people will encounter at least one diagnostic mistake during their lifespan [4]. Various factors may influence the misdiagnosis, which includes:

- Lack of proper symptoms, which often goes unnoticeable
- The condition of rare disease
- the disease is omitted mistakenly from consideration

2.1.1 Artificial Intelligence

Artificial Intelligence (AI) is intelligence demonstrated by machines, as opposed to the natural intelligence displayed by animals including humans. AI research has been defined as the field of study of intelligent agents, which refers to any system that perceives its environment and takes actions that maximize its chance of achieving its goals.

The term "artificial intelligence" had previously been used to describe machines that mimics and display "human" cognitive skills that are associated with the human mind, such as "learning" and "problem-solving". This definition has since been rejected by major AI researchers who now describe AI in terms of rationality and acting rationally, which does not limit how intelligence can be articulated.

2.1.2 Machine Learning

Machine Learning (ML) is a field of inquiry devoted to understanding and building methods that leverage data to improve performance on some set of tasks. It is seen as a part ofartificial intelligence. Machine Learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without explicitly being programmed to do so.

2.1.3 Machine Learning in Health Care

Machine Learning (ML) is used practically everywhere, from cutting-edge technology (such as mobile phones, computers, and robotics) to health care (i.e., disease diagnosis, safety). ML is gaining traction in various fields, including disease diagnosis in health care. Many researchers and practitioners illustrate the promise of Machine Learning- based Disease Diagnosis (MLDD), which is inexpensive and time efficient [5]. Traditional diagnosis processes are costly, time-consuming, and often require human intervention. Whilethe individual's ability restricts traditional diagnosis techniques, ML-based systems have no such limitations, and machines do not get exhausted as humans do. As a result, a method to diagnose disease with outnumbered patients' unexpected presence in health care may be developed. To create MLBDD systems, health care data like images (i.e., X-ray, MRI) and tabular data (i.e., patients' conditions, age, and gender) are employed [6].

Machine Learning (ML) is a subset of AI that uses data as an input resource [7]. The use of predetermined mathematical functions yields a result (classification or regression) thatis frequently difficult for humans to accomplish. For example, using ML, locating malignant cells in a microscopic image is frequently simpler, which is typically challenging to conduct just by looking at the images. Furthermore, since advances in Deep Learning (a form of machine learning), the most current study shows MLBDD accuracy of above 90% [5]. Alzheimer's disease, Heart failure, Breast cancer, and Pneumonia are just a few of the diseases that may be identified with ML. The emergence of Machine Learning (ML) algorithms in disease diagnosis domains illustrates the technology's utility in medical fields.

Recent breakthroughs in ML difficulties, such as unbalanced data, ML interpretation, and ML ethics in medical domains, are only a few of the many challenging fields to handle in a nutshell [8].

The project uses several methods, algorithms and combinations of the algorithms of Machine Learning for early detection and diagnosis of a disease, given the clinical parameters/symptoms of a patient.

2.2 Review of Existing System

Literature study involves conducting studies on various reach analysis techniques and methods that currently in use. Application requirements and functionalities are also defined prior to its development.

The purpose of this literature review is to study and understand all the new and old literatures - articles, research papers, blogs, etc. to find different and better methods orcombination of methods in Machine Learning to achieve an accurate prediction if the patient has that particular disease or not given their clinical parameters/symptoms, the popular datasets used to train that particular disease diagnosis model, the practical domains where this task has been found useful, which method outperforms the others and finally to describe the technicalities behind different preprocessing techniques, training algorithms and evaluation metrics.

2.2.1 Study of Machine Learning and its Algorithms

Machine Learning (ML) is an approach that analyzes data samples to create main conclusions using mathematical and statistical approaches, allowing machines to learn without programming. Arthur Samuel presented machine learning in games and pattern recognition algorithms to learn from experience in 1959, which was the first time the important advancement was recognized. The core principle of ML is to learn from data in order to forecast or make decisions depending on the assigned task [9]. Thanks to Machine Learning (ML) technology, many time-consuming jobs may now be completed swiftly andwith minimal effort. With the exponential expansion of computer power and data capacity, it is becoming simpler to train data-driven machine learning models to predict outcomes with near-perfect accuracy. Several papers offer various sorts of ML approaches [10, 11].

The ML algorithms are generally classified into three categories such as supervised, unsupervised, and semi-supervised [10]. However, ML algorithms can be divided into severalsubgroups based on different learning approaches, as shown in Fig. 1. Some of the popular ML algorithms include Linear Regression, Logistic Regression, Support Vector Machines (SVM), Random Forest (RF), and Naïve Bayes (NB) [10].

2.2.1.1 Decision Tree

The Decision Tree (DT) algorithm follows divide and conquer rules. In DT models, the attribute may take on various values known as classification trees; leaves indicatedistinct classes, while branches reflect the combination of characteristics that result in those class labels. On the other hand, DT can take continuous variables called regression trees. C4.5and EC4.5 are the two famous and most widely used DT algorithms [12]. DT is used extensively by following reference literature: [13, 14, 15, 16].

2.2.1.2 Support Vector Machine

For classification and regression-related challenges, Support Vector Machine (SVM) is a popular ML approach. SVM was introduced by Vapnik in the late 20th century [17]. Apart from disease diagnosis, SVMs have been extensively employed in various other disciplines, including facial expression recognition, protein fold, distant homology discovery, speech recognition, and text classification. For unlabeled data, supervised ML algorithms are unable to perform. Using a hyperplane to find the clustering among the data, SVM can categorize unlabeled data. However, SVM output is not non-linearly separable. To overcome such problems, selecting appropriate kernel and parameters is two key factors when applying SVM in data analysis [11].

2.2.1.3 K-Nearest Neighbor

K-Nearest Neighbor (KNN) classification is a non-parametric classification technique invented in 1951 by Evelyn Fix and Joseph Hodges. KNN is suitable for classification as well as regression analysis. The outcome of KNN classification is class membership. Voting mechanisms are used to classify the item. Euclidean distance techniques are utilized to determine the distance between two data samples. The projected value in regression analysis is the average of the values of the KNN [18].

2.2.1.4 Naïve Bayes

The Naive Bayes (NB) classifier is a Bayesian-based probabilistic classifier. Based on a given record or data point, it forecasts membership probability for each class. The most probable class is the one having the greatest probability. Instead of predictions, the NB classifier is used to project likelihood [11].

2.2.1.5 Logistic Regression

Logistic regression (LR) is an ML approach that is used to solve classification issues. The LR model has a probabilistic framework, with projected values ranging from 0to 1. Examples of LR-based ML include spam email identification, online fraud transaction detection, and malignant tumor detection. The cost function, often known as the Sigmoid function, is used by LR. The Sigmoid function transforms every real number between 0 and 1 [19].

2.2.1.6 AdaBoost

Yoav Freund and Robert Schapire developed Adaptive Boosting, popularly known as AdaBoost. Adaboost is a classifier that combines multiple weak classifiers into a single classifier. AdaBoost works by giving greater weight to samples that are harder to classify and less weight to those that are already well categorized. It may be used for categorization as well as regression analysis [20].

2.2.2 Study of Heart Disease Prediction using ML

With growing development in the field of medical science alongside machine learning various experiments and researches has been carried out in these recent years releasing the relevant significant papers.

Research paper cited as [21] by Purushottam ,et ,al proposed a paper "Efficient Heart Disease Prediction System" using hill climbing and decision tree algorithms .They used Cleveland dataset and preprocessing of data is performed before using classificationalgorithms. The Knowledge Extraction is done based on Evolutionary Learning (KEEL),an open-source data mining tool that fills the missing values in the data set. A decision treefollows top-down order. For each actual node selected by hill-climbing algorithm a node isselected by a test at each level. The parameters and their values used are confidence. Its minimum confidence value is 0.25. The accuracy of the system is about 86.7%.

Research paper cited as [22] by Apurb R., Milan S., Avi A., Dundigalla R., Poonam G., proposed a paper "Heart Disease Prediction Using Machine Learning" using algorithm like Decision tree, Naïve Bayes, Logistic Regression and Random Forest. The Random Forest Algorithm gives the highest accuracy of 90.16%.

Research paper cited as [23] by Santhana Krishnan. J ,et ,al proposed a paper "Prediction of Heart Disease Using Machine Learning Algorithms" using decision tree andNaive Bayes algorithm for prediction of heart disease. In decision tree algorithm the tree is built using certain conditions which gives True or False decisions. The algorithms like SVM, KNN are results based on vertical or horizontal split conditions depends on dependent variables. But decision tree for a tree like structure having root node, leaves and branches base on the decision made in each of tree Decision tree also help in the understating the importance of the attributes in the dataset. They have also used Cleveland data set. Dataset splits in 70% training and 30% testing by using some methods. This algorithm gives 91% accuracy. The second algorithm is Naive Bayes, which is used for classification. It can handle complicated, nonlinear, dependent data so it is found suitable for heart disease dataset as this dataset is also complicated, dependent and nonlinear in nature. This algorithm gives an 87% accuracy.

Research paper cited as [24] Senthil Kumar Mohan et al, proposed “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” in which their main objective is to improve exactness in cardiovascular problems. The algorithms usedare KNN, LR, SVM, NN to produce an improved exhibition level with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with linearmodel (HRFLM).Study of Diabetes Prediction using ML

Many diabetes prediction algorithms have been proposed by the researchers to accurately predict the types 84 of diabetes as well as diabetes of Pima Indians [25, 26, 27]. These works are briefly discussed as follows. Zolfagri et al. [28] have proposed a methodto diagnose diabetes in females’ populations of Pima Indians using an ensemble of neural network and SVM. Pham et al. [29] have predicted diabetes by a new data mining approach that balances using fitting and generalization. Wu et al. [30] have diagnosed diabetes using the prognosis of fuzzy c-means clustering and SVM. Kumari et al. [25] haveused SVM for classification of diabetes. Dey et al. [26] and Zou et al. [27] have implemented a web-based approach to predict diabetes using ML approaches. However, none of the paper has addressed all the well-known supervised learning algorithms at a glance. Karatsiolis et al. have proposed region-based SVM algorithm for medical diagnosisof Pima Indians [32] [33]. Maniruzamman et al. [31] have performed a comparativeanalysis of diabetes mellitus data using ML techniques. Li et al. [25] have proposed a weight-adjusted approach to diagnose diabetes.

In [28] have proposed an accurate diabetes risk stratification using ML techniques. Sivastava et al. [32] have predicted diabetes using ANN approach. Chen et al. have proposed a hybrid prediction model to diagnose type-2 diabetes using decision trees and k-means [34]. In [35] authors have proposed a fuzzy technique to diagnose diabetes accurately. Many such algorithms have been presented in [25, 26, 27, 28, 29, 30, 31, 32,33, 34 and 35]. However, they have not combined addressed most of the supervised learning algorithms.

2.2.3 Study of Diabetes Prediction using ML

Many diabetes prediction algorithms have been proposed by the researchers to accurately predict the types 84 of diabetes as well as diabetes of Pima Indians [25, 26, 27]. These works are briefly discussed as follows. Zolfagri et al. [28] have proposed a methodto diagnose diabetes in females' populations of Pima Indians using an ensemble of neural network and SVM. Pham et al. [29] have predicted diabetes by a new data mining approach that balances using fitting and generalization. Wu et al. [30] have diagnosed diabetes using the prognosis of fuzzy c-means clustering and SVM. Kumari et al. [25] haveused SVM for classification of diabetes. Dey et al. [26] and Zou et al. [27] have implemented a web-based approach to predict diabetes using ML approaches. However, none of the paper has addressed all the well-known supervised learning algorithms at a glance. Karatsiolis et al. have proposed region-based SVM algorithm for medical diagnosisof Pima Indians [32] [33]. Maniruzamman et al. [31] have performed a comparativeanalysis of diabetes mellitus data using ML techniques. Li et al. [25] have proposed a weight-adjusted approach to diagnose diabetes.

In [28] have proposed an accurate diabetes risk stratification using ML techniques. Sivastava et al. [32] have predicted diabetes using ANN approach. Chen et al. have proposed a hybrid prediction model to diagnose type-2 diabetes using decision trees and k-means [34]. In [35] authors have proposed a fuzzy technique to diagnose diabetes accurately. Many such algorithms have been presented in [25, 26, 27, 28, 29, 30, 31, 32,33, 34 and 35]. However, they have not combined addressed most of the supervised learning algorithms.

2.2.4 Study of Breast Cancer Prediction using ML

A large number of machine learning algorithms are available for prediction and diagnosis of breast cancer. Some of the machine learning algorithm are Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5) and K-Nearest Neighbors (KNN Network) etc. A lot of researcher have realized research in breast cancer by using several dataset such as using SEER dataset, Mammogram images as dataset, Wisconsin Dataset and also dataset from various hospitals. By exploiting these dataset authors extract and select various features and complete their research. These are some significant research. The author Sudarshan Nayak [36], demonstrates the use of various supervised machine learning algorithms in classification of breast cancer from using 3Dimages and find out that SVM is the best based on his overall performance. On the otherside, we find that B.M. Gayathri [37], work on comparative study of Relevance vectormachine which provides Low computational cost while comparing with other machinelearning techniques which are used for breast cancer detection, and explain how RVM isbetter than other machine learning algorithms for diagnosing breast cancer even thevariables are reduced and achieved 97% accuracy.

Hiba Asri [38], demonstrated that Support vector Machine (SVM) proves its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate with an accuracy of 97.13%. in recent works, we find that Youness khoudfi and Mohamed Bahaj [39], similarly proposed a comparison between Machine learning algorithms and they found the SVM is the best classifier with an accuracyof 97.9% compared with K-NN, RF and NB, they are based on Multilayer perception with 5layers and 10 times cross validation using MLP.

The author Latchoumiet TP [40] Found a classification value of 98.4% proposing an optimization weighting of the particle swarm (WPSO) based on the SSVM for the classification. Ahmed Hamza Osman [41] proposed a solution for the diagnosis of Wisconsin breast cancer (WBCD) with a prediction of 99.10% found by the SVM algorithm by combining a clustering algorithm with an efficient probabilistic vector support machine. Our research is focused on assessing such machine learning algorithms andapproaches in order to conclude the best methodology for breast cancer prediction and diagnosis.

2.2.5 Study of Liver Disease Prediction using ML

Several machine learning algorithms are available for prediction and diagnosis of liver disease. Some of the machine learning algorithm are Naïve Bayes, ANN, KNN, SVM etc. A lot of researchers have observed successful results in detecting breast cancer by using the UCI ILPD Dataset which contains 10 variables that are age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. and contains 415 as liver disease patients and 167 as non-liver disease patients.

A Gulia et al. [42] in their proposed work researchers have done classification of the liver patient data using the algorithms like Bayesian Network, Support Vector Machine, J48, Multi-Layer Perceptron and Random Forest. The data from the UCI repository which is afforded by Center of Machine Learning and Intelligent Systems has used. After completion of their three-phase analysis, the Random Forest Algorithm is the best one with an accuracy of 71.87% has been concluded.

Y. Kumar et al. [43] in their proposed work researchers have used Rule-Based Classification Model (RBCM) for the prediction of liver diseases. Without the rule-based classification the efficiency of all the common algorithm decreases was analyzed. In their proposed work 20 rules were used for the classification of liver diseases. The decision tree-based algorithm gives the best performance using rule-based classification and accordingly its accuracy decreases when rule-based is not used.

Vijayarani et al. [44] in their research paper classification algorithms are used for the prediction of liver diseases. Famous algorithms like Naïve Bayes and Support Vector Machine (SVM) are used in the proposed work. The dataset from the UCI repository and it is having fields like Gender, Sgot, ALB, ALP, DB etc have taken. Based upon their present work SVM is best in terms of accuracy and Naïve Bayes is good in terms of execution time. Our research is focused on assessing such machine learning algorithms andapproaches in order to conclude the best methodology liver disease prediction and diagnosis.

2.2.6 Study of Chronic Kidney Disease Prediction using ML

With growing development in the field of medical science alongside machine learning various experiments and researches has been carried out in these recent years releasing the relevant significant papers. Different machine-learning techniques have been used for effective classification of chronic kidney disease from patients' data.

Charleonnan et al. [45] did comparison of the predictive models such as K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), and decision tree (DT) on Indians Chronic Kidney Disease (CKD) dataset in order to select best classifier for predicting chronic kidney disease. They have identified that SVM has the highest classification accuracy of 98.3% and highest sensitivity of 0.99.

Salekin and Stankovic [46] did evaluation of classifiers such as K-NN, RF and ANN on a dataset of 400. Wrapper feature selection were implemented and five features were selected for model construction in the study. The highest classification accuracy is 98% by RF and a RMSE of 0.11. S. Tekale et al. [47] worked on "Prediction of Chronic Kidney Disease Using Machine Learning Algorithm" with a dataset consists of 400 instances and 14 features. They have used decision tree and support vector machine. The dataset has been preprocessed and the number of features has been reduced from 25 to 14. SVM is stated as a better model with an accuracy of 96.75%.

Xiao et al. [48] proposed prediction of chronic kidney disease progression using logistic regression, Elastic Net, lasso regression, ridge regression, support vector machine, random forest, XGBoost, neural network and k-nearest neighbor and compared the models based on their performance. They have used 551 patients' history data with proteinuria with 18 features and classified the outcome as mild, moderate, severe. They have concluded that Logistic regression performed better with AUC of 0.873, sensitivity and specificity of 0.83 and 0.82, respectively.

Mohammed and Beshah [49] conducted their research on developing a self-learning knowledge-based system for diagnosis and treatment of the first three stages of chronic kidney have been conducted using machine learning. A small number of data have been used in this research and they have developed prototype which enables the patient to query KBS to see the delivery of advice. They used decision tree in order to generate the rules. The overall performance of the prototype has been stated as 91% accurate.

2.2.7 Study of Parkinson’s Disease Prediction using ML

Shrihari K Kulkarni, K R Sumana, the researchers in [50] used Decision Tree, Logistic Regression, and Naive Bayes, Deep Learning algorithm like Recurrent Neural Networks (RNN) by predicting the Performance Parameters to build the model. Machine learning approaches will be used to construct prediction models that can differentiate early PD from healthy normal using the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDSUPDRS). For Subject and Record Validation, Logistic Regression, Random Forests, and Support Vector Machine were employed. Drawback of this paper were, Data Collection techniques are weakly regulated, resulting in unreliable results such as out-of-range or non-existent This Model purely relies on Evaluation of Motions which is not the only source of Data available on the Disease-bearers or Healthy Citizens.

Yatharth Nakul, Ankit Gupta, Hritik Sachdeva, the researchers in [51] used Supervised Learning Algorithms such as Random Forest, Support Vector and Naïve-Bayes are also compared. Confusion matrix was used for accuracy checking and different Classification methods were used.ML classification technique will improve the accuracy and reduce possible loopholes. Hyper parameter tuning is used to achieve the maximum accuracy. Achieved maximum accuracy of 98.30% using the K nearest neighbor classification The main drawbacks are Delay in Results derived and Output Progression is slow and Best Proposed Methodology used gives Higher error rate when Confusion Matrix is plotted.

SGD (Stochastic Gradient Descent) is utilized for training data models, according to Wu Wang, Junho Lee, Fouzi Harrou, and Ying Sun of [52]. The FNN (Feed-Forward Neural Network) is put into action. The sensitivity of the linear discriminate analysis approach utilized is the best, which means it has the best likelihood of distinguishing a real patient. The proposed deep learning model had a 96.45% accuracy rate. This is owing to the deep learning model's favorable capabilities in learning linear and nonlinear features from PD data without the requirement for handcrafted feature extraction.

2.2.8 Study of Symptoms to Disease Prediction using ML

There are numerous machine learning algorithms available for prediction and diagnosis of various diseases from the symptoms entered by the users. Some of the machine learning algorithm are Support Vector Machine (SVM), Random Forest Classifier, Naïve Bayes, Decision Tree Classifier, Multi-Layer Perceptron (MLP) Classifier etc. Many researchers have carried out extensive researches in this field, releasing the relevant important papers.

Research paper cited as [53] by Talasila Bhanuteja ,et ,al proposed a paper "Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach" , characterizes the illnesses by utilizing different calculations like Random Forest, LightGBM and Decision Tree. The paper shows the research conducted by a robotized framework that can find and separate secret information related with the infections from a historical (diseases-side effects) data set by the standard set of the individual Algorithms and models. The dataset used comprises of 132 indications, the blend or stages of which leads to 41 illnesses. In light of the 4920 documents of various patient samples, mainly to point foster a forecast algorithm that considers in the side effects of various client and forecasts the sickness that the person is bound to be affected.

The proposed machine learning model had a 96.45% accuracy rate using Random Forest algorithm.

Another research paper cited as [55] by Rinkal Keniya, et ,al proposed a paper "Disease prediction from various symptoms using machine learning" has used different ML models to examine the prediction of disease for available input dataset. Out of the 11 models, they have manages to get 50% or above accuracy for 6 models. The highest accuracy was gained for the weighted KNN model of 93.5%.

2.3 Limitations of Existing Systems/ Research Gaps

The medical domain every day gathers an enormous amount of information about patients including clinical tests, imperative boundaries, examination reports, therapy subsequent meet-ups, and medicine choices and so forth. Yet, sadly they aren't examined and mined in a suitable manner. There are a few limitations in the existing systems. They are as follows:

Existing systems primarily focus on detecting only one disease through the inputs entered by the user. This requires the users to know about the disease they are getting tested for. A prior knowledge about the disease along with the constraint that only one disease prediction could happen is not an optimal solution. Predictal provides an easy and user-friendly application, which gives users option for predicting up to seven diseases. By simply entering the values of the asked data in the input fields, the user can get the intensity of the risk of a disease predicted.

Along with entering the values of specific parameters to get the prediction of certain disease, Predictal also provides a symptoms to disease prediction feature where users can simply select symptoms and get the result, which is not available in existing systems.

According to the many research papers referred, not many algorithms have been used to train the existing models. Predictal extensively uses multiple algortihms, after doing a series of rigorous preprocessing, exploratory data analysis and feature selection. This provides better accuracy results over other existing systems. Machine learning algorithms such as Random Forest, Decision Tree, ET classifier, Multi-layer perceptron classifier and many more such algorithms are trained using the datasets, compared with each other based on various performance evaluation metrics and the model with best result is chosen for a particular disease.

Chapter 3

Proposed System

3.1 Analysis/Framework

Machine learning (ML) is an approach that analyzes data samples to create conclusions using mathematical and statistical approaches, allowing machines to learn without programming. Arthur Samuel presented machine learning in games and pattern recognition algorithms to learn from experience in 1959, which was the first time the important advancement was recognized. The core principal of machine learning is to learn from data in order to forecast or make decisions depending on the assigned task [56]. Thanks to machine learning (ML) technology, many time-consuming jobs may be completed swiftly and with minimal efforts. With the exponential expansion of computer power and data capacity, it is becoming simpler to train data-driven ML models to predict outcomes with near-perfect accuracy.

Machine learning (ML) is now being used practically everywhere, from mobile phones, computers and robotics to heath care in disease diagnosis and safety, etc. One such field out of the various other fields gaining popularity would be disease diagnosis in healthcare. ML-based diagnosis systems have no limitations of a traditional diagnosis process like being costly, time-consuming and often requiring human intervention, instead the machine-learning-based disease diagnosis applications neither get exhausted as humans do nor are expensive for generating results in no time. To create such MLBDD (Machine-Learning-Based Disease Diagnosis) systems, heath care data such as images (i.e., X-ray, MRI) and tabular data (i.e. patients' conditions, age and gender) are employed.

Predictal being a multiple MLBDD (Machine-Learning-Based Disease Diagnosis) system, employs all the steps employed in the building of a general machine learning application as described by Yufeng Guo in an online article [57]. These general steps have been applied repeatedly for all the 7 sub-systems that Predictal integrates - the diabetes, heart disease, breast cancer, chronic kidney disease, liver disease, parkinsons and the symptoms to disease prediction system. The *Fig. 3.1* depicts these general seven machine leaning steps in a visual form.

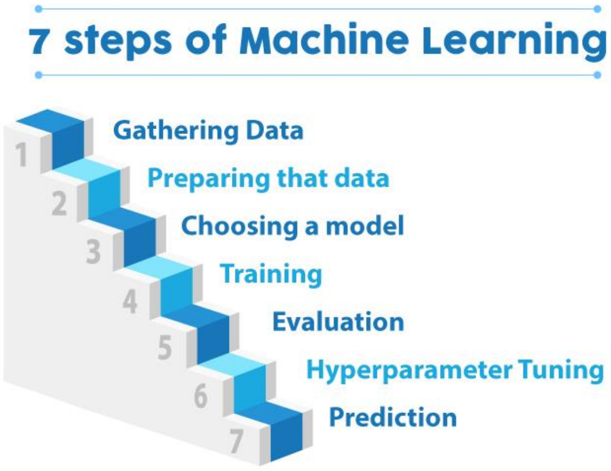


Fig. 3.1 The 7 steps of Machine Learning

These general steps are as follows:

3.1.1 Data Collection

Machines initially learn from the data that is given to them. It is of utmost importance to collect reliable data so that the machine learning model can find the correct patterns. The quality of the data that is fed to the machine will determine how accurate the model is. If the data is incorrect or outdated, the outcomes or predictions will be wrong or non-relevant. Hence, it is important to make sure that the data being used comes from a reliable source, as it will directly affect the outcome of the model. A good data is one which is relevant, contains very few missing and repeated values, and has a good representation of the various subcategories/classes present. The important pointers in the step of "Data Collection" that have been kept in mind while development are as follows:

- The quantity & quality of the data dictates how accurate the model will be.
- The outcome of this step is generally a representation of data which will be used for training.
- Using pre-collected data, by way of datasets from Kaggle, UCI, etc., still fits into this step.

3.1.2 Data Preparation

After the data has been collected and is available in hand, it has to be prepared into a format suitable for further analysis. This is generally carried out by putting together all the data and randomizing it. This helps make sure that the data is evenly distributed, and the ordering does not affect the learning process. Cleaning the data is part of this step which involves removing unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc. The dataset might also require restructuring and changing of rows, columns or their indexes. Visualizing the data could help get a better understanding of how structured the data is and in understanding the relationship between various attributes and the classes present. The final step of the preparation step would be splitting the cleaned data in two sets – a training set and a testing set. The training set is the set the model learns from and a testing set is used to check the accuracy of the model after training. The important sub-steps in the step of "Data Preparation" that have been followed while the development of each sub-system (individual disease diagnosis systems) are as follows:

- Wrangle data and prepare it for training.
- Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)
- Randomize data, which erases the effects of the particular order in which we

collected and/or otherwise prepared our data.

- Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis.
- Split into training and evaluation sets.

3.1.3 Choosing a Model

A machine learning model determines the output one gets after running a machine learning algorithm on the collected data. It is therefore important to choose a model which is relevant to the task at hand. Over the years, scientists and engineers developed various models suited for different tasks like speech recognition, image recognition, prediction, etc. Apart from this, one also has to see if the model is suited for numerical or categorical data and choose accordingly.

The important pointer applied in the step of "Choosing a Model" while the development of each sub-system are as follows:

- Different algorithms are for different tasks; to choose the right and appropriate one.

3.1.4 Training the Model

Training is the considered as the most important step in machine learning. In training, the prepared data is passed to the machine learning model to find patterns and make predictions. It results in the model learning from the data so that it can accomplish the task set. Over time, with training, the model gets better at predicting.

The important pointers in the step of "Training the Model" that have been kept in mind while the development of each sub-system are as follows:

- The goal of training is to answer a question or make a prediction correctly as often as possible.
- Linear regression example: algorithm would need to learn values for *m* (or *W*) and *b* (*x* is input, *y* is output).
- Each iteration of process is a training step.

3.1.5 Evaluating the Model

After model has been trained, the next step is to check how it's performing. This is done by testing the performance of the model on previously unseen data. The unseen data used is the testing set that data split generated earlier. If testing was done on the same data which is used for training, an accurate measure will not be achieved, as the model is already used to the data, and finds the same patterns in it, as it previously did. This will result into disproportionately high accuracy. When used on testing data, an accurate measure of how the model will perform and its speed can be achieved.

The important pointers in the step of "Evaluating the Model" that have been kept in mind and followed while the development of each sub-system are as follows:

- Uses some metric or combination of metrics to "measure" objective performance of model.
- Test the model against previously unseen data.
- The unseen data is meant to be somewhat representative of model performance in the real world, but still helps tune the model (as opposed to test data, which does not).
- Good train/evaluation split? 80/20, 70/30, or similar, depending on domain, data availability, dataset particulars, etc.

3.1.6 Parameter Tuning

Once the model has been created and evaluated, the next step is to see if the model's accuracy can be improved in any way. This is achieved by tuning the parameters present in the model. Parameters are the variables in the model that the programmer generally decides. At a particular value of a model's parameter, the accuracy will be the maximum. Parameter tuning refers to finding these values.

The important pointers in the step of "Parameter Tuning" that have been kept in mind and applied while the development of each sub-system are as follows:

- This step refers to hyperparameter tuning, which is an "artform" as opposed to a science.
- Tune model parameters for improved performance.
- Simple model hyperparameters may include: number of training steps, learning rate, initialization values and distribution, etc.

3.1.7 Making Predictions

In the end, the model is used on unseen data to make predictions accurately. The important pointers in the step of "Making Predictions" that have been kept in mind and applied while the development of each sub-system are as follows:

- Using further (test set) data which have, until this point, been withheld from the model (and for which class labels are known), are used to test the model to get a better approximation of how the model will perform in the real world.

3.2 Design Details

In a multiple disease prediction system, it is possible to predict more than one disease at a time. So the user doesn't need to traverse different sites in order to predict the diseases. Many of the existing models are concentrating on one disease per analysis like one analysis for diabetes analysis, one for cancer analysis, one for skin disease, etc. There is no common system present that can analyze more than one disease at a time. Hence, through *Predictal* we aim to provide a single interface with the option of diagnosing more than one diseases. Not only that, *Predictal* has been developed to diagnose or predict diseases from the various symptoms inputted by its users.

To implement the same, machine learning algorithms and the Streamlit python library has been used. Python pickling is used to save the behavior of the model and then is later loaded on to the main interface to give the final prediction or result to its users.

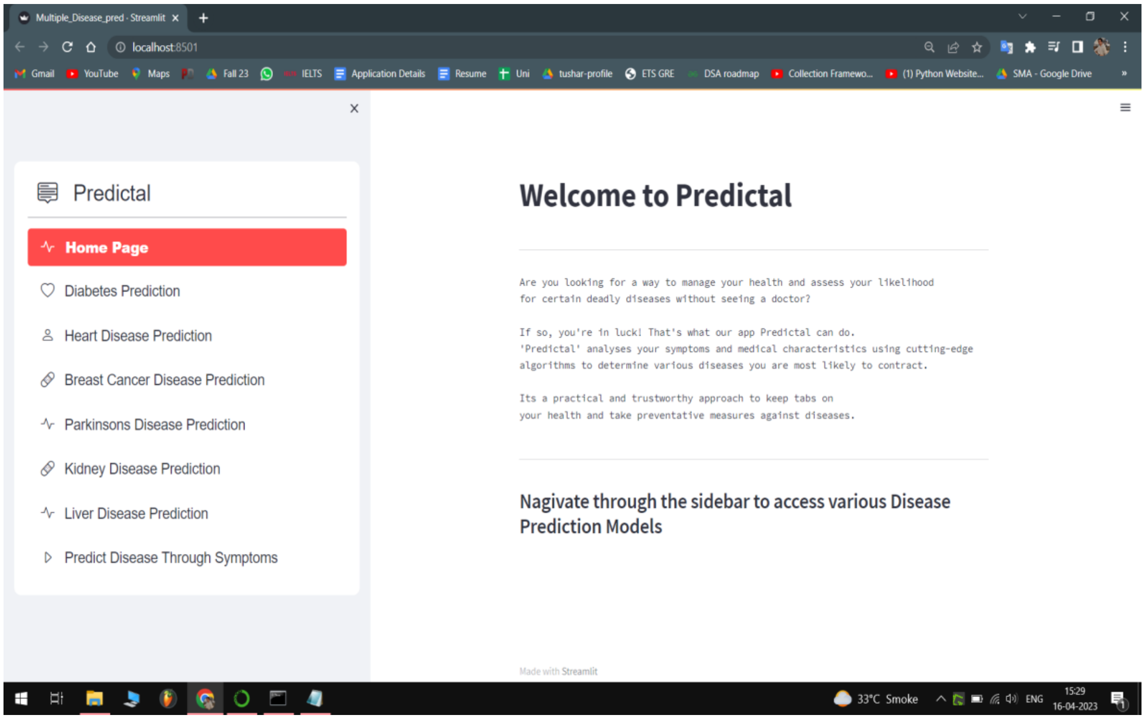
The Graphical User Interface for the developed multiple disease prediction system has been discussed in the following sub-section.

3.2.1 GUI Design

Predictal makes use of the Python's Streamlit library to flaunt a simple and easy-to-use Graphical User Interface (GUI). The aim of this GUI design has been to make it easier and fun for the users to interact with the platform. It has been provided with short, easy-to-understand instructions like a user manual in order to make a user's usage journey easy and fruitful.

3.2.1.1 Web Application's Home Page

The Fig. 3.2 shows the web application's home page, the page where the user is first brought to when it visits the web page. As shown in the figure, the home page has been designed to hold to a small interactive introduction about the web app and instructions for an easy navigation through the app.



Welcome to Predictal

Are you looking for a way to manage your health and assess your likelihood
for certain deadly diseases without seeing a doctor?

If so, you're in luck! That's what our app Predictal can do.
'Predictal' analyses your symptoms and medical characteristics using cutting-edge
algorithms to determine various diseases you are most likely to contract.

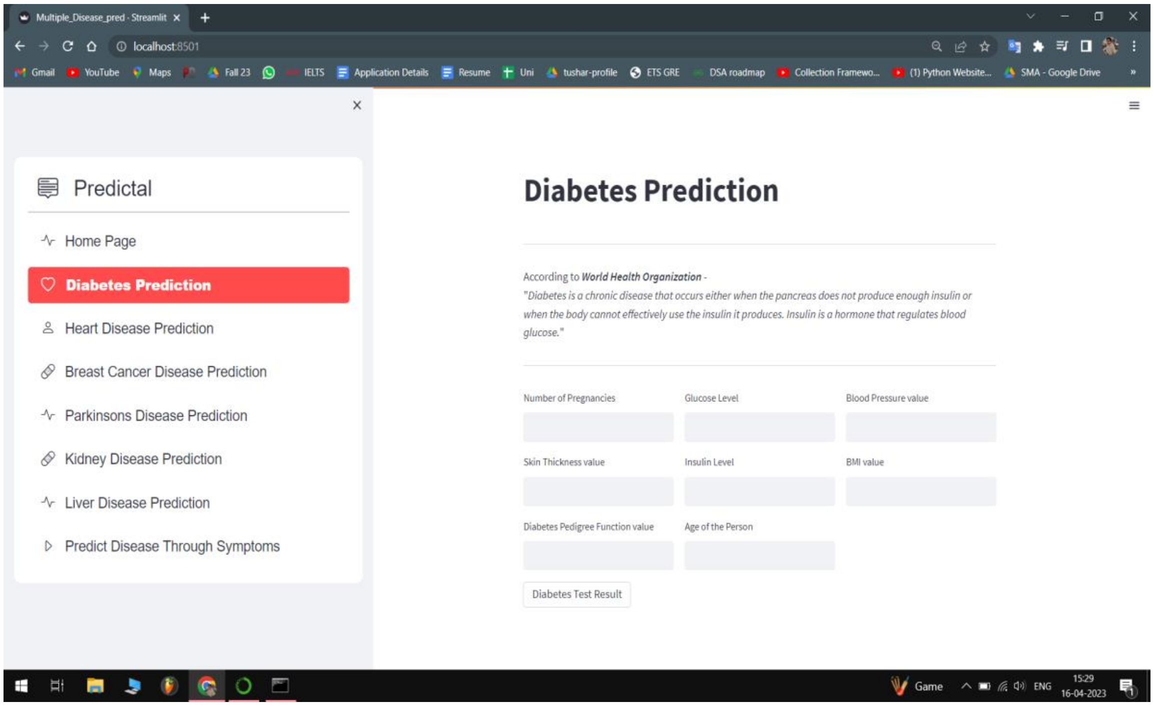
Its a practical and trustworthy approach to keep tabs on
your health and take preventative measures against diseases.

Nagivate through the sidebar to access various Disease Prediction Models

Fig. 3.2 Predictal's Home Page

3.2.1.2 Diabetes Prediction Page

The *Fig. 3.3* shows the web application’s diabetes page. The user design consists of a menu list to the left of the page which consists of a list of diseases *Predictal* is currently capable of providing results related to. The very first disease in this menu is Diabetes. On landing to the diabetes prediction page the user can see a small introduction about the disease and a number of input fields and labels on top of them. These input fields are capable of inputting only float values. These float values are then fed to the backend which generates the final result. These values are fed to the model and the result is generated when clicked on the “Diabetes Test Result” button at the very bottom of the page.



Diabetes Prediction

According to ***World Health Organization*** -

"Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood glucose."

Number of Pregnancies

Glucose Level

Blood Pressure value

Skin Thickness value

Insulin Level

BMI value

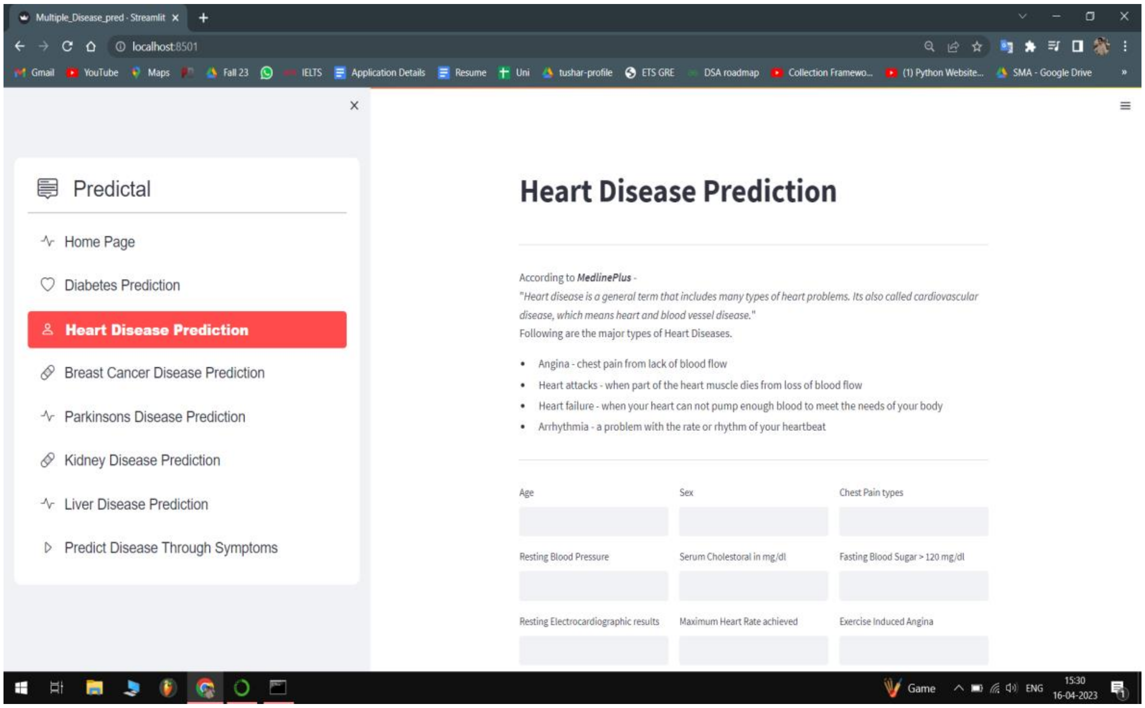
Diabetes Pedigree Function value

Age of the Person

Fig. 3.3 Predictal’s Diabetes Prediction Page

3.2.1.3 Heart Disease Prediction Page

The *Fig. 3.4* shows the web application's heart disease page. The user design consists of a menu list to the left of the page which consists of a list of diseases *Predictal* is currently capable of providing results related to. The second disease in this menu is the heart disease. On landing to the heart disease prediction page the user can see a small introduction about the disease and a number of input fields and labels on top of them. These input fields are capable of inputting only float values. These float values are then fed to the backend which generates the final result. These values are fed to the model and the result is generated when clicked on the "Heart Disease Test Result" button at the very bottom of the page.



Heart Disease Prediction

According to ***MedlinePlus*** -
"Heart disease is a general term that includes many types of heart problems. Its also called cardiovascular disease, which means heart and blood vessel disease."
Following are the major types of Heart Diseases.

- Angina - chest pain from lack of blood flow
- Heart attacks - when part of the heart muscle dies from loss of blood flow
- Heart failure - when your heart can not pump enough blood to meet the needs of your body
- Arrhythmia - a problem with the rate or rhythm of your heartbeat

Age

Sex

Chest Pain types

Resting Blood Pressure

Serum Cholestoral in mg/dl

Fasting Blood Sugar > 120 mg/dl

Resting Electrocardiographic results

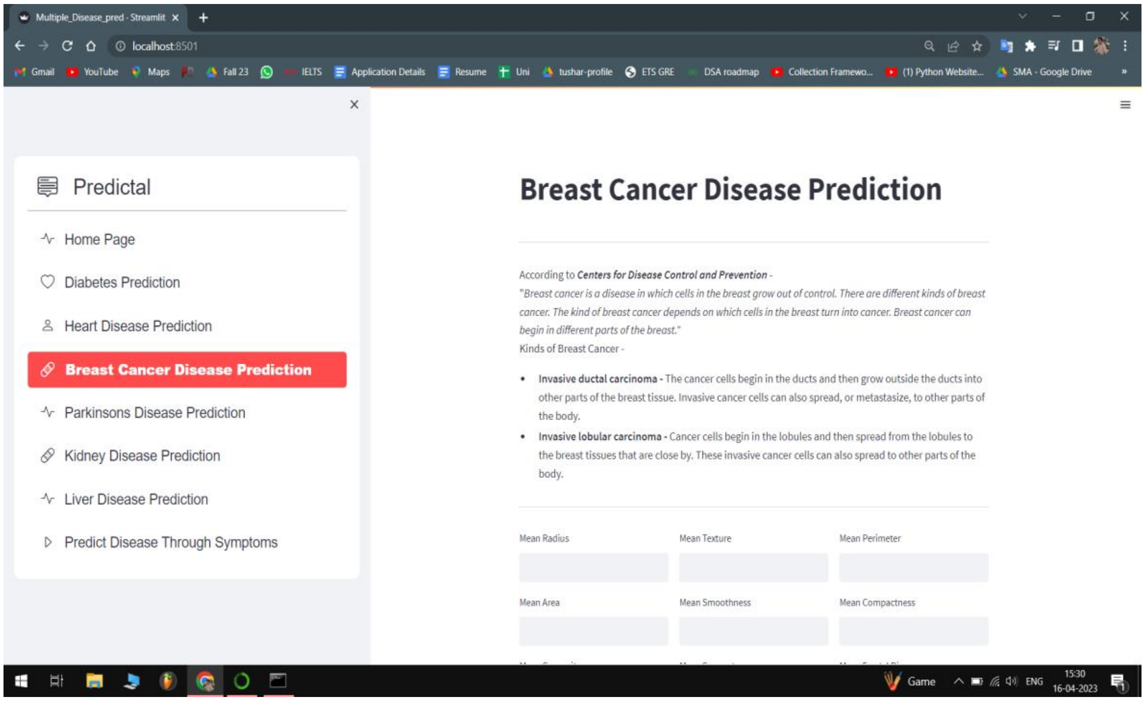
Maximum Heart Rate achieved

Exercise Induced Angina

Fig. 3.4 Predictal's Heart Disease Prediction Page

3.2.1.4 Breast Cancer Prediction Page

The *Fig. 3.4* shows the web application’s breast cancer disease prediction page. The user design consists of a menu list to the left of the page which consists of a list of diseases *Predictal* is currently capable of providing results related to. The third disease in this menu is Breast Cancer. On landing to the breast cancer disease’s prediction page the user can see a small introduction/information about the disease and a number of input fields and labels on top of them. These input fields are capable of inputting only float values. These float values are then fed to the backend which generates the final result. These values are fed to the model and the result is generated when clicked on the “Breast Cancer Test Result” button at the very bottom of the page.



Breast Cancer Disease Prediction

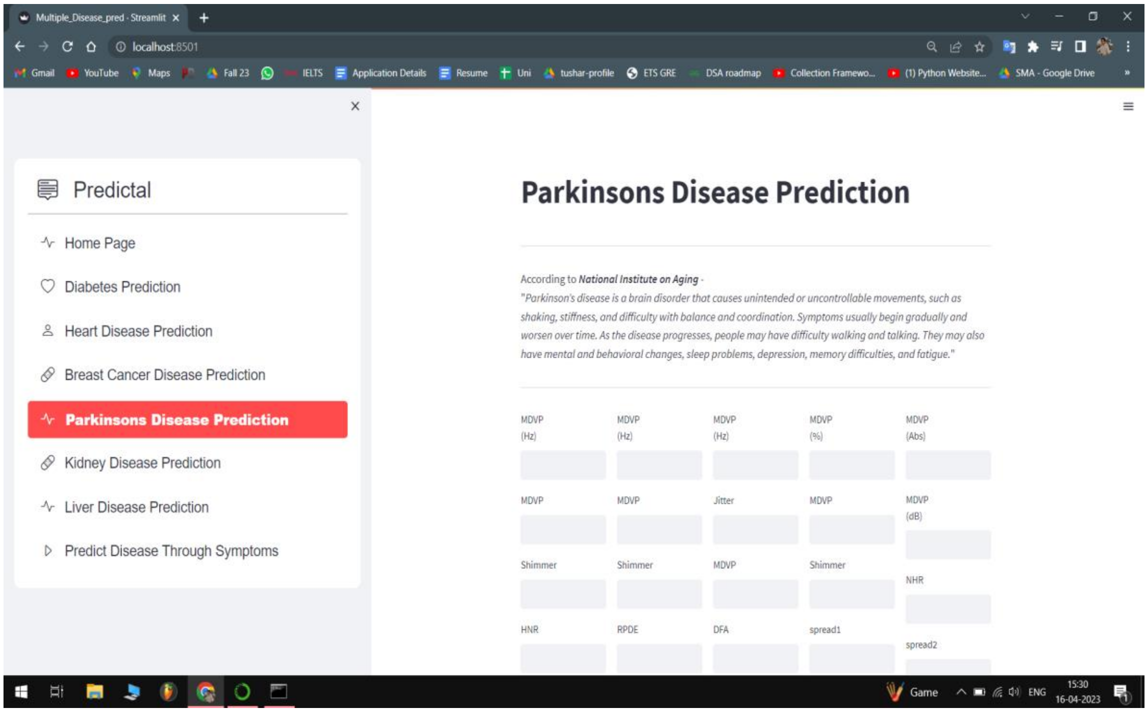
According to ***Centers for Disease Control and Prevention*** -
"Breast cancer is a disease in which cells in the breast grow out of control. There are different kinds of breast cancer. The kind of breast cancer depends on which cells in the breast turn into cancer. Breast cancer can begin in different parts of the breast."
Kinds of Breast Cancer -

- **Invasive ductal carcinoma -** The cancer cells begin in the ducts and then grow outside the ducts into other parts of the breast tissue. Invasive cancer cells can also spread, or metastasize, to other parts of the body.
- **Invasive lobular carcinoma -** Cancer cells begin in the lobules and then spread from the lobules to the breast tissues that are close by. These invasive cancer cells can also spread to other parts of the body.

Fig. 3.5 Predictal’s Breast Cancer Disease Page

3.2.1.5 Parkinson’s Disease Prediction Page

The *Fig. 3.6* shows the web application’s Parkinson’s disease prediction page. The user design consists of a menu list to the left of the page which consists of a list of diseases *Predictal* is currently capable of providing results related to. The fourth disease in this menu is the Parkinson’s disease. On landing to its prediction page the user can see a small introduction about the disease and a number of input fields and labels on top of them. These input fields are capable of inputting only float values. These float values are then fed to the backend which generates the final result. These values are fed to the model and the result is generated when clicked on the “Parkinson’s Disease Test Result” button at the very bottom of the page.



Parkinsons Disease Prediction

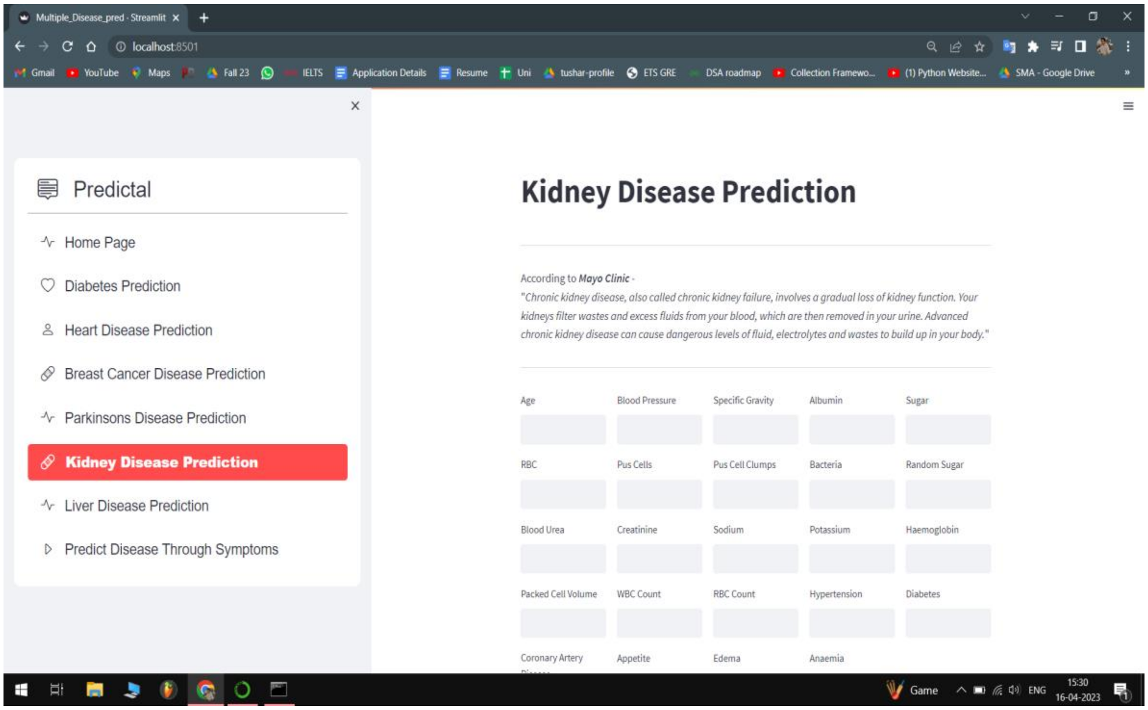
According to ***National Institute on Aging*** -

"Parkinson's disease is a brain disorder that causes unintended or uncontrollable movements, such as shaking, stiffness, and difficulty with balance and coordination. Symptoms usually begin gradually and worsen over time. As the disease progresses, people may have difficulty walking and talking. They may also have mental and behavioral changes, sleep problems, depression, memory difficulties, and fatigue."

Fig. 3.6 Predictal’s Parkinson’s Disease Prediction Page

3.2.1.6 Chronic Kidney Disease Prediction Page

The *Fig. 3.7* shows the web application’s chronic kidney disease prediction page. The user design consists of a menu list to the left of the page which consists of a list of diseases *Predictal* is currently capable of providing results related to. The fifth disease in this menu is the Chronic Kidney disease. On landing to its prediction page the user can see a small introduction about the disease and a number of input fields and labels on top of them. These input fields are capable of inputting only float values. These float values are then fed to the backend which generates the final result. These values are fed to the model and the result is generated when clicked on the “Kidney Disease Test Result” button at the very bottom of the page.



Kidney Disease Prediction

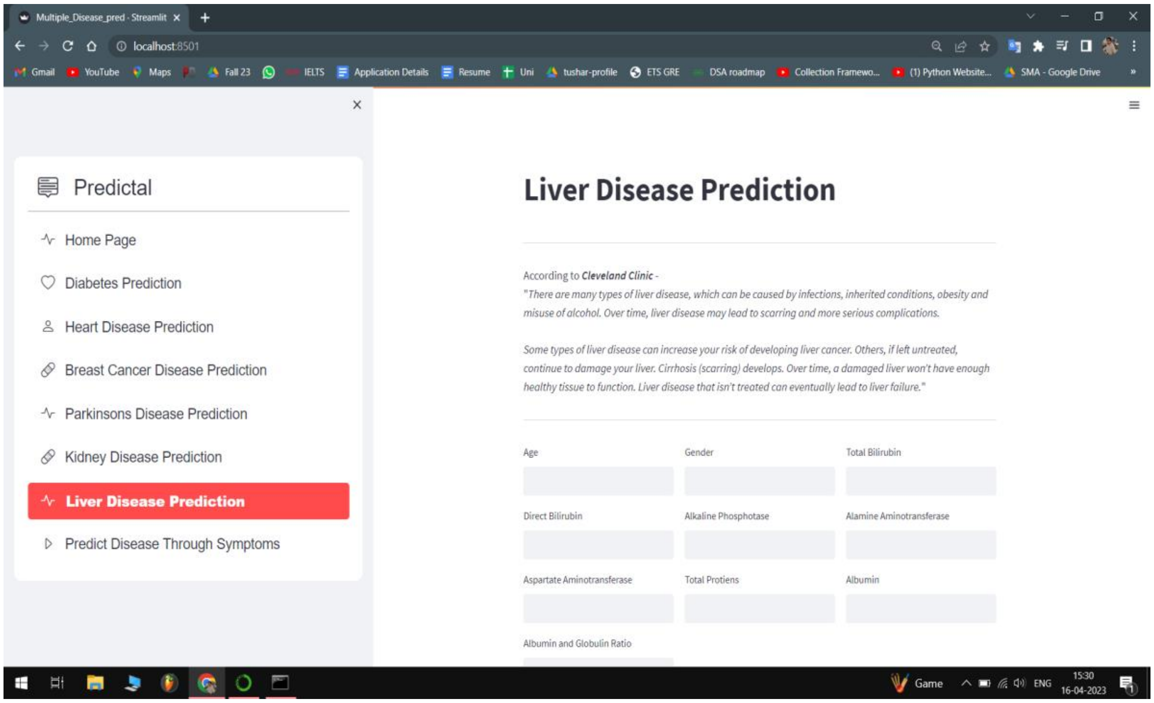
According to ***Mayo Clinic*** -
"Chronic kidney disease, also called chronic kidney failure, involves a gradual loss of kidney function. Your kidneys filter wastes and excess fluids from your blood, which are then removed in your urine. Advanced chronic kidney disease can cause dangerous levels of fluid, electrolytes and wastes to build up in your body."

Age
Blood Pressure
Specific Gravity
Albumin
Sugar
RBC
Pus Cells
Pus Cell Clumps
Bacteria
Random Sugar
Blood Urea
Creatinine
Sodium
Potassium
Haemoglobin
Packed Cell Volume
WBC Count
RBC Count
Hypertension
Diabetes
Coronary Artery Disease
Appetite
Edema
Anaemia

Fig. 3.7 Predictal’s Kidney Disease Prediction Page

3.2.1.7 Liver Disease Prediction Page

The *Fig. 3.8* shows the web application's liver disease prediction page. The user design consists of a menu list to the left of the page which consists of a list of diseases *Predictal* is currently capable of providing results related to. The sixth disease in this menu is the liver disease. On landing to its prediction page the user can see a small introduction about the disease and a number of input fields and labels on top of them. These input fields are capable of inputting only float values. These float values are then fed to the backend which generates the final result. These values are fed to the model and the result is generated when clicked on the "Liver Disease Test Result" button at the very bottom of the page.


Multiple_Disease_pred · Streamlit
localhost:8501
Predictal
Home Page
Diabetes Prediction
Heart Disease Prediction
Breast Cancer Disease Prediction
Parkinsons Disease Prediction
Kidney Disease Prediction
Liver Disease Prediction
Predict Disease Through Symptoms
Game
ENG
15:30
16-04-2023

Liver Disease Prediction

According to ***Cleveland Clinic*** -
"There are many types of liver disease, which can be caused by infections, inherited conditions, obesity and misuse of alcohol. Over time, liver disease may lead to scarring and more serious complications.

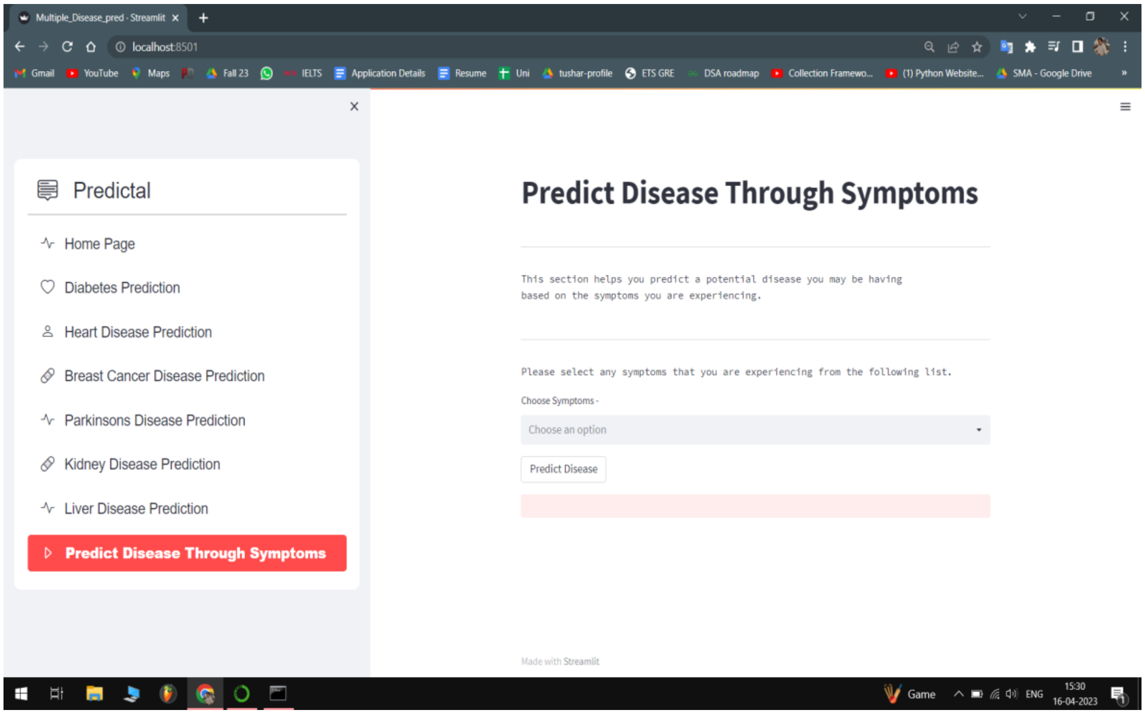
Some types of liver disease can increase your risk of developing liver cancer. Others, if left untreated, continue to damage your liver. Cirrhosis (scarring) develops. Over time, a damaged liver won't have enough healthy tissue to function. Liver disease that isn't treated can eventually lead to liver failure."

Age
Gender
Total Bilirubin
Direct Bilirubin
Alkaline Phosphotase
Alamine Aminotransferase
Aspartate Aminotransferase
Total Protiens
Albumin
Albumin and Globulin Ratio

Fig. 3.8 Predictal's Liver Disease Prediction Page

3.2.1.8 Predict Disease through Symptoms Page

The *Fig. 3.9* shows the web application's predict disease through symptoms page. The user design consists of a menu list to the left of the page which consists of a list of diseases *Predictal* is currently capable of providing results related to. The seventh sub-system and a component that makes it very different from the other existing multiple disease prediction systems is this symptoms to disease prediction. On landing to this page the user can see a small introduction about the section and a dropdown with a label on top of it. The dropdown consists of a total of 132 symptoms, providing users a wide range of options to select from. The selection field takes multiple selection of 3 or more than 3 only. These inputted symptoms are then fed to the backend which generates the final result. The final result is a disease that the model thinks the symptoms entered are related to. These values are fed to the model and the result is generated when clicked on the "Predict" button at the very bottom of the page.



Predict Disease Through Symptoms

This section helps you predict a potential disease you may be having based on the symptoms you are experiencing.

Please select any symptoms that you are experiencing from the following list.

Choose Symptoms -

Choose an option

Multiple_Disease_pred · Streamlit
localhost:8501
Predictal
Home Page
Diabetes Prediction
Heart Disease Prediction
Breast Cancer Disease Prediction
Parkinsons Disease Prediction
Kidney Disease Prediction
Liver Disease Prediction
Predict Disease Through Symptoms
Predict Disease
Made with Streamlit
15:30
16-04-2023

Fig. 3.9 Predictal's Predict Disease through Symptoms Page

3.2.2 Architecture Design

Predictal being a multiple disease detection system contains 7 (seven) sub-systems namely, the heart disease prediction system, diabetes prediction system, breast cancer prediction system, liver failure/disease prediction system, the chronic kidney disease prediction system, Parkinson’s disease prediction system and finally the symptoms to disease prediction system. Now, each of these sub-systems go through the same process flow from data collection to final predictions. Depending on the dataset collected the pre-processing methods differ based on how raw the data is. The data after being pre-processed is split into training and testing data. The training data as the name suggests is used to train the model on. After the model’s learning stage the model is now ready to show its performance by testing it on the test dataset. This test dataset was kept unseen i.e. it was made sure that the test dataset is not used in the model’s learning process. The performance of the model is then evaluated using different performance metrics. The classifier that performs the best and gives the best performance metric results is chosen to be saved. These saved model weights are finally loaded onto the main interface to provide the most accurate and trustable final results to its users. The *Fig. 3.10* below summarizes all this information in the form of an architectural/block diagram.

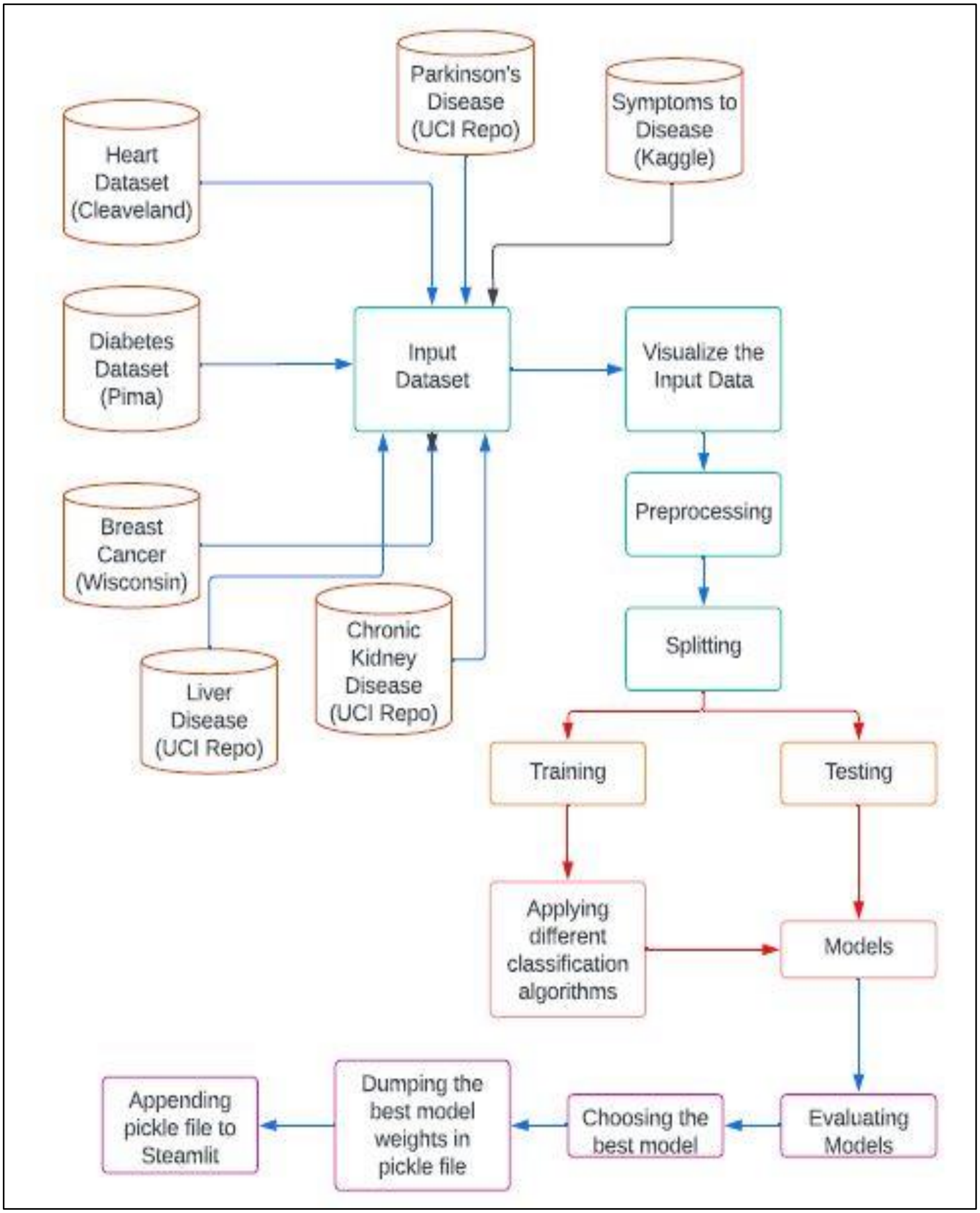

Parkinson's Disease (UCI Repo)
Symptoms to Disease (Kaggle)
Heart Dataset (Cleaveland)
Diabetes Dataset (Pima)
Input Dataset
Visualize the Input Data
Preprocessing
Breast Cancer (Wisconsin)
Splitting
Chronic Kidney Disease (UCI Repo)
Liver Disease (UCI Repo)
Training
Testing
Applying different classification algorithms
Models
Appending pickle file to Steamlit
Dumping the best model weights in pickle file
Choosing the best model
Evaluating Models

Fig. 3.10 Block Diagram

3.3 Methodology

There are different types of research methods either in social science, management, medical, engineering among others. In all the fields of study, there are various research methods available and understanding of these methods will assist an individual to choose the right research methodology in his or her research exercise. There are many ways to categorize different types of research. The words used to describe the research depends on the discipline and field. Generally, the form ones research (types of research methods) approach takes will be shaped by the followings:

- The type of knowledge one aims to produce
- The type of data one will collect and analyze
- The sampling methods, timescale and location of the research.
- Types of research methods

For our multiple disease prediction system - "Predictal", we have considered experimental research, where we first understand or define our problem statement, collect suitable data, prepare it into a format suitable for further analysis, analyze it to look for hidden patterns, trends or relationships, perform data modeling by selecting an appropriated algorithm, splitting the data into training and testing data, building a model on the training data, evaluating it on the testing set or validation set and selecting the model with the best accuracy to be integrated in the main system to generate final predictions on the user data. This pre-processing and the algorithms applied on each sub-system or each disease detection part of Predictal will be elaborated upon disease wise in this section.

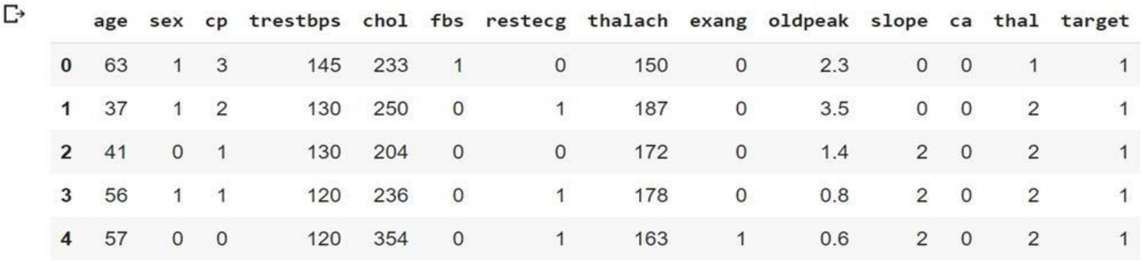
3.3.1 Heart Disease Prediction

Heart diseases have been the primary reason for death all over the world. Majority of the deaths related to cardiovascular problems are caused by heart attacks and strokes. The World Health Organization (WHO) indicates that an approximate 17.9 million people die due to such diseases every year. Therefore, it is essential to find methods to ensure the minimization of these numbers. In order to minimize the detrimental effects of heart diseases, we must try to predict its presence at earlier stages. ML algorithms can help us effectively predict such results with a high degree of accuracy which can in turn help doctors and patients detect the onset of such diseases and reduce their impact or prevent them from occurring. Our objective was to integrate a sub-system that is able to accurately determine the presence of heart disease in a time and cost-efficient manner. This objective has been successfully achieved by applying the following data pre-processing techniques and classification algorithms on the chosen dataset:

3.3.1.1 Data Pre-processing

The data that is to be processed was not clean and contained outliers, noise and missing values. This data if was not processed further would not have given the best results hence the process of cleaning the data collected was carried out and the unnecessary values from it were eliminated or filled according to the remaining values in the dataset. The processed data was further simplified from one format to another to make the model understand it even more by applying scaling on it (the process of transformation) after which that dataset values were brought down to the range of -3 to 3 so as to increase uniformity in the dataset. Next, since working on a complex data would be difficult, time-consuming and resource extensive hence the process of feature selection was carried out using the information gain and chi-squared test in which 13 features were selected out of 14 based on their importance. This process was carried out carefully to make sure that no such feature was eliminated that would have led to a huge loss of information.

The *Fig. 3.11* displays how the heart disease dataset looks like after the pre-processing techniques were applied on it.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig. 3.11 Snippet of Pre – processed Heart Disease Dataset

3.3.1.2 Algorithms and Techniques used

There have been a total of 7 classification algorithms used to model the pre-processed data on. The model accuracies have been compared and the one providing the best accuracy on the testing dataset or the previously unseen dataset was chosen to predict the final result on the user input. The algorithms and techniques used are as follows:

3.3.1.2.1 Logistic Regression

The Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts theoutput of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas logistic regression is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an "S"shaped logistic function called the Sigmoid function, which predicts two maximum values (0 or 1). The *Fig. 3.12* shows the curve for the Sigmoid function.

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

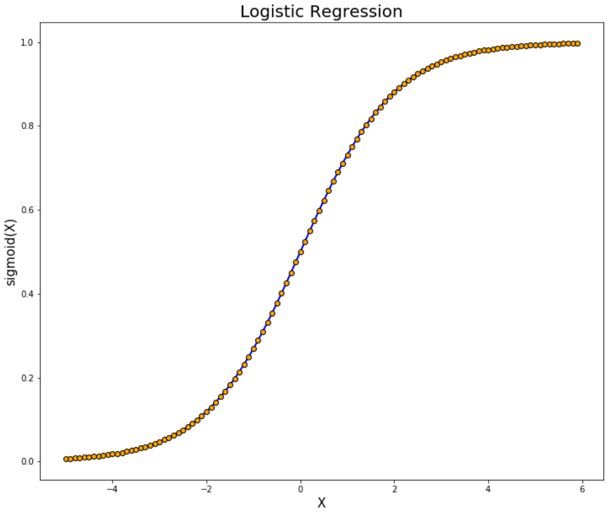


Fig. 3.12 Sigmoid Function

Advantages:

- Logistic Regression is one of the simplest machine learning algorithms and is easy to implement yet provides great training efficiency in some cases. Also due to these reasons, training a model with this algorithm doesn't require high computation power.
- The predicted parameters (trained weights) give inference about the importance of each feature. The direction of association i.e. positive or negative is also given. So we can use Logistic Regression to find out the relationship between the features.
- This algorithm allows models to be updated easily to reflect new data, unlike Decision Tree or Support Vector Machine. The update can be done using stochastic gradient descent.

3.3.1.2.2 Decision Tree

Decision Tree is a supervised learning technique that can be used for both classification and regression problems, but mostly it is preferred for solving classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision Tree, there are two nodes, which are the Decision Node and Leaf Node.

Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a Decision Tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to builda tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A Decision Tree simply asks a question, and based on the answer (Yes/No), it further split the tree into sub trees.

The Decision Tree Algorithm belongs to the family of supervised machine learning algorithms. It can be used for both a classification problem as well as for a regression problem.

The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in whichthe leaf node corresponds to a class label and attributes are represented on the internal node ofthe tree.

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision Tree:

- Decision Trees usually mimic human thinking ability while making a decision, so itis easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree- like structure.

In Decision Tree the major challenge is to identify the attribute for the root node in each level. This process is known as attribute selection. We have two popular attribute selection measures information gain and gini index. The *Fig 3.13* shows how a decision tree looks with its constituent components.

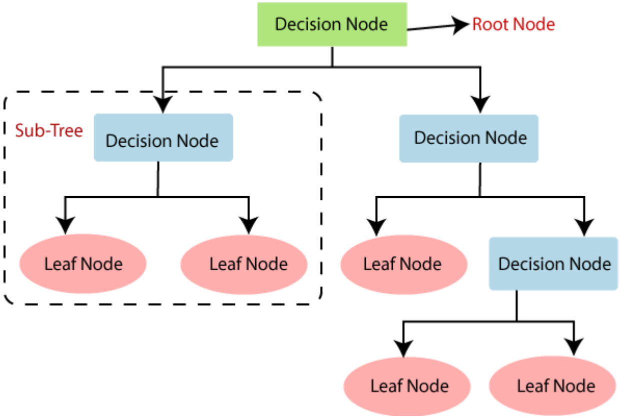


Fig. 3.13 Decision Tree

3.3.1.2.3 Random Forest

Random Forest is a supervised learning algorithm. It is an extension of machine learning classifiers which include the bagging to improve the performance of Decision Tree.It combines tree predictors, and trees are dependent on a random vector which is independently sampled. The distribution of all trees are the same. Random Forests splits nodes using the best among of a predictor subset that are randomly chosen from the node itself, instead of splitting nodes based on the variables. The time complexity of the worst case of learning with Random Forests is O(M(dnlogn)) , where M is the number of growing trees,n is the number of instances, and d is the data dimension.

It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest consists of trees. It is said that the more trees it has, the more robust a forest is. Random Forests create Decision Trees on randomly selected data samples, get predictions from each tree and select the best solution by means of voting. It also providesa pretty good indicator of the feature importance.

Random Forests have a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

Random Forest is a popular machine learning algorithm that belongs to the supervisedlearning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset.

The *Fig. 3.14* displays the working of a general random forest algorithm.

Assumptions:

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

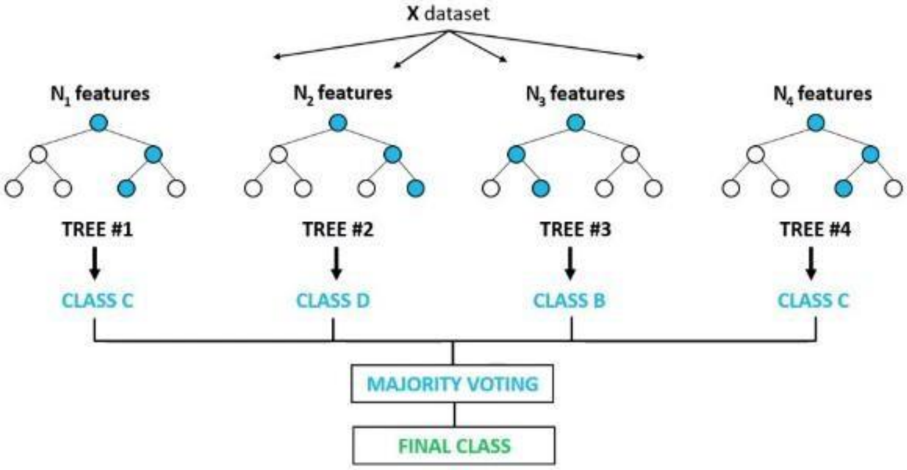


Fig. 3.14 Random Forest Classifier

Advantages:

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

Disadvantages:

- Although Random Forest can be used for both classification and regression tasks, it is notmore suitable for Regression tasks.

3.3.1.2.4 XG-Boost

XG-boost is an implementation of Gradient Boosted decision trees. It is a type of Software library that was designed basically to improve speed and model performance. In thisalgorithm, decision trees are created in sequential form. Weights play an important role in XG-boost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. Weight of variables predicted wrong by the tree is increased and these the variables are then fed to the second decision tree. These individual classifiers/predictors then assemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined predict. The *Fig. 3.15* depicts how the XG-Boost algorithm works.

1. **Regularization:** XG-boost has in-built L1 (Lasso Regression) and L2 (Ridge Regression) regularization which prevents the model from overfitting. That is why, XG-boost is also called regularized form of GBM (Gradient Boosting Machine). While using Scikit Learn libarary, we pass two hyper-parameters (alpha and lambda) to XG-boost related to regularization. alpha is used for L1 regularization and lambda is used for L2 regularization.
2. **Parallel Processing:** XG-boost utilizes the power of parallel processing and that is why it is much faster than GBM. It uses multiple CPU cores to execute the model. While using Scikit Learn libarary, nthread hyper-parameter is used

for parallelprocessing. nthread represents number of CPU cores to be used. If you want to use all the available cores, don't mention any value for nthread and the algorithm will detect automatically.

3. **Handling Missing Values:** XG-boost has an in-built capability to handle missing values. When XG-boost encounters a missing value at a node, it tries both the left and right hand split and learns the way leading to higher loss for each node. It then does the same when working on the testing data.
4. **Cross Validation:** XG-boost allows user to run a cross-validation at each iteration of the boosting process and thus it is easy to get the exact optimum number of boosting iterations in a single run. This is unlike GBM where we have to run a grid- search and only a limited values can be tested.
5. **Effective Tree Pruning:** A GBM would stop splitting a node when it encounters a negative loss in the split. Thus it is more of a greedy algorithm. XG-boost on the other hand make splits upto the max_depth specified and then start pruning the tree backwards and remove splits beyond which there is no positive gain.

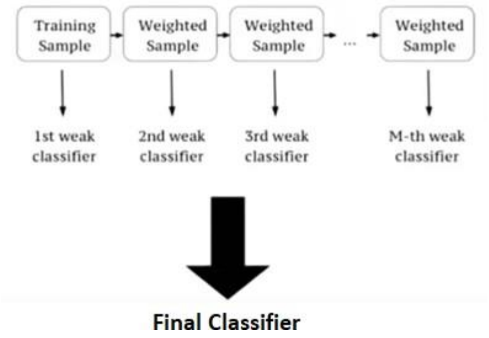


Fig. 3.15 XG-Boost

3.3.1.2.5 Ada-Boost

Adaboost was the first really successful boosting algorithm developed for the purpose of binary classification. Adaboost is short for Adaptive Boosting and is a very popular boosting technique which combines multiple “weak classifiers” into a single “strong classifier”

Algorithm:

1. Initially, Adaboost selects a training subset randomly.
2. It iteratively trains the Adaboost machine learning model by selecting the training setbased on the accurate prediction of the last training.
3. It assigns the higher weight to wrong classified observations so that in the next iteration these observations.
4. Also, it assigns the weight to the trained classifier in each iteration according to the accuracy of the classifier. The more accurate classifier will get high weight.
5. This process iterates until the complete training data fits without any error or until reached to the specified maximum number of estimators.
6. To classify, perform a "vote" across all of the learning algorithms you built Boosting is a system of machine learning boosting, representing a decision tree for large and complex data. It relies on the presumption that the next possible model will minimizethe gross prediction error if combined with the previous set of models. The decision trees are used for the best possible predictions

Advantages:

Adaboost has many advantages due to its ease of use and less parameter tweaking when compared with the SVM algorithms. Plus Adaboost can be used with SVM though theoretically, overfitting is not a feature of Adaboost applications, perhaps because the parameters are not optimized jointly and the learning process is slowed due to estimation stage-wise. This link is useful to understand mathematics. The flexible Adaboost can also be used for accuracy improvement of weak classifiers and cases in image/text classification.

Disadvantages:

Adaboost uses a progressively learning boosting technique. Hence high-quality data is needed in examples of Adaboost vs Random Forest. It is also very sensitive to outliers and noise in data requiring the elimination of these factors before using the data. It is also much slower than the XG-boost algorithm.

3.3.1.2.6 Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. In the 1960s, SVMs were first introduced but later they got refined in 1990. SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.

The followings are important concepts in SVM -

Support Vectors - Data Points that are closest to the hyperplane are called support vectors. Separating line will be defined with the help of these data points.

Hyperplane - As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.

Margin - It may be defined as the gap between two lines on the closest data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin. is an implementation of Gradient Boosted decision trees. It is a type of Software library that was designed basically to improve speed and model performance. In thisalgorithm, decision trees are created in sequential form. Weights play an important role in XG-boost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. Weight of variables predicted wrong by the tree is increased and these the variables are then fed to the second decision tree. The *Fig. 3.16* shows how the SVM algorithm works on a linearly separable 2d data.

Types of SVM:

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier. The predictions from each tree must have very low correlations.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N - the number of features) that distinctly classifies the data points.

Advantages:

- Effective in high dimensional spaces.
- Still effective in cases where the number of dimensions is greater than the number ofsamples.
- Uses a subset of training points in the decision function (called support vectors), so itis also memory efficient.
- Versatile: different kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

Disadvantages:

- If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

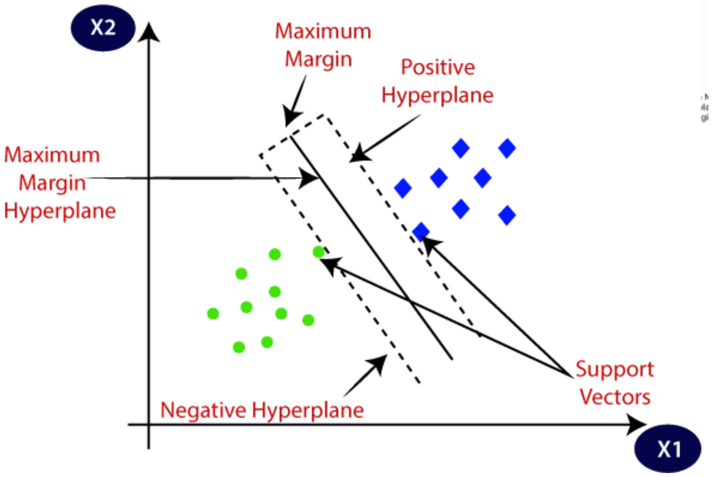


Fig. 3.16 Support Vector Machine

3.3.1.2.7 Stacking Ensemble Technique

Stacking is a way to ensemble multiple classifications or regression model.Stacking (sometimes called Stacked Generalization) is a different paradigm. The point of stacking is to explore a space of different models for the same problem. The idea is that you can attack a learning problem with different types of models which are capable to learn some part of the problem, but not the whole space of the problem. So, you can build multiple different learnersand you use them to build an intermediate prediction, one prediction for each learned model. Then you add a new model which learns from the intermediate predictions the same target.

This final model is said to be stacked on the top of the others, hence the name. Thus, you might improve your overall performance, and often you end up with a model which is better than any individual intermediate model. Notice however, that it does not give you any guarantee, as is often the case with any machine learning technique. The *Fig. 3.17* depicts how the stacking ensemble technique works.

How stacking works?

1. We split the training data into K-folds just like K-fold cross-validation.
2. A base model is fitted on the K-1 parts and predictions are made for Kth part.
3. We do for each part of the training data.

4. The base model is then fitted on the whole train data set to calculate its performanceon the test set.
5. We repeat the last 3 steps for other base models.
6. Predictions from the train set are used as features for the second level model.
7. Second level model is used to make a prediction on the test set.

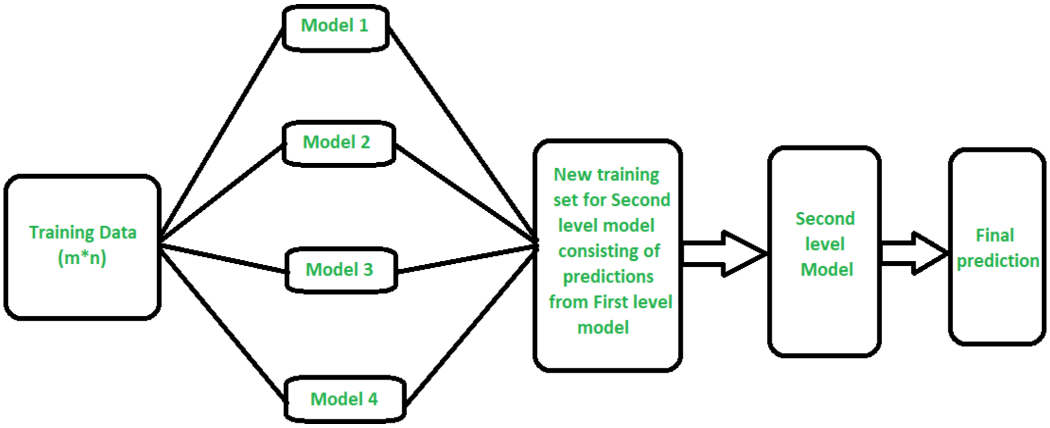

Model 1
Model 2
Training Data (m*n)
Model 3
Model 4
New training set for Second level model consisting of predictions from First level model
Second level Model
Final prediction

Fig. 3.17 Stacking Ensemble Technique

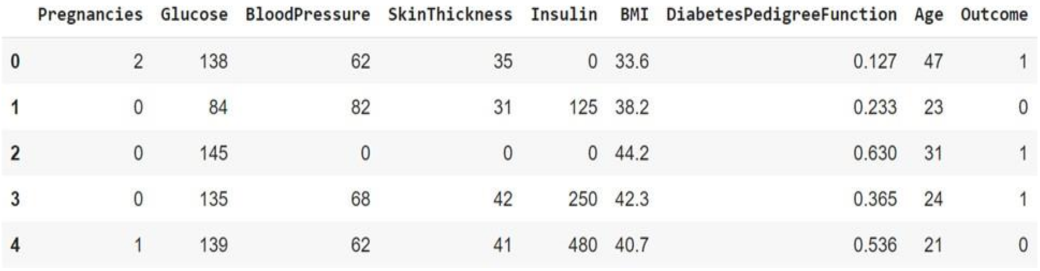
3.3.2 Diabetes Prediction

Diabetes is a condition that is brought on by having abnormally high level of blood glucose in the body. Human bodies are in constant need of power, and sugar is one of the primary sources of vitality that is used in the construction of our muscles and other tissues. In individuals, the primary reasons of type 2 diabetes are often an unhealthy habit combined with a lack of physical activity. Diabetes is a condition that is brought on by an abnormally high level of glucose in the bloodstream. Diabetes occurs when the pancreas is failing to turn the meal into insulin; as a result, sugar is not taken into the body, leading to the condition. Diabetes may cause problems in a variety of body systems, including the kidneys, eyes, neurological system, arteries, and so on. In order to minimize these detrimental effects of diabetes, we must try to predict its presence at earlier stages. ML algorithms can help us effectively predict such results with a high degree of accuracy which can in turn help doctors and patients detect the onset of such diseases and reduce their impact or prevent them from occurring. Our objective was to integrate a sub-system that is able to accurately determine the presence of diabetes in a time and cost-efficient manner. This objective has been successfully achieved by applying the following data pre-processing techniques and classification algorithms on the chosen dataset:

3.3.2.1 Data Pre-processing

The data that is to be processed was not clean and contained outliers, noise and missing values. This data if was not processed further would not have given the best results hence the process of cleaning the data collected was carried out and the unnecessary values from it were eliminated or filled according to the remaining values in the dataset. The outliers were handled using the Winsorisation technique that considers a value as an outlier if it lies beyond the lower and the upper bound. The processed data found no need for further simplified data therefore scaling was not applied on it (the process of transformation). Next, since working on a complex data would be difficult, time-consuming and resource extensive hence the process of feature selection was carried out by visualizing the data as a heatmap to detect multicollinearity, if any. This process was carried out carefully to make sure that highly correlated features do not affect the final results.

The *Fig. 3.18* displays how the heart disease dataset looks like after the pre-processing techniques were applied on it.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0

Fig. 3.18 Snippet of Pre – processed Diabetes Disease Dataset

3.3.2.2 Algorithms and Techniques used

There have been a total of 5 supervised learning algorithms used to model the pre-processed data on. The model accuracies have been compared and the one providing the best accuracy on the testing dataset or the previously unseen dataset was chosen to predict the final result on the user input. The algorithms and techniques used are as follows:

3.3.2.2.1 Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms. It is a classification technique that works on a probability basis. It is much similar to Linear Regression which gives continuous value. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.1*.

3.3.2.2.2 XG-Boost

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.4.*

3.3.2.2.3 Random Forest

Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.3.*

3.3.2.2.4 Support Vector Machine

SVM is a powerful supervised algorithm that works best on smaller datasets but on complex ones. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks, but generally, they work best in classification problems. This learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.6.*

3.3.2.2.5 K-Nearest Neighbor

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based onSupervised Learning technique. It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suitecategory by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. It is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. This algorithmat the training phase just

stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. The *Fig. 3.19* shows the pictorial representation of before and after the application of KNN algorithm situation on a 2d dataset.

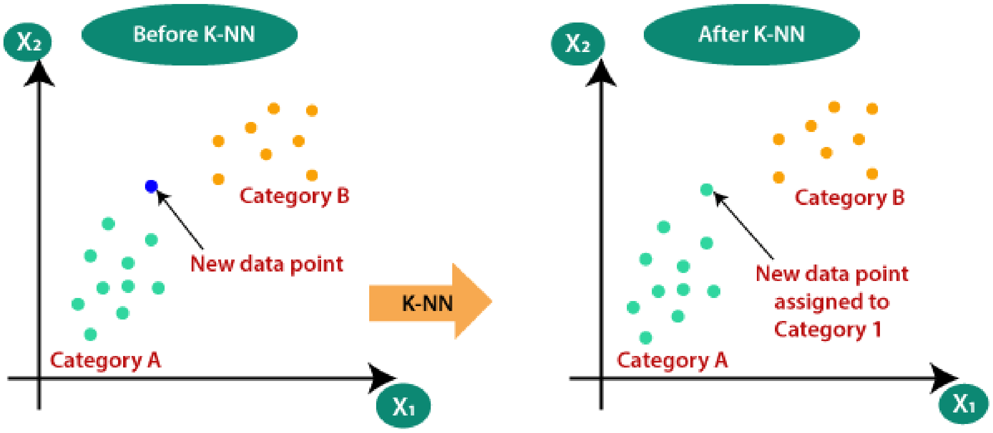


Fig. 3.19 K-Nearest Neighbor Classifier

3.3.3 Breast Cancer Prediction

The second major cause of women's death is breast cancer (after lung cancer) [57] 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of women's death is estimated [58]. Breast cancer represents about 12% of all new cancer cases and 25% of all cancers in women [59]. It is the most common type of all cancers and the main cause of women's deaths worldwide. Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions. In order to minimize these detrimental effects of breast cancer, we must try to predict its presence at earlier stages. ML algorithms can help us effectively predict such results with a high degree of accuracy which can in turn help doctors and patients detect the onset of such diseases and reduce their impact or prevent them from occurring. Our objective was to integrate a sub-system that is able to accurately determine the presence of breast cancer in a time and cost-efficient manner. This objective has been successfully achieved by applying the following data pre-processing techniques and classification algorithms on the chosen dataset:

3.3.3.1 Data Pre-processing

The data was first looked at to get an overview of the attributes in the dataset. An in-depth analysis was carried out next to find the hidden patterns and trends in the same (Exploratory Data Analysis). In the same, correlation between the features were studied to check for multicollinearity, the distribution of the data was visualized and the requirement to create new features were taken into consideration. Next, feature scaling was performed since the range of values of raw data was varying widely and in the ML algorithms, objective functions might not work properly without normalization. Next, the outliers or the extreme values were detected using the Tukey's method and the same were eliminated. The imbalanced data was also taken care of using the Naïve random over-sampling method, Synthetic Minority Oversampling Technique (SMOTE) and the Adaptive Synthetic Sampling (ADASYN) method. At the next stage of pre-processing feature selection was also carried out using the Scikit Learn python package. Multiple classifiers have been applied and the accuracies for each were compared before and after each pre-processing method. The model with the best accuracy over the testing dataset was then saved and was later loaded on the front-end.

3.3.3.2 Algorithms and Techniques used

There have been a total of 8 supervised learning algorithms used to model the pre-processed data on. The model accuracies have been compared and the one providing the best accuracy on the testing dataset or the previously unseen dataset was chosen to predict the final result on the user input. The algorithms and techniques used are as follows:

3.3.3.2.1 Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms. It is a classification technique that works on a probability basis. It is much similar to Linear Regression which gives continuous value. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.1*.

3.3.3.2.2 XG-Boost

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.4.*

3.3.3.2.3 Random Forest

Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.3.*

3.3.3.2.4 Support Vector Machine

SVM is a powerful supervised algorithm that works best on smaller datasets but on complex ones. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks, but generally, they work best in classification problems. This learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.6.*

3.3.3.2.5 K-Nearest Neighbor

SVM K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. This learning technique has been elaborated on and can be referred to from the *Section 3.3.2.2.5.*

3.3.3.2.6 Ada-Boost

AdaBoost, also called Adaptive Boosting, is a technique in Machine Learning used as an Ensemble Method. The most common estimator used with AdaBoost is decision trees with one level which means Decision trees with only 1 split. These trees are also called Decision Stumps.is a powerful supervised algorithm that works best on smaller datasets but on complex ones. This learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.5.*

3.3.2.2.7 Decision Tree

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.2.*

3.3.2.2.8 Stochastic Gradient Descent Classifier

SGD stands for Stochastic Gradient Descent Classifier is a linear classifier (SVM, logistic regression) optimized by the SGD. These are two different concepts. While SGD is a optimization method, Logistic Regression or linear Support Vector Machine is a machine learning algorithm/model. You can think of that a machine learning model defines a loss function, and the optimization method minimizes/maximizes it. This supervised learning technique has been elaborated on and can be hence referred to from the *Section 3.3.1.2.1.*

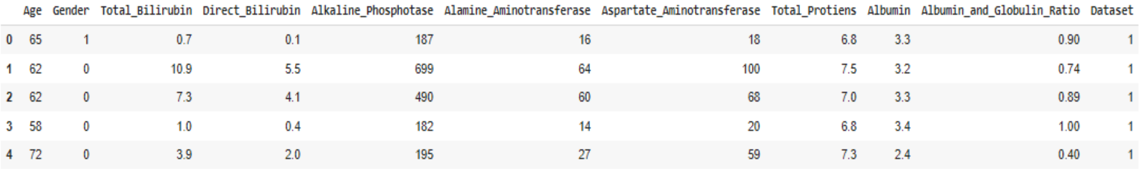
3.3.4 Liver Disease Prediction

The liver is the largest solid organ and the largest gland in the human body, that sits on the right side of the belly. Weighing about 3 pounds, the liver is reddish-brown in color and feels rubbery to the touch. The liver has two large sections, called the right and the left lobes. The gallbladder sits under the liver, along with parts of the pancreas and intestines. The liver and these organs work together to digest, absorb, and process food. Health care and medicine handles huge data on daily basis. Liver failure means that your liver is losing or has lost all of its function. It is a life-threatening condition that demands urgent medical care. In order to minimize these detrimental effects of liver failure, we must try to predict its presence at earlier stages. ML algorithms can help us effectively predict such results with a high degree of accuracy which can in turn help doctors and patients detect the onset of such diseases and reduce their impact or prevent them from occurring. Our objective was to integrate a sub-system that is able to accurately predict chances of liver failure in a time and cost-efficient manner. This objective has been successfully achieved by applying the following data pre-processing techniques and classification algorithms on the chosen dataset.

3.3.4.1 Data Pre-processing

The pre-processing of the data was started by analyzing the distribution of the data i.e. the distribution of the categorical attributes and the numerical or continuous attributes. Next EDA was carried out to check for the relationship between the variables in the dataset to check for multicollinearity. Next, the data that was to be processed was not clean and contained outliers, noise and missing values. This data if was not processed further would not have given the best results hence the process of cleaning the data collected was carried out and the unnecessary values from it were eliminated or filled according to the remaining values in the dataset. The outliers were visualized and handled using the boxplot visualization. The processed data found need for further simplified data therefore scaling was applied on it and it was standardized (the process of transformation).

The *Fig. 3.20.* depicts how the liver disease dataset looks like after the pre-processing techniques were applied on it.



	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Dataset
0	65	1	0.7	0.1	187	16	18	6.8	3.3	0.90	1
1	62	0	10.9	5.5	699	64	100	7.5	3.2	0.74	1
2	62	0	7.3	4.1	490	60	68	7.0	3.3	0.89	1
3	58	0	1.0	0.4	182	14	20	6.8	3.4	1.00	1
4	72	0	3.9	2.0	195	27	59	7.3	2.4	0.40	1

Fig. 3.20 Snippet of Pre – processed Liver Disease Dataset

3.3.4.2 Algorithms and Techniques used

There have been a total of 6 supervised learning algorithms used to model the pre-processed data on. The model accuracies have been compared and the one providing the best accuracy on the testing dataset or the previously unseen dataset was chosen to predict the final result on the user input. The algorithms and techniques used are as follows:

3.3.4.2.1 Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms. It is a classification technique that works on a probability basis. It is much similar to Linear Regression which gives continuous value. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.1.*

3.3.4.2.2 XG-Boost

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.4.*

3.3.4.2.3 Random Forest

Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.3.*

3.3.4.2.4 Support Vector Machine

SVM is a powerful supervised algorithm that works best on smaller datasets but on complex ones. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks, but generally, they work best in classification problems. This learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.6.*

3.3.4.2.5 Decision Tree

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.2.*

3.3.4.2.6 Gradient Boosting Classifier

Gradient Boosting is a system of machine learning boosting, representing a decision tree for large and complex data. It relies on the presumption that the next possible model will minimize the gross prediction error if combined with the previous set of models. The decision trees are used for the best possible predictions.

Gradient boosting is one of the most popular machine learning algorithms for tabular datasets. It is powerful enough to find any nonlinear relationship between your model target and features and has great usability that can deal with missing values, outliers, and high cardinality categorical values on your features without any special treatment. While you can build barebone gradient boosting trees using some popular libraries such as XGBoost or LightGBM without knowing any details of the algorithm, you still want to know how it works when you start tuning hyper-parameters, customizing the loss functions, etc., to get better quality on your model.

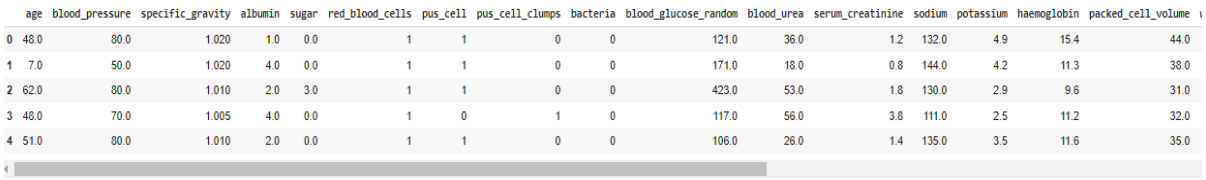
3.3.5 Chronic Kidney Disease Prediction

Chronic Kidney Disease (CKD) is one of the leading causes of morbidity and mortality for individuals with Cardiovascular Disease (CVD). A precise CKD risk prediction model developed from CVD patient data is critical for secondary prevention of CKD In order to minimize these detrimental effects of kidney disease, we must try to predict its presence at earlier stages. ML algorithms can help us effectively predict such results with a high degree of accuracy which can in turn help doctors and patients detect the onset of such diseases and reduce their impact or prevent them from occurring. Our objective was to integrate a sub-system that is able to accurately determine the presence of chronic kidney disease in a time and cost-efficient manner. This objective has been successfully achieved by applying the following data pre-processing techniques and classification algorithms on the chosen dataset:

3.3.5.1 Data Pre-processing

The data that is to be processed was not clean and contained outliers, noise and missing values. This data if was not processed further would not have given the best results hence the process of cleaning the data collected was carried out and the unnecessary values from it were eliminated or filled according to the remaining values in the dataset. The imputation of null values was carried out by filling the higher null values with the values in random sampling and the lower null values were filled using the mean/mode sampling. After which the categorical values were feature encoded using the Label Encoder. The processed data found no need for further simplified data therefore scaling was not applied on it (the process of transformation).

The *Fig. 3.21* displays how the chronic liver disease dataset looks like after the pre-processing techniques were applied on it.



	age	blood_pressure	specific_gravity	albumin	sugar	red_blood_cells	pus_cell	pus_cell_clumps	bacteria	blood_glucose_random	blood_urea	serum_creatinine	sodium	potassium	haemoglobin	packed_cell_volume
0	48.0	80.0	1.020	1.0	0.0	1	1	0	0	121.0	36.0	1.2	132.0	4.9	15.4	44.0
1	7.0	50.0	1.020	4.0	0.0	1	1	0	0	171.0	18.0	0.8	144.0	4.2	11.3	38.0
2	62.0	80.0	1.010	2.0	3.0	1	1	0	0	423.0	53.0	1.8	130.0	2.9	9.6	31.0
3	48.0	70.0	1.005	4.0	0.0	1	0	1	0	117.0	56.0	3.8	111.0	2.5	11.2	32.0
4	51.0	80.0	1.010	2.0	0.0	1	1	0	0	106.0	26.0	1.4	135.0	3.5	11.6	35.0

Fig. 3.21 Snippet of Pre – processed Chronic Liver Disease Dataset

3.3.5.2 Algorithms and Techniques used

There have been a total of 8 supervised learning algorithms used to model the pre-processed data on. The model accuracies have been compared and the one providing the best accuracy on the testing dataset or the previously unseen dataset was chosen to predict the final result on the user input. The algorithms and techniques used are as follows:

3.3.5.2.1 Extra Decision Tree (ETree Classifier)

Extremely Randomized Trees Classifier (Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a “forest” to output it’s classification result. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest.

Each Decision Tree in the Extra Trees Forest is constructed from the original training sample. Then, at each test node, each tree is provided with a random sample of k features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria (typically the Gini Index). This random sample of features leads to the creation of multiple de-correlated decision trees.

To perform feature selection using the above forest structure, during the construction of the forest, for each feature, the normalized total reduction in the mathematical criteria used in the decision of feature of split (Gini Index if the Gini Index is used in the construction of the forest) is computed. This value is called the Gini Importance of the feature. To perform feature selection, each feature is ordered in descending order according to the Gini Importance of each feature and the user selects the top k features according to his/her choice.

3.3.5.2.2 XG-Boost

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.4.*

3.3.5.2.3 Random Forest

Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.3.*

3.3.5.2.4 Gradient Boosting Classifier

Gradient boosting is one of the most popular machine learning algorithms for tabular datasets. It is powerful enough to find any nonlinear relationship between your model target and features and has great usability that can deal with missing values, outliers, and high cardinality categorical values on your features without any special treatment. is a powerful supervised algorithm that works best on smaller datasets but on complex ones. This learning technique has been elaborated on and can be referred to from the *Section 3.3.4.2.6.*

3.3.5.2.5 Decision Tree

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.2.*

3.3.5.2.6 Ada-Boost

AdaBoost, also called Adaptive Boosting, is a technique in Machine Learning used as an Ensemble Method. The most common estimator used with AdaBoost is decision trees with one level which means Decision trees with only 1 split. These trees are also called Decision Stumps.is a powerful supervised algorithm that works best on smaller datasets but on complex ones. This learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.5.*

3.3.5.2.7 K-Nearest Neighbor

SVM K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. This learning technique has been elaborated on and can be referred to from the *Section 3.3.2.2.5.*

3.3.5.2.8 Stochastic Gradient Boosting

It is also called Gradient Boosting Machines. Hence, this learning technique can be referred to from the *Section 3.3.4.2.6.*

3.3.6 Parkinson's Disease Prediction

Parkinson's disease is a neuro-degenerative disorder which affects quality of life of an estimated 10 million people worldwide. A tell-tale marker of this disease is a decrease in the dopamine levels in the brain which could be attributed to the degeneration of dopaminergic neurons. The onset of the disease may be suggested by tremor, rigidity, slowness of movement and postural instability. Such symptoms may not present itself in the same format in all the cases but rather vary in combinations and severity but generally is chronic and degenerative. What interests us here is that among all diagnosed cases of PD, 90% of the cases show some sort of vocal impairment which consists of deterioration of normal production of vocal sounds, which is medically termed as Dysphonia. The impact of this disease stands at a whopping 1-2% of people worldwide in the age range of 60 years and above. In order to minimize these detrimental effects of Parkinson's disease, we must try to predict its presence at earlier stages. ML algorithms can help us effectively predict such results with a high degree of accuracy which can in turn help doctors and patients detect the onset of such diseases and reduce their impact or prevent them from occurring. Our objective was to integrate a sub-system that is able to accurately determine the presence of Parkinson's disease in a time and cost-efficient manner. This objective has been successfully achieved by applying the following data pre-processing techniques and classification algorithms on the chosen dataset:

3.3.6.1 Data Pre-processing

The data was first looked at to get an overview of the attributes in the dataset. An in-depth analysis was carried out next to find the hidden patterns and trends in the same (Exploratory Data Analysis). In the same, correlation between the features were studied to check for multicollinearity, the distribution of the data was visualized and the requirement to create new features were taken into consideration. Next, feature scaling was performed since the range of values of raw data was varying widely and in the ML algorithms, objective functions might not work properly without standardization.

3.3.6.2 Algorithms and Techniques used

Support Vector Machine was used to model the pre-processed data on. The model accuracies have been evaluated on the testing dataset or the previously unseen dataset to check if the model would work properly in the real world and was finally loaded on the main interface to predict the final result on the user input. The algorithm and technique used has been elaborated in the following sub-section.

3.3.6.2.1 Support Vector Machine

SVM is a powerful supervised algorithm that works best on smaller datasets but on complex ones. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks, but generally, they work best in classification problems. This learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.6.*

3.3.7 Symptoms to Disease Prediction

Our medical care area every day gathers an enormous information worried about patients including clinical assessment, imperative boundaries, examination reports, therapy subsequent meet-ups, and drug choices and so forth. Yet, tragically it isn't examine and mine in a suitable manner. It is taken care of either in record room as bunches of paper sheet or devouring hard circle space. The experts similarly as investigators are hasty stressed over this huge data. It will supportive in different illnesses the executives including viability of surgeries, clinical trials, drug, and the disclosure of connections among clinical and determination information to utilize Data Mining systems. The medical care and clinical area are more needing data mining today. The soul point is utilizing the grouping so that it can help doctor. Illnesses also good fitness associated issues such as intestinal sickness, Chickenpox, Migraine, Diabetes, Impetigo, Jaundice, dengue and so forth, tend to critical impact on person’s wellbeing and at times may likewise prompt passing whenever disregarded. In order to minimize the detrimental effects of these life-threatening diseases, we must try to predict its presence at earlier stages. ML algorithms can help us effectively predict such results with a high degree of accuracy which can in turn help doctors and patients detect the onset of such diseases and reduce their impact or prevent them from occurring. Our objective was to integrate a sub-system that is able to accurately determine the presence of a disease in a time and cost-efficient manner from the symptoms provided as an input. This objective has been successfully achieved by applying the following data pre-processing techniques and classification algorithms on the chosen dataset:

3.3.7.1 Data Pre-processing

The data that is to be processed was not clean and contained outliers, noise and missing values. This data if was not processed further would not have given the best results hence the process of cleaning the data collected was carried out and the unnecessary values from it were eliminated or filled according to the remaining values in the dataset. The bar plots for each column or symptom was analyzed to check error of inconsistent data such as not comparable numerical measurement formats and data types. The outliers were handled using the Winsorisation technique where outlier values are replaced with the minimum or maximum non-outlier value identified using the interquartile range (IQR) method. Duplicate records were also looked for. The processed data found no need for further simplified data therefore scaling was not applied on it (the process of transformation). Next, since working on a complex data would be difficult, time-consuming and resource extensive hence the process of feature selection was carried out by visualizing the data as a heatmap to detect multicollinearity, if any and 132 attributes were reduced to 33 and the accuracy on the reduced data was evaluated again for the best choice of model. This process was carried out carefully to make sure that highly correlated features do not affect the final results.

3.3.7.2 Algorithms and Techniques used

There have been a total of 3 supervised learning algorithms used to model the pre-processed data on. The model accuracies have been compared and the one providing the best accuracy on the testing dataset or the previously unseen dataset was chosen to predict the final result on the user input. The algorithms and techniques used are as follows:

3.3.7.2.1 Multi-Layer Perceptron Classifier

Multi-layer perception is also known as MLP. It is fully connected dense layers, which transform any input dimension to the desired dimension. A multi-layer perception is a neural network that has multiple layers. To create a neural network we combine neurons together so that the outputs of some neurons are inputs of other neurons.

A multi-layer perceptron has one input layer and for each input, there is one neuron(or node), it has one output layer with a single node for each output and it can have any number of hidden layers and each hidden layer can have any number of nodes. The *Fig. 3.22* depicts a schematic diagram of a Multi-Layer Perceptron (MLP).

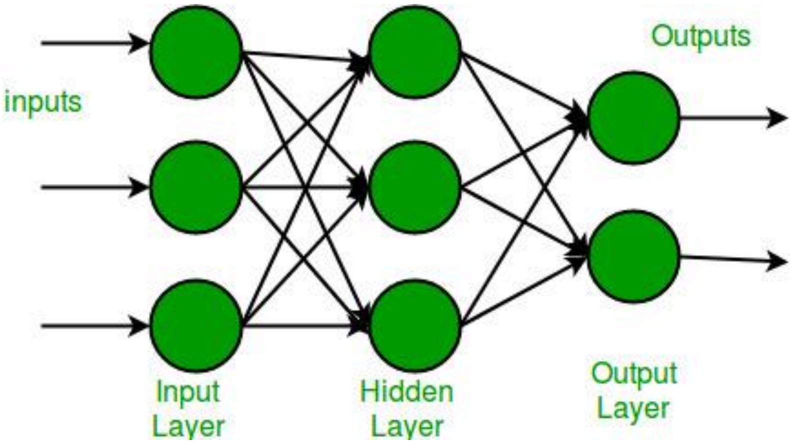


Fig. 3.22 Schematic diagram of a Multi-Layer Perceptron

In the multi-layer perceptron diagram above, we can see that there are three inputs and thus three input nodes and the hidden layer has three nodes. The output layer gives two outputs, therefore there are two output nodes. The nodes in the input layer take input and forward it for further process, in the diagram above the nodes in the input layer forwards their output to each of the three nodes in the hidden layer, and in the same way, the hidden layer processes the information and passes it to the output layer.

Every node in the multi-layer perception uses a sigmoid activation function. The sigmoid activation function takes real values as input and converts them to numbers between 0 and 1 using the sigmoid formula.

3.3.7.2.2 Decision Tree

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.2*.

3.3.7.2.3 Random Forest

Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. This supervised learning technique has been elaborated on and can be referred to from the *Section 3.3.1.2.3.*

Chapter 4

Implementation Details

4.1 Experimental Setup

The following section discusses about the setup used or required by the system. The section would basically include the dataset description of all the datasets used to build the model on and eventually to generate the final results.

4.1.1 Dataset Description

Thanks to the never-ending efforts of the researchers to make crucial metadata available to the common public, the following repositories culminated over the years have been chosen to make real-time disease specific predictions.

4.1.1.1 Heart Disease Dataset

For predicting the occurrence of heart diseases, we have used the "Heart Disease Dataset" by UCI. This dataset consists of 13 medical predictor features and one target feature. The dataset consists of 303 instances and 75 attributes like:

- age - age in years
- sex - (1 = male; 0 = female)
- cp - chest pain type

- 0: Typical angina: chest pain related decrease blood supply to the heart
    - 1: Atypical angina: chest pain not related to heart
    - 2: non-anginal pain: typically, esophageal spasms (non-heart related)
    - 3: Asymptomatic: chest pain not showing signs of disease
- trestbps - resting blood pressure (in mm Hg on admission to the hospital) anything above 130-140 is typically cause for concern
- chol - serum cholestoral in mg/dl
    - serum = LDL + HDL + .2 * triglycerides
    - above 200 is cause for concern

- fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
    - '>126' mg/dL signals diabetes
- restecg - resting electrocardiographic results
    - 0: Nothing to note
    - 1: ST-T Wave abnormality
        - can range from mild symptoms to severe problems
        - signals non-normal heart beat
    - 2: Possible or definite left ventricular hypertrophy
        - Enlarged heart's main pumping chamber
- thalach - maximum heart rate achieved
- exang - exercise induced angina (1 = yes; 0 = no)
- oldpeak - ST depression induced by exercise relative to rest looks at stress of heart during excercise unhealthy heart will stress more
- slope - the slope of the peak exercise ST segment
    - 0: Upsloping: better heart rate with excercise (uncommon)
    - 1: Flatsloping: minimal change (typical healthy heart)
    - 2: Downslopins: signs of unhealthy heart
- ca - number of major vessels (0-3) colored by flourosopy
    - colored vessel means the doctor can see the blood passing through

    - the more blood movement the better (no clots)
- thal - thalium stress result
    - 1,3: normal
    - 6: fixed defect: used to be defect but ok now
    - 7: reversable defect: no proper blood movement when excercising
- target - have disease or not (1=yes, 0=no) (= the predicted attribute)

4.1.1.2 Diabetes Dataset

For predicting the occurrence of diabetes, we have used the "PIMA Indians Diabetes Dataset" by UCI. This dataset consists of 9 medical predictor features and one target feature. The dataset consists of 2000 instances and 9 attributes like:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1)

4.1.1.3 Breast Cancer Dataset

For predicting the occurrence of breast cancer we have used the "Breast Cancer Wisconsin (Diagnostic) Dataset" from the UCI Machine Learning repository. This dataset consists of 31 medical predictor features and one target feature. The dataset consists of 569 instances and 32 attributes like id, radius-mean, texture-mean, perimeter-mean, area-mean, etc.

4.1.1.4 Liver Dataset

This data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India. The "Dataset" column is a class label used to divide groups into liver patient (liver disease) or not (no disease). This data set contains 441 male patient records and 142 female patient records.

Any patient whose age exceeded 89 is listed as being of age "90". The attributes are as follows:

- Age of the patient
- Gender of the patient
- Total Bilirubin
- Direct Bilirubin
- Alkaline Phosphotase
- Alamine Aminotransferase
- Aspartate Aminotransferase
- Total Protiens

4.1.1.5 Chronic Kidney Disease Dataset

For predicting the occurrence of Chronic Kidney diseases we have used the "Chronic Kidney Disease Data Set" by UCI. This dataset consists of 11 numeric and 14 nominal attributes in which one attribute is a class attribute. The dataset consists of 400 instances.

4.1.1.6 Parkinson’s Disease Dataset

For predicting the occurrence of Parkinson’s disease we will be using the "Parkinsons Data Set" by UCI. This dataset consists of 23 medical attributes related to vocal fundamental frequencies, health status of the subject, etc. The dataset consists of 197 instances.

4.1.1.7 Symptoms to Disease Dataset

The dataset is taken from Kaggle. It contains 4920 rows of observations and 134 columns of attributes. The data types of the attributes consist of 1 qualitative discrete categorical, 132 quantitative discrete binary and 1 quantitative continuous numerical float with 64-digit placings. The testing dataset contains 42 rows of observations and 133 columns. After performing the Feature selection step, the original dataset was reduced to contain 33 columns of predictors, out of the original 132.

The symptoms available on the system that help us to predict diseases are: 'itching', 'skin_rash', 'nodal_skin_eruptions' 'continuous_sneezing',' shivering','chills', 'joint_pain', 'stomach_pain', 'acidity',' 'ulcers_on_tongue', 'muscle_wasting', 'vomiting', ''burning_micturition', 'spotting_ urination', 'fatigue', 'weight_gain', 'anxiety', 'cold_hands_and_feets', 'mood_swings', 'weight_loss',, 'restlessness', 'lethargy', 'patches_in_throat', 'irregular_sugar_level', 'cough', 'high_fever', 'sunken_eyes', 'breathlessness', 'sweating', 'dehydration', 'indigestion', 'headache', 'yellowish_skin', 'dark_urine', 'nausea', 'loss_of_appetite', 'pain_behind_the_eyes', back_pain', constipation','abdominal_pain', 'diarrhoea', 'mild_fever', etc.

The system is able to predict the following diseases: (vertigo) Paroymsal Positional Vertigo', 'AIDS','Acne','Alcoholic hepatitis','Allergy', 'Arthritis','Bronchial Asthma', 'Cervical spondylosis', 'Chicken pox', 'Chronic cholestasis', 'Common Cold', 'Dengue', 'Diabetes ', 'Dimorphic hemmorhoids(piles)', 'Drug Reaction' 'Fungal infection', 'GERD','Gastroenteritis', 'Heart attack','Hepatitis B', 'Hepatitis C', 'Hepatitis D', 'Hepatitis E', 'Hypertension','Hyperthyroidism', 'Hypoglycemia','Hypothyroidism', 'Impetigo', 'Jaundice', 'Malaria', 'Migraine','Osteoarthristis', 'Paralysis (brain hemorrhage)', 'Peptic ulcer diseae', 'Pneumonia', 'Psoriasis', 'Tuberculosis', etc.

4.2 Software and Hardware Setup

The hardware and the software requirements for the proposed system/work are as follows:

4.2.1 Hardware Requirements

Processor - Minimum 1 GHz; Recommended 2GHz or more.
RAM - Minimum 1 GB; Recommended 4 GB or above.

4.2.2 Software Requirements

Operating System - Windows 7 or higher versions.
Technology - Python3.7.
IDE - Google Colaboratory Notebook.
Browser - Firefox / Chrome/ Internet Explorer / Safari or any other web browser.

4.2.2.1 Why Python?

- General purpose programming language
- Increasing popularity for use in data science
- Easy to build end-to-end products like web applications
- Since the goal of this project is to build a web application, Python is a better choice. Though frameworks like Shiny can be used with R to create web applications, it is extremely slow.

4.2.3 Libraries Used

Python libraries that are used in Machine Learning are:

- Numpy
- Scipy
- Scikit-learn
- TensorFlow
- Pandas
- Matplotlib
- Seaborn
- Keras

4.2.3.1 NumPy

NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra, Fourier transform, and random number capabilities. High-end libraries like TensorFlow uses NumPy internally for manipulation of Tensors.

4.2.3.2 Scipy

SciPy is a very popular library among Machine Learning enthusiasts as it contains different modules for optimization, linear algebra, integration and statistics. There is a difference between the SciPy library and the SciPy stack. The SciPy is one of the core packages that make up the SciPy stack. SciPy is also very useful for image manipulation.

4.2.3.3 Scikit-learn

Scikit-learn is one of the most popular ML libraries for classical ML algorithms. It is built on top of two basic Python libraries, viz., NumPy and SciPy. Scikit-learn supports most of the supervised and unsupervised learning algorithms. Scikit-learn can also be used for data-mining and data-analysis, which makes it a great tool who is starting out with ML.

4.2.3.4 TensorFlow

TensorFlow is a very popular open-source library for high performance numerical computation developed by the Google Brain team in Google. As the name suggests, Tensorflow is a framework that involves defining and running computations involving tensors. It can train and run deep neural networks that can be used to develop several AI applications. TensorFlow is widely used in the field of deep learning research and application.

4.2.3.5 Pandas

Pandas is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and wide variety tools for data analysis. It provides many inbuilt methods for grouping, combining and filtering data.

4.2.3.6 Matplotlib

Matplotlib is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It particularly comes in handy when a programmer wants to visualize the patterns in the data. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualization, viz., histogram, error charts, bar chats, etc,

4.2.3.7 Seaborn

Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions. Seaborn library aims to make a more attractive visualization of the central part of understanding and exploring data. It is built on the core of the matplotlib library and also provides dataset-oriented APIs. Seaborn is also closely integrated with the Panda's data structures, and with this, we can easily jump between the various different visual representations for a given variable to better understand the provided dataset.

4.2.3.8 Keras

It provides many inbuilt methods for groping, combining and filtering data. Keras is a very popular Machine Learning library for Python. It is a high-level neural networks API capable of running on top of TensorFlow, CNTK, or Theano. It can run seamlessly on both CPU and GPU. Keras makes it really for ML beginners to build and design a Neural Network.

Chapter 5

Results and Discussion

5.1 Performance Evaluation Parameters

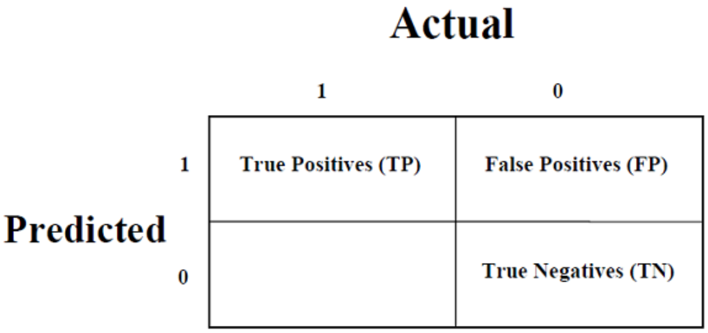
There are various ways to check the performance of our machine learning or deep learning model and why to use one in place of the other. We will discuss terms like:

- Confusion matrix
- Accuracy
- Precision
- Recall
- Specificity
- F1 score
- Precision-Recall or PR curve
- ROC (Receiver Operating Characteristics) curve
- PR vs ROC curve.

5.1.1 Confusion Matrix

It is the easiest way to measure the performance of a classification problem where the output can be of two or more type of classes. A confusion matrix is nothing but a table with two dimensions viz. "Actual" and "Predicted" and furthermore, both the dimensions have "True Positives (TP)," "True Negatives (TN)", "False Positives (FP)", "False

Negatives (FN)”. The same can be depicted in the *Fig. 5.1*.



		Actual	
		1	0
Predicted	1	True Positives (TP)	False Positives (FP)
	0		True Negatives (TN)

Fig 5.1 Confusion Matrix

Explanation of the terms associated with confusion matrix are as follows −

- True Positives (TP) − It is the case when both actual class & predicted class of data point is 1.
- True Negatives (TN) − It is the case when both actual class & predicted class of data point is 0.
- False Positives (FP) − It is the case when actual class of data point is 0 & predicted class of data point is 1.
- False Negatives (FN) − It is the case when actual class of data point is 1 & predicted class of data point is 0.

We can use confusion_matrix function of sklearn.metrics to compute Confusion Matrix of our classification model.

5.1.2 Accuracy

The most commonly used metric to judge a model and is actually not a clear indicator of the performance. The worse happens when classes are imbalanced.

Accuracy= (TP+TN)/(TP+FP+TN+FN)

Take for example a cancer detection model. The chances of actually having cancer are very low. Let’s say out of 100, 90 of the patients don’t have cancer and the remaining 10 actually have it. We don’t want to miss on a patient who is having cancer but goes undetected (false negative). Detecting everyone as not having cancer gives an accuracy of 90% straight. The model did nothing here but just gave cancer free for all the 100 predictions.

It is good to use the Accuracy metric when the target variable classes in data are approximately balanced. For example, if 60% of classes in a fruit image dataset are of Apple, 40% are Mango. In this case, if the model is asked to predict whether the image is of Apple or Mango, it will give a prediction with 97% of accuracy.

It is recommended not to use the Accuracy measure when the target variable majorly belongs to one class. For example, Suppose there is a model for a disease prediction in which, out of 100 people, only five people have a disease, and 95 people don't have one. In this case, if our model predicts every person with no disease (which means a bad prediction), the Accuracy measure will be 95%, which is not correct.

5.1.3 Precision

Percentage of positive instances out of the total predicted positive instances. Here denominator is the model prediction done as positive from the whole given dataset. Take it as to find out 'how much the model is right when it says it is right'. Precision, used in document retrievals, may be defined as the number of correct documents returned by our ML model. We can easily calculate it by confusion matrix with the help of following formula –

$$\text{Precision}= TP/(TP+FP)$$

5.1.4 Recall/Sensitivity/True Positive Rate

Percentage of positive instances out of the total actual positive instances. Therefore denominator (TP + FN) here is the actual number of positive instances present in the dataset. Take it as to find out 'how much extra right ones, the model missed when it showed the right ones'.

$$\text{Recall}= TP/(TP+FN)$$

From the above definitions of Precision and Recall, we can say that recall determines the performance of a classifier with respect to a false negative, whereas precision gives information about the performance of a classifier with respect to a false positive.

So, if we want to minimize the false negative, then, Recall should be as near to 100%, and if we want to minimize the false positive, then precision should be close to 100% as possible.

In simple words, if we maximize precision, it will minimize the FP errors, and if we maximize recall, it will minimize the FN error. The most commonly used metric to judge a model and is actually not a clear indicator of the performance. The worse happens when classes are imbalanced.

5.1.5 Specificity

Percentage of negative instances out of the total actual negative instances. Therefore denominator (TN + FP) here is the actual number of negative instances present in the dataset. It is similar to recall but the shift is on the negative instances. Like finding out how many healthy patients were not having cancer and were told they don’t have cancer. Kind of a measure

to see how separate the classes are. It can be calculated using the formula:

Specificity= TN/(TN+FP)

5.1.6 F1 - Score

It is the harmonic mean of precision and recall. This takes the contribution of both, so higher the F1 score, the better. See that due to the product in the numerator if one goes low, the final F1 score goes down significantly. So a model does well in F1 score if the positive predicted are actually positives (precision) and doesn't miss out on positives and predicts them negative (recall).

$$F1 - score = 2 * \frac{precision * recall}{precision + recall}$$

As F-score make use of both precision and recall, so it should be used if both of them are important for evaluation, but one (precision or recall) is slightly more important to consider than the other. For example, when False negatives are comparatively more important than false positives, or vice versa.

One drawback is that both precision and recall are given equal importance due to which according to our application we may need one higher than the other and F1 score may not be the exact metric for it. Therefore either weighted-F1 score or seeing the PR or ROC curve can help.The most commonly used metric to judge a model and is actually not a clear indicator of the performance. The worse happens when classes are imbalanced.

5.1.7 ROC Curve

ROC stands for receiver operating characteristic and the graph is plotted against TPR and FPR for various threshold values. As TPR increases FPR also increases. There are four categories and we want the threshold value that leads us closer to the top left corner. Comparing different predictors (here 3) on a given dataset also becomes easy, one can choose the threshold according to the application at hand. ROC AUC is just the area under the curve, the higher its numerical value the better. *Fig 5.2* shows an ROC Curve.

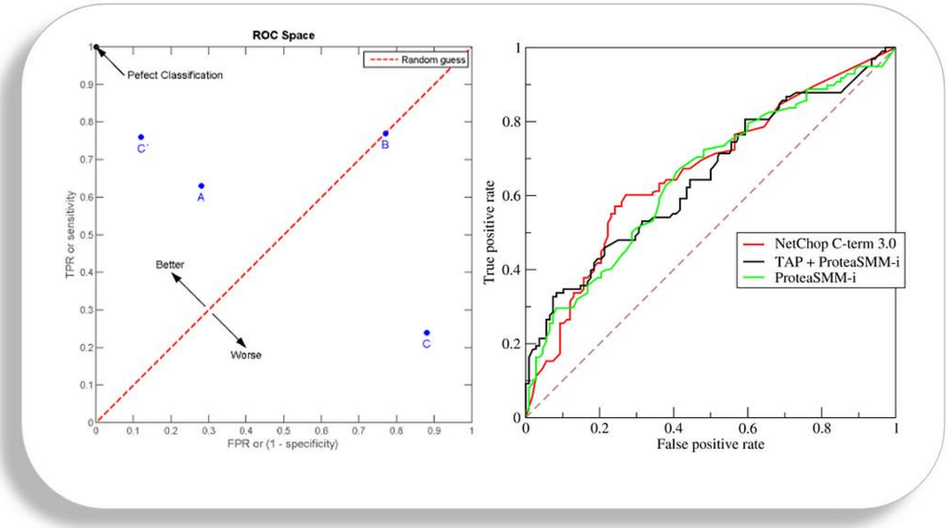


Fig 5.2 ROC Curve

5.2 Implementation Results

This section discussed in detail the performance of each algorithm when trained on each sub-system or a single disease prediction system. This is done to compare and choose the best model in order to get the most accurate results.

5.2.1 Prediction Results Accuracy

This sub-section discusses the performance of the models in terms of the performance evaluation metric called accuracy. The accuracy evaluation metric has been elaborated on and can be referred to in the *Section 5.1.2*.

5.2.1.1 Heart Disease

For predicting the occurrence of heart diseases, we have used the "HeartDisease Dataset" by UCI. This dataset consists of 13 medical predictor features and one target feature. The dataset consists of 303 instances and 75 attributes like chol, cp, trestbps, age, fbs, sex, restecg, etc. We have run our dataset through seven machine learning models. The *Table 5.1* gives the accuracies achieved for each machine learning model.

Table 5.1 – Accuracy Results for Heart Disease Detection

Model	**Accuracy**
Logistic Regression	85.24%
Decision Tree	84.78%
Random Forest	80.43%
XG Boost	78.26%
Adaboost	67.39%
Support Vector Machine	89.13%
Stacking Logistic Regression, KNN, Decision Tree Classifier	**91.30%**

Out of all these, we are getting the best results for Stacking Ensemble Technique with an accuracy of 91.3% on the test results.

5.2.1.2 Diabetes

For predicting the occurrence of Diabetes diseases, we have used the "Pima Indians Diabetes Dataset" by Kaggle. This dataset consists of 8 medical predictor features and one target feature. The dataset consists of 768 instances and 9 attributes like BloodPressure, pregnancies, glucose, SkinThickness, BMI, etc. We have run our dataset through five machine learning models. They are: For predicting the occurrence of heart diseases, we will be using the "HeartDisease Dataset" by UCI. This dataset consists of 13 medical predictor features and one target feature. The dataset consists of 303 instances and 75 attributes like chol, cp, trestbps, age, fbs, sex, restecg, etc. We have run our dataset through six machine learning models. The *Table 5.2* gives the accuracies achieved for each machine learning model.

Table 5.2 – Accuracy Results for Diabetes Disease Detection

Model	**Accuracy**
Logistic Regression	78.20%
XGBoost	79.90%
Random Forest Classifier	78.80%
Support Vector Machine	**80.25%**
KNN	77%

Out of all these, we are getting the best results for Support Vector Machine with an accuracy of 80.25% on the test results.

5.2.1.3 Breast Cancer

For predicting the occurrence of breast cancer, we have used the "Breast Cancer Wisconsin (Diagnostic) Dataset" by Kaggle. This dataset consists of 31 medical predictor features and one target feature. The dataset consists of 569 instances and 32 attributes like id, radius-mean, texture-mean, perimeter-mean, area-mean, etc. Several techniques were used. The *Table 5.3* gives the accuracies achieved for each machine learning model when individual pre-processing techniques were applied on it and the Table 5.4 gives the accuracies achieved for the same machine learning models when the combinations of these preprocessing techniques were applied.

Table 5.3 – Accuracy Results for Breast Cancer Detection (individual techniques applied)

Classifier	**Original**	**Scaled**	**Outliers Removed**	**12 Features**	**10 Features**	**Resampled Random**	**SMOTE**	**ADASYN**
Random Forest	**97.36%**	96.47%	96.16%	95.60%	**98.24%**	95.10%	95.81%	95.13%
Extra DecisionTrees	**97.36%**	96.47%	**97.10%**	95.60%	96.45%	95.80%	96.50%	96.52%
Decision Tree	94.71%	95.60%	91.24%	92.86%	94.74%	94.41%	93.00%	93.02%
Support Vector	94.65%	97.36%	89.21%	91.88%	91.88%	82.52%	83.21%	77.03%
AdaBoost Classifier	96.47%	96.47%	94.21%	95.58%	93.84%	97.20%	95.80%	96.52%
Gradient Boosting	95.60%	95.60%	95.20%	94.71%	96.47%	95.80%	**96.50%**	96.53%
SGD Classifier	94.65%	**97.37%**	90.23%	93.77%	84.45%	83.16%	82.26%	75.83%
Logistic Regression	95.58%	97.36%	93.18%	**96.47%**	98.25%	**97.90%**	95.10%	95.83%
XGB Classifier	**97.36%**	97.36%	96.16%	93.81%	95.62%	96.50%	95.80%	**97.22%**
KNN Classifier	96.45%	96.47%	89.21%	92.86%	92.86%	86.02%	86.00%	85.41%

Table 5.4 – Accuracy Results for Breast Cancer Detection (combination of techniques applied)

Classifier	Scaled + Outliers Removed	Scaled + Resampled	Scaled+Outliers +Resampled	Scaled+Feature +Outlier+Resampled
Random Forest	95.10%	97.90%	**99.28%**	**93.01%**
Extra DecisionTrees	95.16%	**98.60%**	96.38%	92.31%
Decision Tree	91.24%	95.80%	96.38%	90.91%
Support Vector	97.10%	97.90%	97.83%	90.91%
AdaBoost Classifier	94.21%	97.20%	98.55%	90.91%
Gradient Boosting	92.18%	97.90%	98.55%	92.31%
SGD Classifier	96.14%	95.08%	97.10%	90.91%
Logistic Regression	96.12%	97.90%	97.83%	91.61%
XGB Classifier	96.16%	97.90%	98.55%	91.61%
KNN Classifier	**98.06%**	94.40%	98.55%	90.91%

Out of all these, we are getting the best results for Random Forest Technique applied on Scaled, Outliers handled and resampled dataset with an accuracy of 99.28% on the test results.

5.2.1.4 Liver Disease

For predicting the occurrence of liver diseases, we have used the "ILPD (Indian Liver Patient Dataset)" by UCI. This data set contains 416 liver patient records and 167 non liver patient records (583 in total). It consists of 10 medical attributes like age, gender, total_bilirubin, total_proteins, etc. We have run our dataset through six machine learning models. The *Table 5.5* gives the accuracies achieved for each machine learning model.

Table 5.5 – Accuracy Results for Liver Disease Detection

Model	**Accuracy**
Logistic Regression	69.4%
Gradient Boosting Classifier	70.7%
Support Vector Machine	68.4%
Random Forest Classifier	**74.3%**
Decision Tree Classifier	68.4%
XGBoost	68.4%

Out of all these, we are getting the best results for Random Forest Classifier with an accuracy of 74.3% on the test results.

5.2.1.5 Kidney Disease

For predicting the occurrence of chronic kidney diseases, we have used the "Chronic Kidney Disease Data Set" by UCI. This dataset consists of 11 numeric and 14 nominal attributes in which one attribute is a class attribute. The dataset consists of 400 instances.etc. We have run our dataset through eight machine learning models. The *Table 5.6* gives the accuracies achieved for each machine learning model.

Table 5.6 – Accuracy Results for Chronic Kidney Disease Detection

Model	Accuracy
Extra Trees Classifier	**98.33%**
Gradient Boosting Classifier	96.67%
Stochastic Gradient Boosting	96.67%
Decision Tree Classifier	95%
XGBoost	94.16%
Adaboost	90.83%
KNN	61.67%

Out of all these, we are getting the best results for ET Classifier with an accuracy of 99.33% on the test results.

5.2.1.5 Parkinson’s Disease

For predicting the occurrence of Parkinsons disease we have used the "Parkinsons Data Set" by UCI. This dataset consists of 23 medical attributes related to vocal fundamental frequencies, health status of the subject, etc. The dataset consists of 197 instances., etc. Out of several tecniques applied, we are getting the best results for SVM Classifier with an accuracy of 87.17% on the test results.

5.2.1.7 Symptoms to Disease Prediction

For predicting the disease from the symptoms, we have used the "Symptoms to Disease Prediction dataset" by Kaggle. This dataset consists of 4000+ instances and 133 attributes with 132 symptoms and 1 target variable called “prognosi”. We have run our dataset through three machine learning models. The *Table 5.7* gives the accuracies achieved for each machine learning model.

Table 5.7 – Accuracy Results for Symptoms to Disease Prediction sub-system

Model	Training Accuracy	Testing Accuracy
Multi layer perception classifier	100%	90.67%
Random Forest Classifier	100%	100%
Decision Tree Classifier	100 %	100%
Accuracy after feature selection	**90.97%**	**90.47%**

Out of all these, we are getting the best results for DT Classifier with an accuracy of 100% on the test results. The RF classifier also provides an accuracy of 100% on the test dataset but it is costly in terms of computation hence DT is chosen.

5.3 Results Discussion

This section discusses the final best results obtained on applying multiple algorithms to the data for each sub-section or the single disease prediction systems. The model with the best accuracy was saved and was loaded to the main interface to obtain the most accurate final results on the user's input.

5.3.1 Heart Disease Prediction Results

With the increasing number of deaths due to heart diseases, it has become significant to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This project predicts people with cardiovascular disease by extracting the patient medical history that leads to a fatal heart disease from a dataset that includes patients' medical history such as chest pain, sugar level, blood pressure, etc. This study compares the accuracy score of various models, namely Decision Tree, Logistic Regression, XGBoost, Adaboost, Support Vector Machine, Random Forest and Stacking Ensemble Technique for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the for Stacking Ensemble Technique is the most efficient algorithm with accuracy score of 91.3% for prediction of heart disease, as seen in *Table 5.1*. By using these, computer aided techniques we can predict the patient fast and better and the cost can be reduced very much.

5.3.2 Diabetes Prediction Results

The main aim of this project was to design and implement Diabetes Prediction Using ML Methods and Performance Analysis of that methods. The proposed approach uses various classification and ensemble learning method in which SVM, KNN, Random Forest, XGBoost, Logistic Regression are used. Out of all these, we are getting the best results for Support Vector Machine with an accuracy of 80.25% on the test results. The Experimental results can help health care section to make early predictions and make early decision to cure diabetes and save humans life.

5.3.3 Breast Cancer Prediction Results

Breast cancer if found at an early stage will help save lives of thousands of women or even men. These projects help the real-world patients and doctors to gather as much information as they can. The research on nine papers has helped us gather the data for the project proposed by us. By using machine learning algorithms, we will be able to classify and predict the cancer into being or malignant. Machine learning algorithms can be used for medical oriented research, it advances the system, reduces human errors and lowers manual mistakes. For predicting the occurrence of breast cancer, we will be using the "Breast Cancer Wisconsin (Diagnostic) Dataset" by Kaggle. This dataset consists of 31 medical predictor features and one target feature. The dataset consists of 569 instances and 32 attributes like id, radius-mean, texture-mean, perimeter-mean, area-mean, etc. The experimental results in *Table 5.4* illustrate that the Random Forest algorithm achieves the highest accuracy of 99.28% and thus successfully achieving the objective of improving the prediction accuracy.

5.3.4 Liver Disease Prediction Results

The prediction of liver illness in patients has been examined by analyzing the "ILPD (Indian Liver Patient Dataset)" by UCI which contains 416 liver patient records and 167 non liver patient records (583 in total). It consists of 10 medical attributes like age, gender, total_bilirubin, total_proteins, etc. We have run our dataset through six machine learning models, namely SVM, Decision Tree, Random Forest, XGBoost, Logistic Regression & Gradient Boosting. Best results for Random Forest Classifier with an accuracy of 74.3% on the test results as seen in *Table 5.5*.

5.3.5 Chronic Kidney Prediction Results

Chronic kidney disease is a global health threat and becoming silent killer in Ethiopia. Many people die or suffer severely by the disease mainly due to lack of awareness about the disease and inability to detect early. Thus, early prediction of chronic kidney disease is believed to be helpful to slow the progress of the disease. Machine learning plays a vital role in early disease identification and prediction. It supports the decision of medical experts by enabling them to diagnose the disease fast and accurately. For predicting the occurrence of chronic kidney diseases, we will be using the "Chronic Kidney Disease Data Set" by UCI. This dataset consists of 11 numeric and 14 nominal attributes in which one attribute is a class attribute. The dataset consists of 400 instances. We have run our dataset through eight machine learning models. The experimental results illustrated in *Table 5.6* that the Extra Trees Classifier achieves the highest accuracy of 98.33%.

5.3.6 Parkinson’s Disease Prediction Results

Parkinson's disease is a brain disorder that affects the central nervous system (CNS), and there is currently no cure for it unless it is diagnosed early. Late detection results in no therapy and death. As a result, early detection is critical. We used several machine learning algorithms for early disease detection because they are known for their efficiency and quick retrieval. For predicting the occurrence of Parkinson’s disease, we used the "Parkinson’s Data Set" by UCI. This dataset consists of 23 medical attributes related to vocal fundamental frequencies, health status of the subject, etc. The dataset consists of 197 instances., etc. Out of several techniques applied, we are getting the best results for SVM Classifier with an accuracy of 87.17% on the test results.

5.3.7 Symptoms to Disease Prediction Results

We have run our dataset through several machine learning models, however, DT model has the best performance metrics of 100% for all four metrics of test accuracy, precision, recall, and F1-score. The RF model also achieved 100% for all, but this is not preferred as the time complexity will usually be larger. On the other hand, the MLP model provides a testing accuracy of 97.62%. Hence, for all three chosen models, all the prognosis are almost perfectly classified and predicted.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In conclusion, there are many disease-predictive models available in the market. However, only ‘Predictal’ provides a unified platform where a user can predict the risk of having various diseases, with high accuracy. The user does not need to traverse through different websites to check for different symptoms, which saves time as the user can simply enter some parameter values and find out if they are at risk of the disease.

The user need not go to a doctor to interpret the results of a medical test performed. Instead, just input the values into the input fields and get results. The user interacts with the prediction system by filling out a form that holds the parameter set provided as input to the trained models. The Prediction system provides an optimal performance compared to other state of art approaches.

The diseases being covered by Predictal are Heart Disease, Diabetes, Breast Cancer, Parkinson’s, Kidney Disease, and Liver Disease. There also exists a section for the prediction of various kinds of potential diseases based on the selected symptoms. There are in total 132 symptoms in the list for the user to choose from.

Diseases, if predicted early can increase your life expectancy. For this purpose, we have developed a Prediction Engine using different models and chosen the one with the highest accuracy which enables the user to check whether he/she has a higher risk of diabetes or breast cancer, or heart disease.

This section discussed in detail the performance of each algorithm when trained on each sub-system or a single disease prediction system. This is done to compare and choose the best model in order to get the most accurate results.

6.2 Future Work

There are many things we intend to develop in the future -

- Although, the created project has Machine Learning models with high accuracies, to make them more accurate we intend to improve our data sources. Connecting our product with trustworthy medical data sources such as Hospital Databases, and EHRs (Electronic Health Records) can provide us with more accurate as well as detailed data about the patients. This will help us in achieving reliable data and hence improve prediction accuracy.
- Adding a section where a patient can provide feedback about the results achieved in detail can help us evaluate the models used in a better way. We can also take feedback from industry professionals regarding the accuracy of the predictions made by our platform.
- Currently. Our model only provides a statement regarding the possibility of having a certain disease. In the future, we can provide the output to the user with an explanation regarding the prediction made. Providing the user with an explanation in detail can help us build more trust with the user and also help users understand the precautions to take in order to avoid the worsening of diseases.
- We will be developing our platform in terms of visualization. Users can understand the results in a better way if the presented output is visual. This section discusses the final best results obtained on applying multiple algorithms to the data for each sub-section or the single

References

[1] Stephen J McPhee, Maxine A Papadakis, Michael W Rabow, et al. Current medical diagnosis and treatment 2010. McGraw-Hill Medical New York:, 2010.

[2] Md Manjurul Ahsan, Md Tanvir Ahad, Farzana Akter Soma, Shuva Paul, Ananna Chowdhury, Shahana Akter Luna, Munshi Md Shafwat Yazdan, Akhlaqur Rahman, Zahed Siddique, and Pedro Huebner. Detecting sars-cov-2 from chest x-ray using artificial intelligence. Ieee Access, 9:35501–35513, 2021.

[3] Eric R Coon, Ricardo A Quinonez, Virginia A Moyer, and Alan R Schroeder. Overdiagnosis: how our compulsion for diagnosis may be harming children. Pediatrics, 134(5):1013–1023, 2014.

[4] Erin P Balogh, Bryan T Miller, and John R Ball. Improving diagnosis in health care. The National Academies of Sciences Engineering Medicine, 2015.

[5] Md Manjurul Ahsan and Zahed Siddique. Machine learning-based heart disease diagnosis:A systematic literature review. arXiv preprint arXiv:2112.06459, 2021.

[6] Md Manjurul Ahsan, Tasfiq E Alam, Theodore Trafalis, and Pedro Huebner. Deep mlp- cnn model using mixed-data to distinguish between covid-19 and non-covid-19 patients.Symmetry, 12(9):1526, 2020.

[7] IS Stafford, M Kellermann, E Mossotto, RM Beattie, BD MacArthur, and S Ennis. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. NPJ digital medicine, 3(1):1–11, 2020.

[8] Md Manjurul Ahsan, Kishor Datta Gupta, Mohammad Maminur Islam, Sajib Sen, Md Rahman, Mohammad Shakhawat Hossain, et al. Covid-19 symptoms detection based on nasnetmobile with explainable ai using various imaging modalities. Machine Learning and Knowledge Extraction, 2(4):490–504, 2020.

[9] Arthur L Samuel. Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3):210–229, 1959.

[10] Jason Brownlee. Machine learning mastery with python. Machine Learning Mastery Pty Ltd, 527:100–120, 2016.

[11] Essam H Houssein, Marwa M Emam, Abdelmgeid A Ali, and Ponnuthurai Nagaratnam Suganthan. Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. Expert Systems with Applications, 167:114161, 2021.

[12] Mr Brijain, R Patel, MR Kushik, and K Rana. A survey on decision tree algorithm for classification. CiteSeer, 2014.

[13] Rajesh S Walse, Gajanan D Kurundkar, Santosh D Khamitkar, Aniket A Muley, Parag U Bhalchandra, and Sakharam N Lokhande. Effective use of naïve bayes, decision tree, and random forest techniques for analysis of chronic kidney disease. In International Conference on Information and Communication Technology for Intelligent Systems, pages 237–245.Springer, 2020.

[14] Keerthana Rajendran, Manoj Jayabalan, and Vinesh Thiruchelvam. Predicting breast cancer via supervised machine learning methods on class imbalanced data. Int. J. Adv. Comput. Sci. Appl., 11(8):54–63, 2020.

[15] Hsin-Yi Tsao, Pei-Ying Chan, and Emily Chia-Yu Su. Predicting diabetic retinopathyand identifying interpretable biomedical features using machine learning algorithms. BMC bioinformatics, 19(9):111–121, 2018. [16] A Nurrohman, S Abdullah, and H Murfi. Parkinson's disease subtype classification: Application of decision tree, logistic regressionand logit leaf model. In AIP Conference Proceedings, volume 2242, page 030015. AIP Publishing LLC, 2020.

[17] Harris Drucker, Donghui Wu, and Vladimir N Vapnik. Support vector machines for spam categorization. IEEE Transactions on Neural networks, 10(5):1048–1054, 1999.

[18] Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique, 57(3):238–247, 1989.

[19] Raymond E Wright. Logistic regression. APA PsycNet, 1995.

[20] Robert E Schapire. Explaining adaboost. In Empirical inference, pages 37–52. Springer, 2013.

[21] Purushottam, et al. "Efficient Heart Disease Prediction System." *Procedia Computer Science*, vol. 85, Elsevier BV, 2016, pp. 962–69. *Crossref*, https://doi.org/10.1016/j.procs.2016.05.288.

[22] Apurb R., Milan S., Avi A., Dundigalla R., Poonam G., O., 2020. Heart Disease Prediction using Machine Learning. *International Journal of Engineering Research & Technology* (IJERT), [online] Vol. 9 Issue 04, ISSN: 2278-0181. Available at: https://www.researchgate.net/publication/341870785_Heart_Disease_Prediction_using_Mach ine_Learning

[23] Santhana Krishnan J and Geetha S, "Prediction of Heart Disease using Machine LearningAlgorithms" ICIICT, 2019.

[24] Harris Drucker, Donghui Wu, and Vladimir N Vapnik. Support vector machines for spam categorization. IEEE Transactions on Neural networks, 10(5):1048–1054, 1999.

[25] Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique, 57(3):238–247, 1989.

[26] Raymond E Wright. Logistic regression. APA PsycNet, 1995.

[27] Robert E Schapire. Explaining adaboost. In Empirical inference, pages 37–52. Springer, 2013.

[28] Purushottam, et al. "Efficient Heart Disease Prediction System." *Procedia Computer Science*, vol. 85, Elsevier BV, 2016, pp. 962–69. *Crossref*, https://doi.org/10.1016/j.procs.2016.05.288.

[29] Apurb R., Milan S., Avi A., Dundigalla R., Poonam G., O., 2020. Heart Disease Prediction using Machine Learning. *International Journal of Engineering Research & Technology* (IJERT), [online] Vol. 9 Issue 04, ISSN: 2278-0181. Available at: https://www.researchgate.net/publication/341870785_Heart_Disease_Predicti

on_using_Mach ine_Learning

[30] Santhana Krishnan J and Geetha S, “Prediction of Heart Disease using Machine LearningAlgorithms” ICIICT, 2019.

[31] Harris Drucker, Donghui Wu, and Vladimir N Vapnik. Support vector machines for spam categorization. IEEE Transactions on Neural networks, 10(5):1048–1054, 1999.

[32] Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique, 57(3):238–247, 1989.

[33] Raymond E Wright. Logistic regression. APA PsycNet, 1995.

[34] Robert E Schapire. Explaining adaboost. In Empirical inference, pages 37–52. Springer, 2013.

[35] Purushottam, et al. “Efficient Heart Disease Prediction System.” *Procedia Computer Science*, vol. 85, Elsevier BV, 2016, pp. 962–69. *Crossref*, https://doi.org/10.1016/j.procs.2016.05.288.

[36] Apurb R., Milan S., Avi A., Dundigalla R., Poonam G., O., 2020. Heart Disease Prediction using Machine Learning. *International Journal of Engineering Research & Technology* (IJERT), [online] Vol. 9 Issue 04, ISSN: 2278-0181. Available at: https://www.researchgate.net/publication/341870785_Heart_Disease_Predicti on_using_Mach ine_Learning

[37] Santhana Krishnan J and Geetha S, “Prediction of Heart Disease using Machine LearningAlgorithms” ICIICT, 2019.

[38] Mohan, Senthilkumar, et al. “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques.” *IEEE Access*, vol. 7, Institute of Electrical and Electronics Engineers (IEEE), 2019, pp. 81542–54. *Crossref*, https://doi.org/10.1109/access.2019.2923707.

[39] V. A. Kumari and R. Chitra, “Classification of diabetes disease using support vector machine,” International Journal of Engineering Research and Applications, vol. 3, no.2, pp. 1797–1801, 2013.

[40] S.K. Dey, A. Hossain and M.M. Rahman, “Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm,” In 21st international conference a. of computer and information technology (ICCIT) IEEE, pp–5, 2018.

[41] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju and H. Tang, “Predicting diabetes mellitus with machine learning techniques,” Frontiers in genetics, vol. 9, pp. 515, 2018.

[42] R. Zolfaghari, “Diagnosis of diabetes in female population of pima indian heritage with ensemble of bp neural network and svm,” Int. J. Comput. Eng. Manag, vol. 15, pp. 2230– 7893, 2012.

[43] H. N. A. Pham and E. Triantaphyllou, “Prediction of diabetes by employing a new data a.mining approach which balances fitting and generalization,” In Computer and Information Science, b. Springer, pp.11–26, 2008.

[44] R. Sanakal and T. Jayakumari, “Prognosis of diabetes using data mining approach-fuzzyc means clustering and support vector machine,” International Journal of Computer Trendsand Technology, vol. 11, no. 2, pp. 94–8, 2014.

[45] M. Maniruzzaman, M.J. Rahman, B. Ahammed, and M.M Abedin, “Classification and prediction of diabetes disease using machine learning paradigm,” Health Information Science and Systems, vol. 8, no. 1, pp. 7, 2020.

[46] R. Sivanesan and K.D. R. Dhivya, “A review on diabetes mellitus diagnoses using classification on Pima Indian diabetes data set,” International Journal of Advance Research in Computer Science and Management Studies, vol. 5, no. 1, 2017.

[47] S. Karatsiolis and C.N. Schizas, “Region based support vector machine algorithm for medical diagnosis on pima indian diabetes dataset,” in 12th International Conference on Bioinformatics & Bioengineering (BIBE), IEEE, pp. 139–1442, 2012.

[48] W. Chen, S. Chen, H. Zhang and T. Wu, T, “A hybrid prediction model for type 2 diabetes using k-means and decision tree,” in 2017 8th IEEE International Conference on Software Engineering a. and Service Science (ICSESS), pages 386–390, 2017.

[49] S. Lekkas and L. Mikhailov, “Evolving fuzzy medical diagnosis of pima indians diabetes and of dermatological diseases. Artificial Intelligence in

Medicine,” vol. 50, no. 2, pp.117– 126, 2010.

[50] S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017, pp.

[51] B.M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEEInternational Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5.

[52] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, ‘Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis’, Procedia Computer Science, vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[53] Y. khoudfi and M. Bahaj, Applying Best Machine Learning Algorithms for BreastCancer Prediction and Classification, 978-1-5386- 4225- 2/18/$31.00 ©2018 IEEE.

[54] L. Latchoumi, T. P., & Parthiban, “Abnormality detection using weighed particle swarm optimization and smooth support vector machine,” Biomed. Res., vol. 28, no. 11, pp. 4749– 4751, 2017.

[55] A. H. Osman, “An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 4, pp. 158–165, 2017.

[56] A. Gulia ,R. Vohra and P. Rani . (2014), Liver Patient Classification Using Intelligent Techniques, International Journal of Computer Science and Information Technologies. 5 (4) : 5110-5115

[57] Y. Kumar, “Prediction of different types of liver diseases using rule based classification model”, Technology and health care: official journal of the European Society for Engineering and Medicine 21(5)DOI:10.3233/THC-130742, August 2013.

[58] Dr. S. Vijayarani & Mr. S. Dhayanand, “ Liver Disease Prediction using SVM and Naïve Bayes Algorithm”, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 4, April 2015

[59] Charleonnan A, Fufaung T, Niyomwong T, Chokchueypattanakit W, Suwannawach S, Ninchawee N. Predictive analytics for chronic kidney disease using machine learning techniques. Manag Innov Technol Int Conf MITiCON. 2016;80–83:2017.

[60] Salekin A, Stankovic J. Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes. In: Proc. - 2016 IEEE Int. Conf. Healthc. Informatics, ICHI 2016, pp. 262–270, 2016

[61] Tekale S, Shingavi P, Wandhekar S, Chatorikar A. Prediction of chronic kidney disease using machine learning algorithm. Disease. 2018;7(10):92–6.

[62] Xiao J, et al. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. J Transl Med. 2019;17(1):1–13.

[63] Almasoud M, Ward TE. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. Int J Adv Computer. 2019;10(8):89–96.

[64] Shrihari K Kulkarni1, K R Sumana2,"Detection of Parkinson’s Disease Using Machine Learning and Deep Learning Algorithms" International Journal of Engineering Science Invention (IJESI) ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726 VOLUME 8 ISSUE:8, (AUG 2021)-Page No :1189-1192.

[65] Yatharth Nakul1 , Ankit Gupta2 , Hritik Sachdeva3,,,” Parkinson Disease Detection Using Machine Learning Algorithms” International Journal of Science and Research (IJSR) ISSN: 2319-7064 SJIF (2020): 7.803 Volume 10 Issue 6, June 2021-Page no 314-318

[66] Wu Wu Wang1 , Junho Lee2 , Fouzi harrou3 and Ying sun4,,,” Early Detection of Parkinson’s Disease Using Deep Learning and Machine Learning” IEEE ACCESS Digital Object Identifier 10.1109/ACCESS.2020.3016062 Volume 8,2020-Page no 147635- 147646.

[67] Talasila Bhanuteja, K. V. Narendra Kumar and K. S. Poornachand, “ Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach”, Article in International Journal of Innovative Technology and Exploring Engineering · August 2021

[68] Rinkal Keniya, Aman Khakharia and Vruddhi Shah, “Disease prediction from

various symptoms using machine learning", July 27, 2020

[69] Samuel A.L. Some studies in machine learning using the game of checkers. IBM J. Res. Dev. 1959;3:210–229. doi: 10.1147/rd.33.0210. [https://doi.org/10.1147%2Frd.33.0210] [Google Scholar] [Ref list].

[70] "The 7 Steps of Machine Learning", towardsdatascience.com. [Online]. Available: https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e [Accessed: Jan. 14, 2023].

[71] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.

[72] Siegel RL, Miller KD, Jemal A. Cancer Statistics , 2016. 2016;00(00):1-24. doi:10.3322/caac.21332.

[73] "Globocan 2012 - Home." [Online]. Available: http://globocan.iarc.fr/Default.aspx. [Accessed: 15-Nov-2022].

Acknowledgement

Foremost, we would like to express our sincere gratitude to our advisor and project guide **Dr. Archana B. Patankar** for the continuous support in the study and research of this project, for her patience, motivation, enthusiasm, and immense knowledge. Her dynamism, vision, sincerity and motivation have deeply inspired us. She has taught us the methodology to carry out the research and to present the research work as clearly as possible. It was a great privilege and honor to work and study under her guidance. We could not have imagined having a better guide and mentor for our BE project. We are really thankful to her.

We would also like to thank our Head of Department (HOD), **Ms. Tanuja Sarode**, for her encouragement, insightful comments, and hard questions. Our sincere thanks also goes to **Ms. Darakshan Khan**, **Ms. Nabanita Mandal** and **Dr. Vaibhav Ambhire**, for offering the much required insights in our data collection and domain understanding. We would like to express our special thanks of gratitude to our principal **Dr. G.T. Thampi** who gave us the golden opportunity to do this wonderful project on the topic "Predictal – A One-Stop Medical Solution for Early Diagnosis of Multiple Diseases" , which also helped us in doing a lot of Research. We are overwhelmed in all humbleness and gratefulness to acknowledge our depth to all those who have helped me to put these ideas, well above the level of simplicity and into something concrete.

Any attempt at any level can't be satisfactorily completed without the support and guidance of parents, friends and closed ones that support you with their prayers, empathy, and a greatsense of humor. Finally, our thanks go to all the people who have supported us to complete the research work directly or indirectly.

Tushar Budhwani

Esha Datwani

Anushree Dutt

Yukta Jain