

Impact of Lifestyle Factors on Life Expectancies in Pakistan



By

Syeda Aiman Farrukh

00000275744

Supervisor

Dr. Seemab Latif

Department of Computing

A thesis submitted in the partial fulfillment of the requirements for the degree of
MS (IT)

In

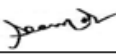
School of Electrical Engineering and Computer Sciences,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

December, 2021.

Approval

It is certified that the contents and form of the thesis entitled "Impact of Healthy Lifestyle Factors on Life Expectancies in the Pakistan " submitted by SYEDA FARRUKH have been found satisfactory for the requirement of the degree

Advisor: Dr. Seemab Latif

Signature:  _____

Date: 09-Dec-2021

Committee Member 1: Dr. Rabia Irfan

Signature:  _____

Date: 09-Dec-2021

Committee Member 2: Pakeeza Akram

Signature:  _____

Date: 13-Dec-2021

Committee Member 3: Dr. Mehdi Hussain

Signature:  _____

Date: 08-Dec-2021

Dedication

I dedicate this piece of work to my amazing parents, Mr. and Mrs. Syed Farrukh Naseem Tirmizi and my biggest support, my siblings, Yumna and Shahzeb. As Daddy liked to say, “Nothing ever happens unless you make it happen”. Together, we made it to the end.

And as Sirius says,


“The ones that love us, never really leave us. You can always find them in here.”

-Sirius Black

Certificate of Originality

I hereby declare that this submission titled "Impact of Healthy Lifestyle Factors on Life Expectancies in the Pakistan" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECs or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECs or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name: SYEDA FARRUKH

Student Signature: 

Acknowledgement

First and Foremost, I would like to exhibit my gratitude to my Research Supervisor, Dr. Seemab Latif whose support, guidance and confidence in me, encouraged me to do better. I am indebted to her for giving me the research freedom and helping me in exploring different paths in Machine Learning and Deep Learning. I would like to thank my mother for pushing me and motivating me throughout this journey. I would also like to acknowledge my friends who supported me and helped me whenever I was stuck somewhere.

Table of Contents

1 Introduction	1
1.1 <u>Lifestyle factors and diseases</u>	1
1.2 <u>Role of artificial Intelligence</u>	3
1.3 <u>Applications of machine learning in health care</u>	5
1.4 <u>Challenges</u>	5
1.5 <u>Research Statement</u>	6
2 Literature Review	7
2.1 Diabetes	7
2.1.1 <u>Factors Contributing to T2DM</u>	8
2.2 Cardiovascular Diseases	9
2.2.1 <u>Factors Contributing to CVD</u>	9
2.3 Metabolic Syndrome – Obesity	10
2.3.1 <u>Factors Contributing to Obesity</u>	10
2.4 <u>Machine Learning and Deep Learning</u>	11
2.5 <u>Applications of Machine Learning in the identification of lifestyle diseases</u>	12
3 Methodology	18
3.1 <u>Proposed Methodology</u>	18
3.2 <u>Applications of Machine Learning Algorithms</u>	20
3.3 <u>System Architecture</u>	21
3.4 <u>Detection of Diabetes Mellitus Type 2 (T2DM)</u>	21
3.4.1 <u>Dataset</u>	21
3.4.2 <u>Preprocessing of NHANES dataset</u>	24
3.4.3 <u>Splitting the dataset</u>	25
3.4.4 <u>Algorithms and Hyper parameters</u>	26
3.5 <u>Detection of Cardiovascular Diseases</u>	27
3.5.1 <u>Dataset</u>	27
3.5.2 <u>Splitting the Dataset</u>	29
3.5.3 <u>Algorithms and hyper parameters</u>	29
3.5.3.1 <u>Deep Learning Model and Hyper Parameters</u>	30

<u>3.6 Detection of Metabolic Syndrome – Obesity</u>	31
<u>3.6.1 Dataset</u>	31
<u>3.6.2 Preprocessing the Dataset</u>	32
<u>3.6.3 Splitting the Dataset</u>	32
<u>3.6.4 Algorithms and Hyper parameters</u>	32
<u>3.6.5 Application of Feed Forward Neural Network</u>	34
3.7 Evaluation metrics.....	35
<u>4 Experiments and Results</u>	38
<u>4.1 Experiment 1 – Detection of Diabetes Mellitus – type 2</u>	38
<u>4.1.1 Heat Map for NHANES Diabetes Dataset</u>	39
<u>4.1.2 Results</u>	39
<u>4.2 Experiment 2 – Detection of Cardiovascular Diseases – CVD</u>	42
<u>4.2.1 Heat Map for Cleveland Dataset</u>	42
<u>4.2.2 Results</u>	43
<u>4.2.3 Application of Neural Network</u>	45
<u>4.3 Experiment 3 – Metabolic Syndrome – Obesity</u>	47
<u>4.3.1 Heat Map for Obesity Dataset</u>	48
<u>4.3.2 Results</u>	49
<u>4.3.3 Endpoint Generation</u>	50
<u>5 Conclusion</u>	51
<u>5.1 Future Works</u>	51
<u>6 References</u>	52

List of Figures

Figure Number	Figure Description	Page Number
Figure 1.1	Incidence of lifestyle diseases in Karachi. Taken from [5]	3
Figure 2.1	A brief history of machine learning	11
Figure 2.2	Breakdown of diabetes in patients. Taken from [26]	13
Figure 2.3	System Architecture. Taken from [30]	14
Figure 2.4	Cluster formation after application of PCA. Taken from [31]	15
Figure 3.1	Proposed System	19
Figure 3.2	Branches of Machine Learning	20
Figure 3.3	Machine learning hierarchy taken from [40]	21
Figure 3.4	System Architecture Diagram	21
Figure 3.5	Missing Values in the dataset	25
Figure 3.6	Hyper-parameters for the deep learning model	30
Figure 3.7	List of hyper parameters for neural network	34
Figure 4.1	Heat Map for NHANES dataset	39
Figure 4.2	AUC-ROC analysis for different models for the identification of diabetes	41
Figure 4.3	Heat map for the Cleveland dataset	42
Figure4.4	AUC-ROC analyses for different models on Cleveland dataset.	44
Figure 4.5	Model Accuracy for Feed Forward Neural Network	45
Figure 4.6	Model precision for feed forward neural network	46
Figure 4.7	Boxplot for neural network on obesity dataset	46
Figure 4.8	Heat map for the obesity dataset	48

Figure 4.9	Error rate for different values of K	49
Figure 4.10	Prediction of Obesity in a patient	50

List of Tables

Table ID	Table Description	Page Number
Table 3.1	Dataset Description for NHANES	22
Table 3.2	Hyper-parameters for NHANES Dataset	26
Table 3.3	Dataset Description for Cleveland Data	28
Table 3.4	Hyper-parameter tuning for Cleveland dataset	29
Table 3.5	Dataset Description for Obesity data	31
Table 3.6	Hyper parameter tuning for obesity Dataset	33
Table 3.7	Confusion Matrix	35
Table 4.1	Results for Diabetes identification models	40
Table 4.2	Results of different models on the Cleveland dataset for the identification of cardiovascular diseases	43
Table 4.3	Results of different models for the identification of Obesity	49

Abstract

The main idea of my research work is to perform data analytics on the data collected by monitoring the daily life styles of people. These analytics will help identify and timely detect any disease.

With the recent advancements in sensor technologies, wearable devices, Internet of Things and wireless communication the research on mobile health care and monitoring systems has reached new levels.

This system enables continuous monitoring of patients' physiological and health conditions by sensing and transmitting measurements like heart rate, physical activity, blood pressure, blood sugar levels and nutritional values.

My thesis idea revolves around collecting the real time data of people, cleaning it and then performing medical big data analytics on it. This research will help in diagnosis and treatment of patients with chronic diseases like depression, hypertension and diabetes. This research can also be used for monitoring the health of elderly people.

After the collection of real time data for analysis, the next step will be Data Fusion. In these systems, the data collected is then integrated with informational databases, knowledge based systems and other similar sources. The basic purpose of Data Fusion is to integrate all types of information in a uniform format which can be used for computational purposes.

In the next step, I will be performing data analytics on the processed data using different big data tools and algorithms. In this study, I will be merging health care with technology to create smart and cost effective solution for early stage detection and prevention of diseases. My thesis work includes acquiring the data, processing it, storing it and then through advance algorithms of information retrieval and data mining perform data analytics on the preprocessed dataset.

The daily routine factors that will be the main focus of my work are:

- a) Sleep
- b) Nutrition
 - Fruit Intake
 - Water Intake
- c) Physical Activity
- d) Heart rate

The health of patients can be now monitored through sensing wearable. The data collected through these wearable can be further utilized to extract patterns and form associations which can be used for early diagnosis of many diseases.

Most of these illnesses are a result of our daily life routines. Making certain amendments in our routines can lead us towards a healthy life. The work of my thesis revolve around eliciting certain patterns from a real time data set of people's daily routine and identifying the factors that can lead to certain illnesses or could disrupt the mental and physical health.

CHAPTER 1

Introduction

1.1. Lifestyle Factors and diseases

Lifestyle is a manner in which an individual or a group of people follow their daily routines. The choices we make are characterized by geographical, socioeconomically, political, cultural, or religious grounds. According to the World Health Organization, an individuals' health and quality of life are correlated for up to 60% [1].

Lifestyle factors are the modifiable habits and ways of life that can greatly influence overall health and well-being. Not following a good and recommended lifestyle can lead to many health issues and non-communicable diseases (NCDs). The NCDs encompass the diseases that are non-transmittable from one individual to another. The risk factors of prolonged exposure to smoking, unhealthy diet, and physical inactivity are similar to having a lifestyle disease [2].

The most common lifestyle methods that affect health are [3]:

- Diet
- Sleep
- Water Intake
- Physical activity
- Smoking/ alcohol consumption

An imbalance in these lifestyle methods and not complying with a healthy routine can lead to many non-communicable diseases like

- Diabetes
- Hypertension
- Cardiovascular diseases
- Mental health issues
- Sleep disorders, accidents, injuries
- Cancer

These diseases are emerging as a serious threat to public health and causing a distressing impact on mortality, morbidity, and the overall economy. The lower-income countries and the developing countries are the ones that are most badly affected by this wave of NCDs. World Health Organization states that Non Communicable diseases are responsible for 71% of death globally – about 41 million annually die from these diseases. 77% of these deaths are in low and middle-income countries. Annually more than 15 million people aged between 30-69 years, die while 85% of these premature deaths occur in low and middle-income countries [2]

Cardiovascular disease is the leading cause of death claiming 17.9 million deaths annually, followed by cancer – 9.3 million deaths, pulmonary diseases 4.1 million and diabetes 1.5 million [2]

As Pakistan is a low-income developing country, Non-communicable diseases are among the top 10 causes of morbidity and mortality. 25% of deaths that occur here are due to these NCDs. A sedentary lifestyle, unhealthy dietary choices, and not enough physical activity have caused diabetes type 2 and cardiovascular diseases to rise exponentially in Pakistan. According to a survey Pakistan stands 6 in line to be the most diabetic country. These diseases are also adding a strain to the already strained economy and increasing the Burden of Disease (BoD) [4]. Fig 1. Shows the results of a study held in Karachi.

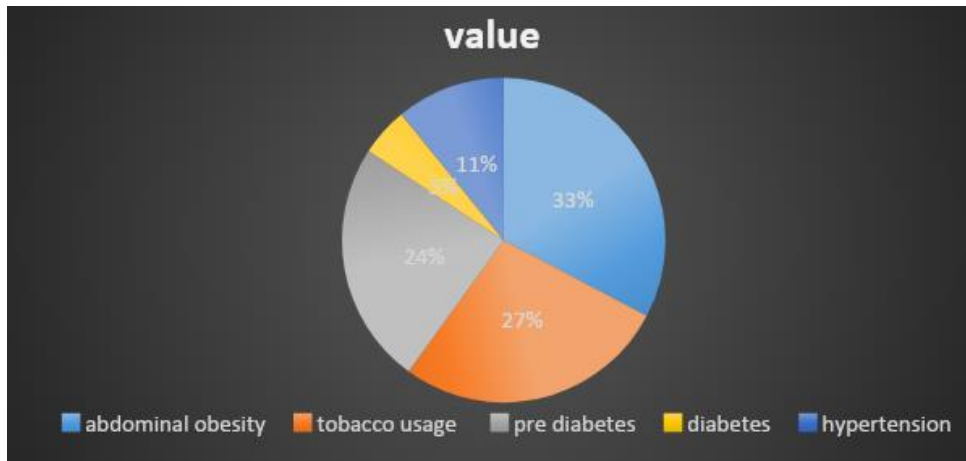


Figure 2.1 Incidence of lifestyle diseases in Karachi. Taken from [5]

According to a survey by Global Youth Tobacco, among school-going children aged between 13-15 years in Pakistan, nearly 11% of them are smokers [5].

A community-based survey done in Karachi showed physical inactivity among 60% of individuals. A food study was done in Karachi, Lahore, and Quetta among school-going children showed 80% of them were following unhealthy diets. Owing to unhealthy lifestyles, the prevalence of hypertension among individuals aged 15 years and above was recorded at about 19%. The issue of hypertension occurred in about 33% of individuals of age 45 years and above. [4]

Owing to the negligence of preventable causes of death, we lose millions of people. These diseases are emerging as a serious threat to public health and causing a distressing impact on mortality, morbidity, and the overall economy. The lower-income countries and the developing countries are the ones that are most badly affected by this wave of NCDs.

1.2 Role of Artificial Intelligence

Artificial intelligence works to make the machine capable enough to make real-time human-like decisions. When a machine can perform functions that are associated with how a human mind works in terms of learning and problem solving – that is referred to as artificial intelligence. Machine learning is a subfield

of artificial intelligence. Since 1950, when it was introduced, till now tremendous efforts have been put into this field and machine learning is now being excessively used in almost every aspect of our daily lives [6].

With massive datasets and a surge in computing power, another subfield of machine learning has surfaced, known as deep learning. Based on the concepts of neurons in a human brain, deep learning has shown much more promising results on large datasets as compared to its shallow learning models. A huge number of fields have been benefited from the deep learning concept especially in the medical and health care sector [7].

Data mining is another subfield of artificial intelligence where datasets are mined and unknown patterns or associations are discovered. Data mining is the process of finding meaningful, interesting, valuable insights in large datasets [8]. With the usage of sensors, electronic health records the health care data is increasing exponentially however, it is still being underutilized. The application of data mining and machine learning tools on these datasets can help doctors in making clinical decisions [9].

As the incidence of lifestyle diseases in Pakistan is increasing exponentially, drastic measures need to be taken. The application of machine learning and data mining on the publically available datasets can yield meaningful information which can be used to make decisions at a much faster pace. Moreover, we can even create models that can pre-identify the onset of any disease.

Lifestyle diseases are much dependent on the lifestyle of an individual. By monitoring the lifestyle of an individual, a machine learning model can predict whether he/she can fall prey to certain lifestyle diseases. Sensors, electronic health records, and public health repositories can help in creating such models.

With the help of sophisticated algorithms and techniques, the healthcare sector is now moving towards the preventive medicine concept. This concept allows the pre-identification of diseases and helps to nullify the cause of the disease at a much earlier stage.

By monitoring the daily lifestyle of the user and keeping a record of his routine, every time the user falls short in some aspect of a healthy lifestyle or does not comply with the recommended values of certain factors, the application will notify the user and remind him to fulfill the certain criteria. This application can also help doctors identify the diseases in a much shorter period. The co-morbidity of different diseases can also be identified by the data analytics algorithm. This will reduce the cost of medical facilities and the time that is taken in conventional approaches [10].

1.3 Applications of machine learning in healthcare

The sheer volume and variety of medical data make it an ideal candidate for the application of machine learning. Healthcare data exists in many formats including text data, numeric data, and image data. All over the world, thousands of researchers are now using healthcare data along with machine learning models for prediction analysis, diagnosis, and prognosis of diseases. These algorithms are being used for the prediction of many diseases including diabetes, hypertension, cardiovascular diseases, tumors, cancer, depression, etc.

Munira et al. [11] have done a review of all the machine learning algorithms that are being applied to different medical datasets for the prediction or diagnosis of multiple diseases.

Over the last couple of years, the interest in machine learning for the diagnosis of diabetic retinopathy [12], diagnosis and prognosis of lymph nodes for the cases of breast cancer [13], detection of autism [14] and depression [15], insulin-producing and detection of cardiovascular diseases has increased immensely.

The application of machine learning algorithms in the healthcare sector will not only help in the diagnosis and prognosis of disease but will help in cost reduction is often the diagnosis of diseases taken an exhaustive amount of testing and time. This will help in cutting down both: cost and time

1.4 Challenges

A model can only be as good as the data it has been trained upon. Although the medical data exists in abundance, usually it contains a large of missing, incomplete, biased, imbalanced, or wrongly labeled data.

Machine learning algorithms can create a bias in the model due to missing, incomplete, underestimation, sample dataset size, class imbalanced data, or even some measurement error. These biases can contribute to socioeconomic disparities in the health care sector [17].

1.5 Research statement

Lifestyle diseases can be avoided if proper measures are taken timely. With rapid, reliable, and efficient decision-making, the mortalities due to these diseases can be avoided. The main challenge however is creating a screening tool that is cost-effective, reliable, and fast. Such a tool can expedite the decision-making process without the need for costly blood work and labor. Developing countries can especially make use of this tool where although the economies are strained yet the incidence of lifestyle diseases is on the rise. World Health Organization also advises creating simple strategies to identify individuals who have risk to develop any lifestyle disease.

As Pakistan is a developing country, lifestyle diseases are one of the highest causes of mortality and morbidity here. Not only are these diseases responsible for the death of a huge number of people every year, but these ailments are also causing an additional strain on the already strained economy by doubling the burden of disease (BoD).

As research and various studies suggest there exists a strong relationship between lifestyle factors and comorbidity [18], In our research, we propose a machine learning framework that is capable of detection of the lifestyle disease based on a patients' invasive and noninvasive parameters. Our major focus is on the monitoring of lifestyle and making very little use of invasive parameters. We believe by focusing on just the lifestyle, lifestyle diseases can be avoided to a great extent. The key problems in current lifestyle diseases detection models are high computing and cost and low accuracy. Thus, we want to develop a system that has high accuracy and uses non-invasive parameters to identify diseases.

CHAPTER – 2

Literature Review

2.1 Diabetes

Diabetes Mellitus (DM) is a disease in which the body loses its ability to normalize the blood glucose levels, the Diabetes Mellitus can be divided into 2 categories: Diabetes Mellitus 1 and Diabetes Mellitus 2. In type 1 DM (T1DM) due to an abnormality, the immune system starts attacking the insulin-producing pancreatic cells. This leads to a complete deficiency of insulin secretion. Whereas in Type 2 DM (T2DM), the resistance on body cells to insulin is increased. This starts to limit insulin secretion.

T2DM progresses over time and begins by first showing the pre-diabetic symptoms and then if not properly attended to, it makes the patient a diabetic. As the body is not able to regulate the glucose levels, continuous exposure to elevated blood sugar levels can cause major damage to small and large blood vessels which in the long run can lead to cardiovascular, neuropathic, nephropathy, and retinopathy issues. Obesity is the number 1 risk factor in the American Diabetes Association standards of medical care, upon which T2D status in asymptomatic patients is based [19].

Of both DM types, 90% of people suffering from this disease are the ones who have type 2 DM.

According to International Diabetes Federation, one in every 11 adults aged between 80-79 years suffered from this disease in 2015. That accumulated to a total of 415 million adults. This number is expected to rise to a staggering 642 million by 2040 with the strongest impact in countries of low income to middle-income levels.

According to a study by Global Burden of Disease in 2013, DM was identified as the ninth leading cause of reduced life expectancy. One of the reasons for such a large number of deaths by this lifestyle disease is that it remains undiagnosed for a long period. In a study, it was estimated that globally 45.8% of all DM cases remain undiagnosed. And this leads to the occurrence of major complications in patients.

A sad reality is that the number of patients having diabetes has exponentially increased over time with its strongest impact on the developing regions. Only between 2010 and 2030, a 20% increase in patients having this disease in developed countries and a huge 69% increase in the developing countries is expected to occur.

Especially in the Asian specific region, china, India, and Pakistan are the countries that are majorly hit with this disease. As compared to the western countries, the individuals here are exposed to DM at much a younger age. Increased weight gain is closely associated with more chances of Diabetes Mellitus type 2 (T2DM) [20].

Following a healthy lifestyle has been proven to reduce the risk of falling prey to diabetes and cardiovascular diseases. In a survey done by the Indian Diabetes Prevention Program, by adapting to a healthier lifestyle, the incidence of diabetes among patients was reduced to 28.5%.

2.1.1 Factors contributing to T2DM

According to World Health Organization, to prevent diabetes the most important factors that contribute to being a diabetic are:

- Sleep
- BMI \geq 24KG/m²
- Age > 45
- Not doing adequate physical activity
- Family history of Diabetes Mellitus
- Imbalance in cholesterol levels

- Occurrence of hypertension or other cardiovascular issues
- Not consuming enough nutrients

Of the 3 primary risk factors of family history, age, and obesity, obesity is the only modifiable cause [21] and is hence a major target of T2D prevention.

2.2 Cardiovascular Diseases

Cardiovascular diseases are the leading cause of death worldwide. Only in 2019, around 17.9 million people lost their lives to CVD. This is 32% of total worldwide deaths. 85% of these deaths were due to heart attack and stroke. These diseases are most common in low and middle-income countries. These diseases usually occur in the fifth decade of human life. However, research proves that CVD risk factors during childhood potentially increase the chances of incidence of this disease in adulthood.

2.2.1 Factors contributing to CVD

World Health Organization identifies the following lifestyle style patterns as the potential risk factors apart from the clinical risk factors [22]:

- Tobacco use
- Unhealthy diet
- Obesity
- Physical Inactivity
- Alcohol abuse

The clinical risk factors include the combined influence of healthy diet and active lifestyle on cardiovascular disease risk factors in adolescents:

- Elevated levels of total cholesterol (TC)
- Triglycerides (TGs)
- Insulin resistance
- Blood pressure

- Total and central body fat
- Low levels of high-density lipoprotein cholesterol
- Cardiorespiratory fitness

2.3 Metabolic Syndrome – Obesity

Obesity is another lifestyle disease that is increasing exponentially throughout the world. The incidence of Obesity is also increasing rapidly in Pakistan. According to Data World obesity, 10 out of 6.5 adults and 6 Out of 11 children are at risk of becoming obese. The incidence of childhood obesity is also increasing at An alarming pace [23].

2.3.1 Factors contributing to Obesity

The major contributors of metabolic syndrome are [24]

- Sedentary lifestyle
- Inadequate physical activity
- Poor dietary habits like overconsumption of salt, sugar, and high saturated fat

Obesity and overweight are the major lifestyle diseases and are potential risk factors for other non-communicable diseases including:

- Cardiovascular diseases
- Diabetes type II
- Hypertension
- Depression
- Chronic Obstructive Pulmonary Disease

Obesity also causes the premature occurrence of other lifestyle diseases including hypertension, diabetes, and cardiovascular diseases.

2.4 Machine Learning and Deep Learning

Fig 2. presents a brief timeline of the evolution of machine learning.

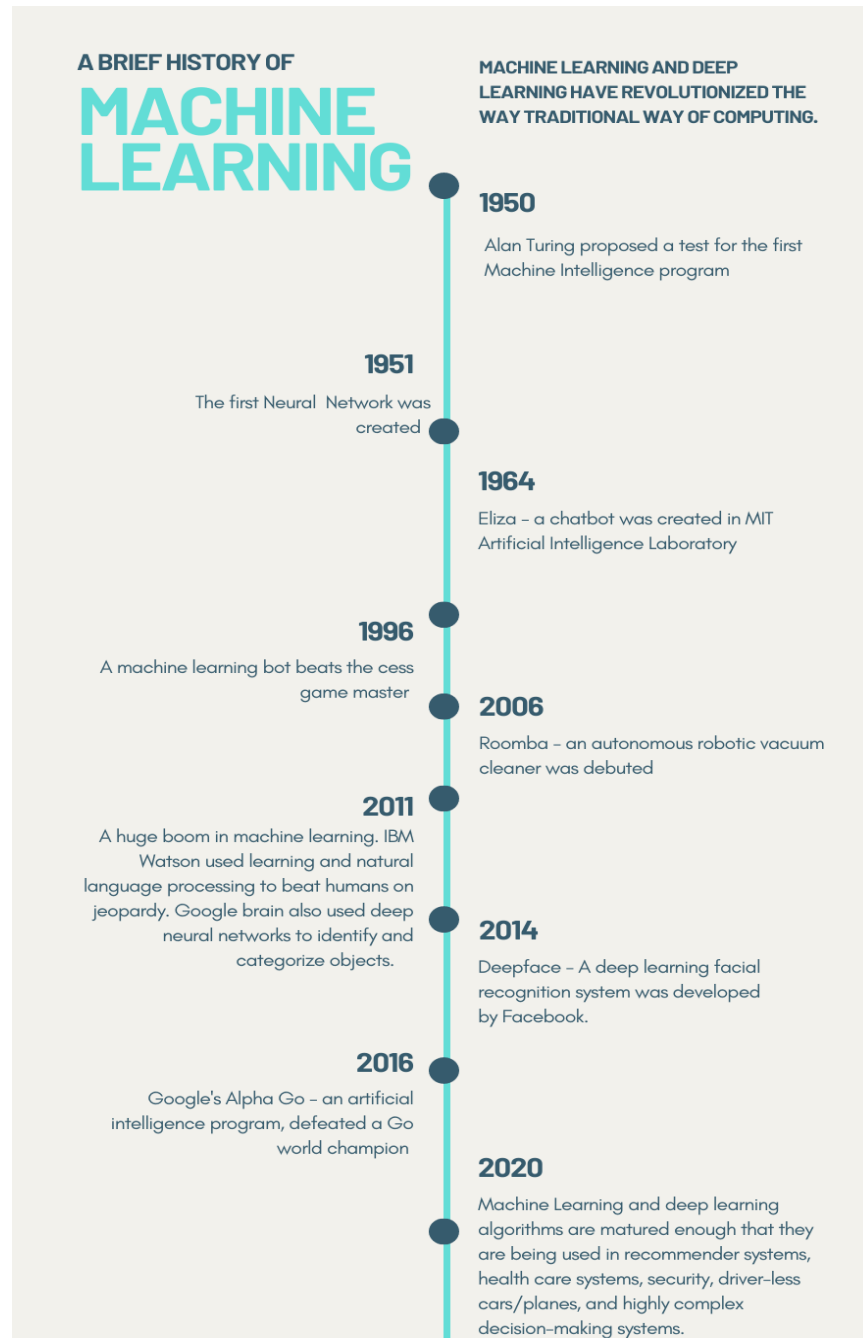


Figure 2.1 A brief history of machine learning

2.5 Applications of machine learning in the identification of lifestyle diseases

Diabetes and cardiovascular diseases are one of the leading causes evolution of death worldwide and usually are correlated. Dinh. Et al [25] have proposed a model that uses survey data as well as clinical data to create a model for the prediction of cardiovascular diseases, diabetes, and pre-diabetes. Their model suggests that waist size, age, leg length, sodium intake, and self-reported weight as the top 5 indicators for diabetes. Attributes like systolic blood pressure, diastolic pressure, self-reported weight, the occurrence of chest pain, and age as the top contributors to cardiovascular diseases. The proposed model is trained on the NHANES data set and uses clinical and self-reported attributes. The model gives the highest AUC-ROC curve when trained on clinical data using the Extreme Gradient Boosting algorithm

Pei. Et al [26] presented a model for the early diagnosis of diabetes using the annual physical examination reports of 4,205 adults in Sheng Jing Hospital, China. The dataset contains parameters including age, gender, BMI, hypertension, family history of diabetes, physical activity, stress, and salty food preference. The model is trained on several classifiers including J48, AdaBoost, SMO, Bayes Net, and Naïve Bayes. Their models show that the decision tree classifiers show the highest accuracy and indicates, age as the most significant feature followed by family history of diabetes, work, stress, BI, salty food preference, hypertension, gender, and any history of cardiovascular disease. The following figures show the breakdown of diabetic patients using the decision tree classifier.

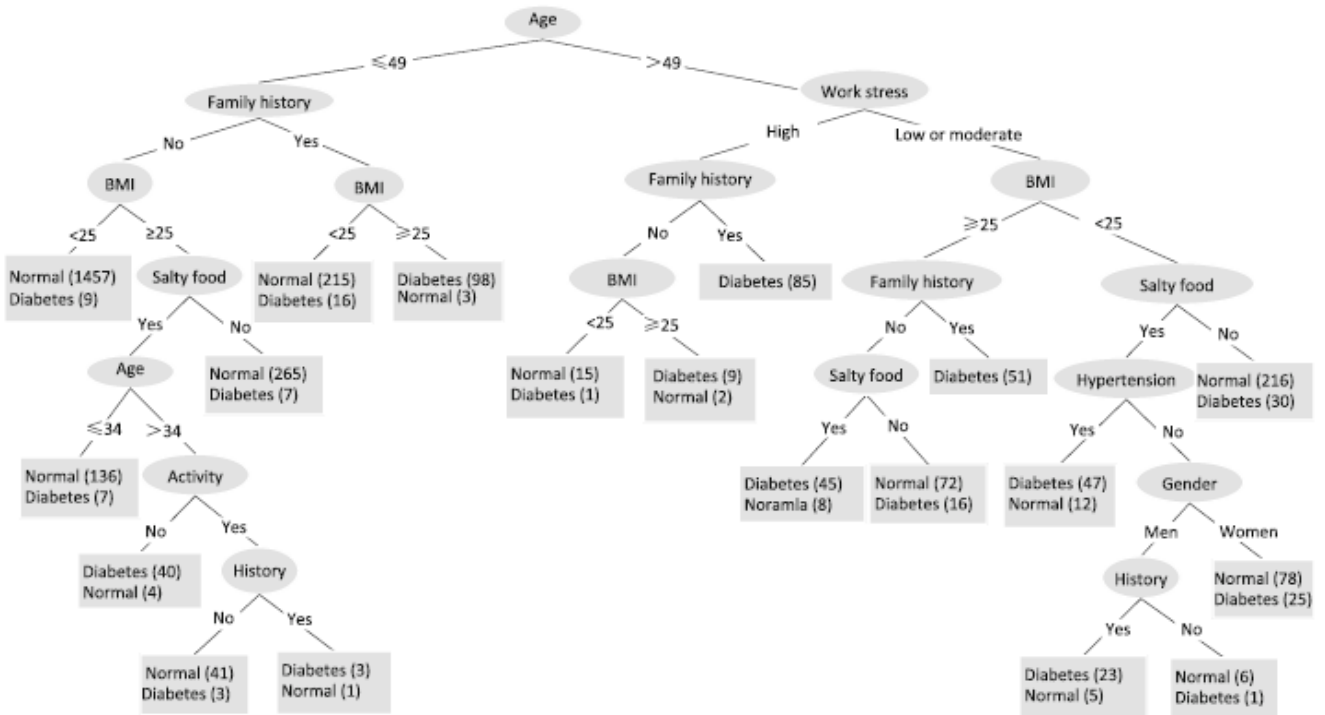


Figure 2.2 Breakdown of diabetes in patients. Taken from [26]

Haq et al [27] proposed a diabetic detection model using clinically obtained data from Frankfurt Hospital Germany. The authors have proposed a filtering model using the decision tree classifier. They have also trained the model using random forest and Ada boost. For validation they have a cross-validation method and the performance of the model is evaluated using measures like accuracy, specificity, sensitivity, F-1 Measure, and ROC curve.

The one drawback of this paper is that it is solely based on clinical data and does not account for the daily lifestyle factors. Owing to the vast amount of Electronic Health Records, researchers are widely using these publically available datasets for the early prediction of diseases. With the implementation of appropriate Machine Learning algorithms, the health care providers can get valuable insights that help them in decision making and even forecasting [29].

One of these publically available datasets that are used extensively for research is the Cleveland heart disease dataset. It is stored in a UCI machine learning repository [28]. The Cleveland heart disease dataset is one of the most accurate, complete, and clean data. That is why many researchers have used this dataset to generate their Machine Learning models.

Amin et al. [29] proposed a machine learning model for the prediction of heart diseases using the Cleveland and stat log dataset. The model is trained by seven different classifiers including K Nearest Neighbor, Decision trees, Naïve Bayes network, Logistic regression and Support Vector Machine, and a voting model based on logistic regression and naïve Bayes algorithm. The model trained on the voting classifier gives the highest accuracy. One of the drawbacks of this paper is that the authors have not used cross-validation or bootstrap techniques. No feature engineering has been done. HDPM – Heart Disease Prediction Model is a clinical decision support system proposed by Fitriyani et al. [30] the proposed model is based on Cleveland and stat log datasets. DBSCAN is used for outlier detection and SMOTE-ENN for data balancing. The model is trained on XGBoost to achieve optimal results. The purpose of this system is to create a clinical decision support system that can be used to filter out the patients who have a high prospect of some cardiovascular disease for prompt treatment. The figure shows their proposed architecture

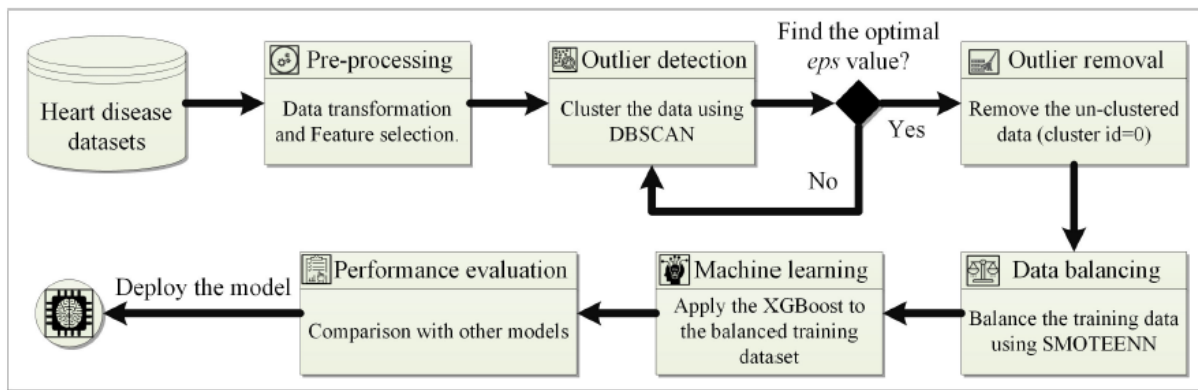


Figure 2.3 System Architecture. Taken from [30]

Naheeda et al [31] proposed a cost-effective cardiovascular disease prediction model based on data collected through a medical camp held at the National University of Science and Technology. The model is based on data of 161 individuals and consists of noninvasive and non-clinical factors. The model uses the daily lifestyle factors including age gender, BMI, waist, diet, mental health, stress, physical activity, and sleeping patterns. The authors have used the QRISK risk estimator score for calculating the probability. A drawback of this paper is that the dataset is of very limited size thus more prone to over fitting. Moreover, the application of algorithms like XGBoost, Random forest can yield more accurate results. Principal Component Analysis has been used for dimensionality reduction. The authors have identified the following clusters as a result of PCA.

z

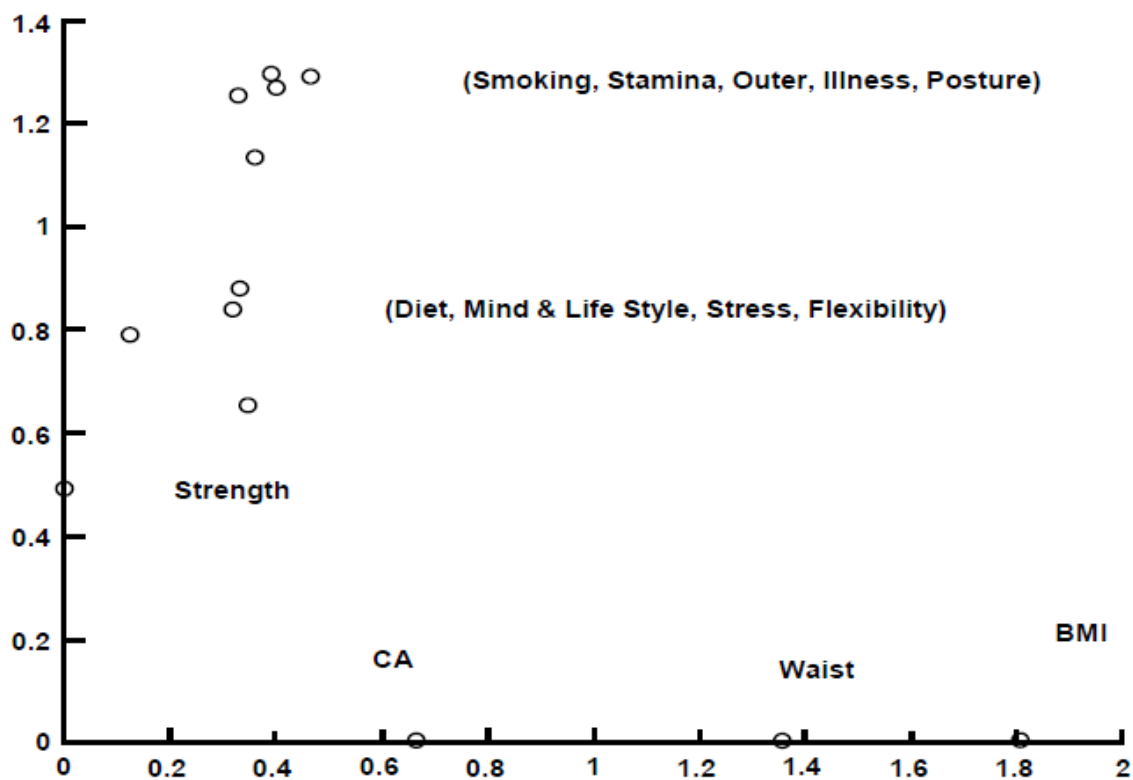


Figure 2.4 Cluster formation after application of PCA. Taken from [

As the diagnosis of heart disease is complex, [30] have proposed an ensemble method based upon feature fusion approaches. They have collected the data through sensors as well as the medical records of patients. In [31], the authors have proposed a hybrid approach by combining classifiers like NN, DT, SVM, and Neural Networks. Their proposed approach gives an F-Measure of 86.8%.

Another aspect of using machine learning in the field of healthcare is to identify the already susceptible patients. This can help in medical cost reduction as high-risk patients require expensive treatments. Now researchers are actively working on predictive and preventive measures for non-communicable diseases [32]. In the paper [33], the authors have proposed a disease prediction model using different classifiers and feature extraction algorithms. The best performance is given by the voting classifier.

In the paper [34], the authors have proposed a neural network approach for the detection of heart disease using the Cleveland dataset. They have used back-propagation and logistic regression to increase the performance of their model. In the paper [35], the authors have presented a novel Multi-Layer- Pi Sigma Neuron Model for the identification of this disease. For data processing, they have used PCA, LDA, and data normalization approaches.

All of these works have used some form of performance metrics. In our proposed method, we have tried to create a model that is simple, easy to use, and gives the highest accuracy.

As obesity is a worldwide health issue, researchers are using multivariant and diverse data including biomedical data, behavioral data, sensor data, and longitudinal data to create effective models for the detection of such issues. In a study [38] the authors have developed an app Feedforward. The app collects data through the pedometer sensor in a smartphone. Any missing values in the data are replaced by the average values. The authors have used Linear Regression and Random Forrest as baseline classifiers and Long Short Term Memory network for the prediction model. The collected data features include gender,

age, height, disease, step count, blood pressure, blood sugar, weight, heart rate, activity pace, education, communication, and weather information.

This paper lacked in covering the daily lifestyle factors including dietary habits, fruits and vegetable consumption, and any family history of obesity.

Zheng et al [39] used the 2015 Youth Risk Behavior Surveillance System (YBRFS) and applied logistic regression, k weighted the nearest neighbor, improved decision tree, and artificial neural network. The YRBFS focuses on high school students. They have used nine health-related factors for the survey to make their prediction model. Their IDT model gave an accuracy of 80.23% with a specificity of 99.44%. The weighted KNN models gave an accuracy of 88.82% and a specificity of 90.74% and the ANN model was able to achieve, an accuracy of 84.22% with a specificity of 99.46%.

CHAPTER 3

Methodology

3.1 Proposed Methodology

The main aim of our research work is to pre-identify and detect major lifestyle diseases. We have worked with multiple datasets for the identification of diabetes mellitus type 2 (T2DM), cardiovascular diseases (CVD), and Metabolic syndrome (Obesity) in populations close to Pakistanis in demographics and socioeconomic backgrounds.

In our research work, we have collected data from multiple online repositories. We have gathered data from the UCI Machine Learning repository and the National Health and Nutrition Examination Survey (NHANES). To achieve the optimal results we have used the benchmarked datasets. The NHANES includes questions about demographics, lifestyle, dietary habits, and socioeconomic conditions.

We have used Pearson correlation and Heat Maps to find relations among different data attributes. After feature selection, we have split our data using a 70 30 ratio. 70% of the data is used in the training phase and 30% of the data is used in the testing phase. We have also implemented a 10 fold cross-validation technique. For the performance evaluation of our model, we used various evaluation metrics for the classifiers. Our proposed system works in five different stages:

1. Data collection
2. Data Preprocessing
3. Feature Extraction
4. Data partitioning and K-Fold Cross-Validation
5. Performance evaluation

Fig. 1 shows our proposed system architecture. The main idea of our system is to make a prediction model for lifestyle diseases.

The data preprocessing steps include data cleaning, removing noisy data, data anomalies, filling out any missing data, data normalization, and data encoding. After doing a correlation analysis we have also done feature extraction and split the data for training and testing.

The post-processing steps include creating visuals, pattern evaluations, pattern selection, creating models, and performance evaluations.

Our proposed system uses supervised machine learning algorithms to identify the diseases in patients by using small, medium, and large publically available datasets. Supervised learning algorithms are used to solve: regression problems – where a numerical value is predicted and classification problems – where a prediction is made.

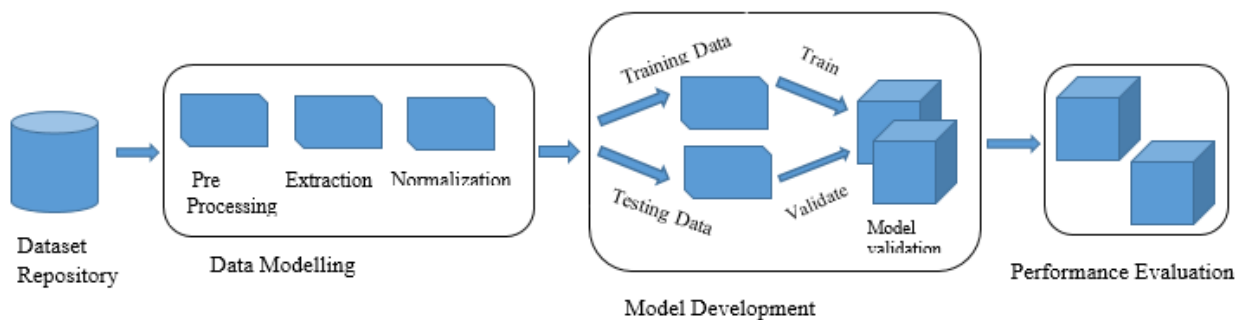


Figure 3.1 Proposed System

We have applied multiple preprocessing techniques to our data to remove misleading data, fill out the incomplete data and select the required features. We have applied multiple machine learning models including K Nearest Neighbor (kNN), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Gaussian Naïve Bayes (NB), Support Vector Mache (SVM), Boosting (XG Boost), Gradient Descent, Voting Classifier, Multi-Layer Perceptron Neural Network (MLP) and Feed Forward Neural Network (NN).

3.2 Application of machine learning algorithms

Machine learning algorithms can be broken down into supervised, unsupervised, and reinforcement learning algorithms. 70% of healthcare machine learning models work on supervised learning algorithms. Numerical and textual healthcare data usually makes use of the supervised algorithms. We have used different machine learning algorithms, deep learning algorithms, and ensemble methods for accurate and real-time results.

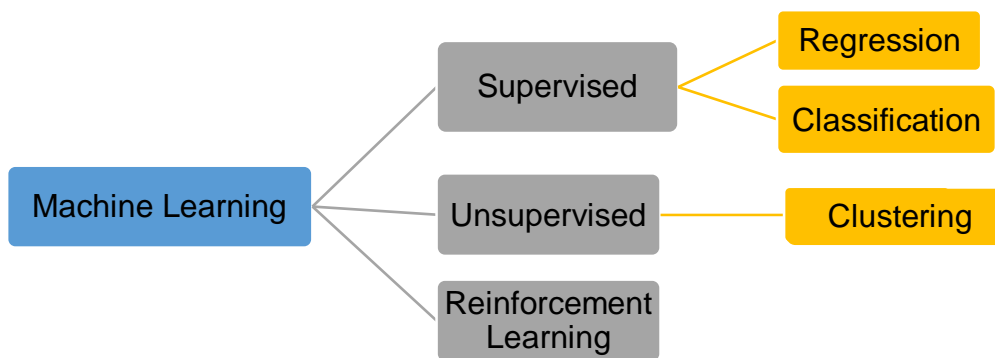


Figure 3.2 Branches of Machine Learning

Supervised learning algorithms refer to learning a function that maps an input to an output based on sample input-output pairs. These algorithms use a labeled and structured dataset for learning purposes. It's comprised of target values or predictions on the historical data which the model learns.

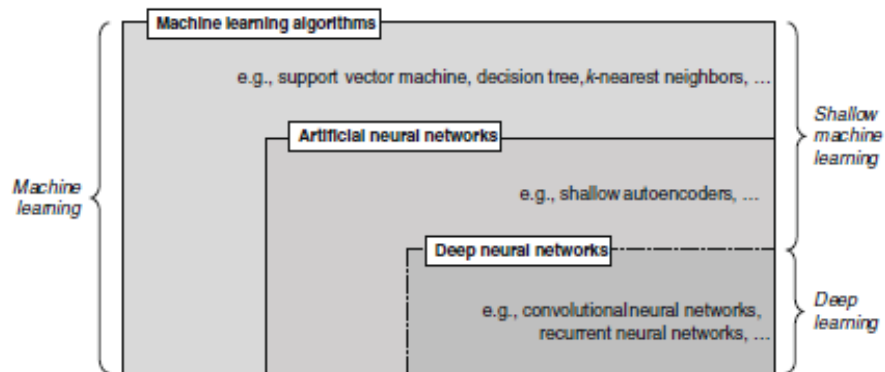


Figure 3.3 Machine learning hierarchy taken from [40]

3.3 System Architecture Diagram

Fig. 9 depicts our proposed system architecture.

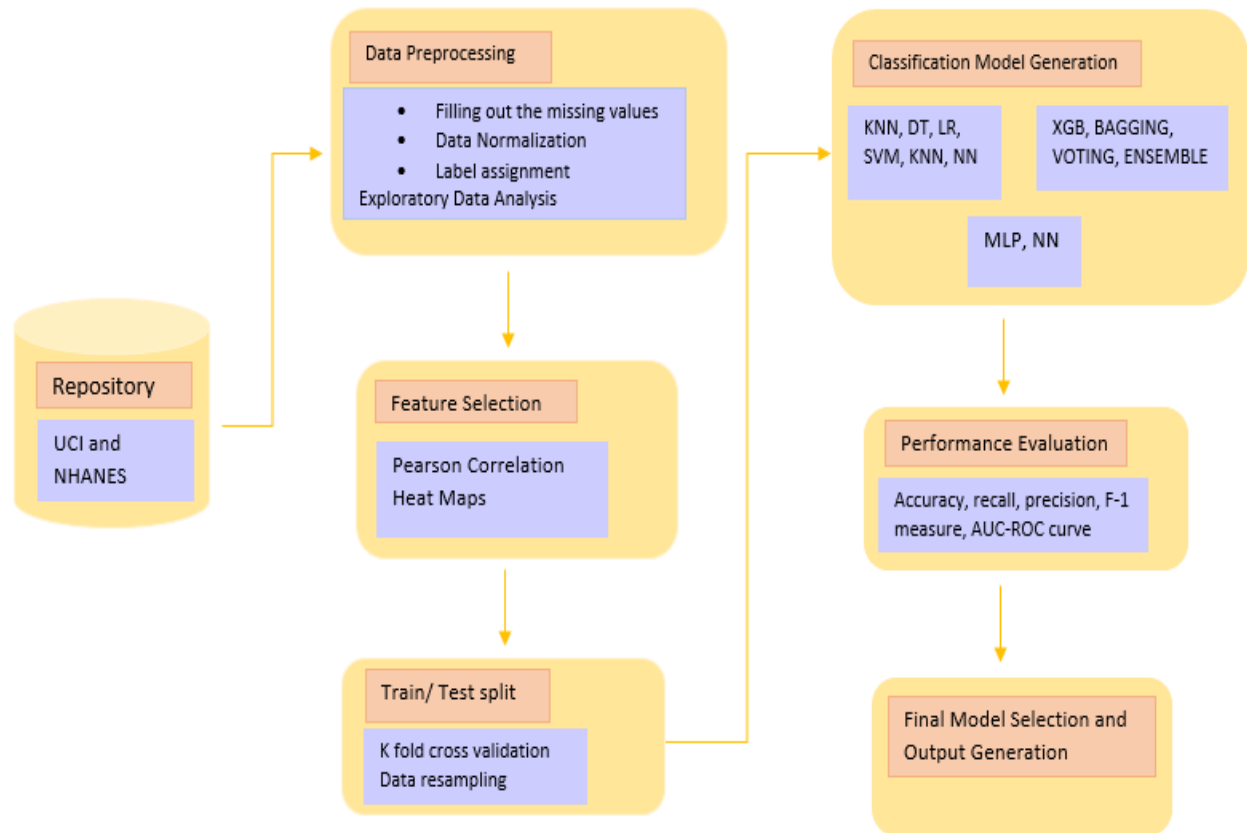


Figure 3.4 System Architecture Diagram

3.4 Detection of Diabetes Mellitus Type 2 (D2TM)

3.4.1 Dataset

The dataset that we have used for the detection of diabetes is collected from the National Health and Nutritional Estimation Survey of 1999-2004 [41]. It contains 17 columns and 5,514 records. This dataset covers the daily lifestyle of individuals, their socio-economic conditions, demographics, and educational levels. A detailed description of data attributes is given in table I. the following points represent some key points about the data.

- a) The target variable is “Status” which indicates the presence or absence of Diabetes. 0 denotes the absence of diabetes whereas 1 denotes the presence of diabetes.
- b) The individuals considered are both males and females aged anywhere between 0 – 150 years.
- c) The starting age varies for each attribute
- d) Each attribute has a code or a value to represent the individuals answer to a particular answer
- e) .A dot Sign indicates a missing value. Almost every attribute has some missing value.
- f) The attribute INDHHINC - annual household income is divided into 13 categories depending upon the income where the value 77 represents the individuals who refused to answer this question and the value 99 represents that the individuals didn’t know the income.
- g) For the attribute, MCQ250 A – Any family member who has diabetes; only his blood relatives (grandparents, parents, brothers, sisters) are considered.

Table 3.2 Dataset Description for NHANES

Data Attribute	Attribute Description	Values
ALQ120Q	Frequency of alcohol consumption in 1 year	0 for none No. of drinks if yes
BMXBMI	BMI of an individual	Values between 11.49 – 66.44
BMXHT	Standing height in cm	Values between 81.8 – 201.3 cm. (contains some missing values)
BMXLEG	Upper leg length in cm	Values between 23.5 to 55 cm
BMX WAIST	Waist circumference in cm	Values between 38.2 to 173.4cm
BMXWT	Weight in kg	Values between 3.1 to 193.3 Kg
BPQ020	Ever been told that you have a high blood pressure	1 = yes 2 = no

		7 = refused 9 = don't know
DMDECUC2	Educational level for adults who are above 20 in age	1 = less than 9 th grade 2 = 9-11 th grade 3 = high school grad / GED 4 = some college or AA degree 5 = college graduate or above 7 = refused 9 = don't know
INDHHINC	Annual household income	Range value between 0\$ 75,000\$
LBXTC	Total cholesterol	Values between 72 – 575
MCQ250A	Any blood relatives who have diabetes	1 = yes 2 = no 7 = refused 9 = don't know
PAQ180	Avg. level of physical activity each day	1 = sits during the day and does not walk much 2 = stands or walks much but does not lift heavy things often 3 = lift(s) heavy loads or climb(s) hills or stairs often 4 = do/does heavy work or carry heavy loads 7 = refused 9 = don't know
RIAGENDER	Gender	1 = Male 2 = Female The dataset has 4883 males and 5082 females
RIDAGYER	Age	0 – 84 = any numeric value to represent the age 85 = represents age equal to or greater than 85

RIDRETH1	Race	1 = Mexican American 2 = other Hispanic 3 = Non-Hispanic white 4 = Non-Hispanic Black 5 = other race - Including Multi-Racial
SMD030	Age when started smoking	Age between 7 – 77 years 0 = never smoked 777 = refused 999 = didn't know
Status	Target variable	0 = Non Diabetic patient 1 = Diabetic patient

3.4.2 Preprocessing of the NHANES dataset

I. Handling the missing values

The NHANES dataset as a whole contained 4,805 missing values. The fig. shows the number of null values in each attribute.

frequency_of_alcohol	909
BMI	122
HEIGHT	93
LEG_LENGTH	170
WAIST	155
WEIGHT	81
BLOOD_PRESSURE	56
EDUCATIONAL_LEVEL	0
ANNUAL_HOUSEHOLD_INCOME	517
CHOLESTEROL	107
DM_IN_FAMILY	0
PHYSICAL_ACTIVITY_IN_A_DAY	0
RIAGENDR	0
RIDAGEYR	0
RACE	0
SMOKING_AGE	2595
STATUS	0

Figure 3.5 Missing Values in the dataset

To solve the problem of missing values, we have used KNN Imputer. Nearest Neighbors imputation methods are efficient methods for missing data values. Each missing value is replaced by related cases

among the whole set of records. This method makes sure that the imputed values are closest to the actual values while keeping the inherent structure of the dataset secure [42]. In our research work we have used:

$$K = 5$$

II. Renaming the Columns

As the features in the dataset are specified by a certain keyword, to make things a little bit easy and friendly for us, we have renamed the columns into some meaningful names.

III. Dimensionality Reduction

we have dropped down the features which have little or no effect on the target variable. We have dropped down these columns:

- Educational Level
- Annual Household Income
- Race

Also, we have dropped the columns Amount of Alcohol consumed annually, since, in our culture, the assumption of Alcohol is very rare and forbidden in Pakistani culture.

3.4.3 Splitting the dataset

I. Training Data

We have used 70% of our data as the training data where X represents the feature set and y represents the target variable.

II. Cross-Validation

We have used k-fold cross-validation with k=10

3.4.4 Algorithms and Hyper Parameters

To find the optimum results, we have applied the following algorithms to FBS the training data:

Table 3.2 Hyper-parameters for NHANES Dataset

Algorithm applied	Hyper parameters
Logistic Regression (LR)	C = 0.01 solver = libliner max_iter = 100
Support vector machine (SVM)	C = 10 Kernel = rbf
Decision Tree (DT)	Criterion = entropy Max_depth = 4
Adaboost Classifier	Max_ddepth = 2 n_estimators = 100 Learning_rate = 0.5
Gradient boosting	n_estimators = 100
Random Forest (RF)	n_estimators = 300 max_depth = 3 criterion = entropy n_jobs = -1 random state = 50 max_features = auto FBS min_sample_leaf = 50
Bagging Classifier	min_samples_split = 10 max_depth = 3 max_Samples = 0.5 max_features = 1.0 n_estimators = 10
K Nearest Neighbors (Knn)	N_neighbors = 4
Gaussian Naïve Bayes	Default parameters

3.5 Detection of Cardiovascular Disease

3.5.1 Dataset

The dataset that we have used in his research is collected from UCI Irvine Machine Learning Repository and **Cleveland dataset** [28]. This dataset comprises 303 records and 73 attributes, although the researchers, both in the field of Healthcare and Machine Learning agree that only 14 of these attributes are required for the detection of disease. Table II shows a detailed description of these attributes.

- a) The target variable is “Num” which is a multivariate column. It has values ranging from 0-4. The value 0 indicates the absence of the disease and values 1-4 represent the presence of the disease.
- b) For the application of deep learning neural networks, we have synthetically generated 20,000 rows based upon the parent data.

Table 3.3 Dataset Description for Cleveland Data

Attribute	Attribute Description	Values
Age	This represents age in years	Max age Min age
Sex	This represents gender.	0 = Female 1 = Male
Cp	This represents chest pain.	1 = typical angina 2 = atypical angina 3 = non anginal pain 3 = asympatotic
Trestbps	This shows the resting blood pressure at the time of hospital admission.	Values in mm/Hg
Chol	This shows the cholesterol levels	Value in mg/dl
Fbs	This denotes the fasting blood sugar.	Value of fbs > 120 mg/dl, fbs = 1 else fbs = 0
Restecg	This shows electrocardiographic results in 3 values.	0 = normal 1 = ST-T wave abnormality 2 = probable or definite left ventricular hypertrophy by Estes' criteria.
Thalach	This shows the maximum heart rate achieved	Numerical value indicating the heart rate
Exang	This shows whether the patient has exercise-induced asthma or not.	1 = the patient has exercise-induced asthma 0 = the patient doesn't have asthma
Oldpeak	This shows the ST depression induced by exercise relative to rest	
Slope	This shows the slope of the peak exercise relative to rest.	1 = up-sloping 2 = flat 3 = down sloping
Ca	Number of major vessels (0-3) colored by fluoroscopy	
Thal	This shows the thallium heart scan.	Value 3 = normal Value 6 = fixed defect Value 7 = reversible defect
Target	This tells whether the patient has the disease or not.	0 = No disease 1 = Disease is present

3.5.2 Splitting the dataset

I. Training Data

We have used 70% of our data as the training data where X represents the feature set and y represents the target variable.

II. Cross-Validation

We have used k-fold cross-validation with k=10

3.5.3 Algorithms and Hyper Parameters

To find the optimum results, we have applied the following algorithms to the training data:

Table 3.4 Hyper-parameter tuning for Cleveland dataset

Algorithm applied	Hyper parameters
Logistic Regression (LR)	C = 0.01 solver = libliner max_iter = 100
Support vector machine (SVM)	C = 10 Kernel = rbf Probability = true
Decision Tree (DT)	Criterion = entropy Max_depth = 4
Adaboost Classifier	Max_ddepth = 2 n_estimators = 100 Learning_rate = 0.5
Gradient boosting	n_estimators = 100
Random Forest (RF)	n_estimators = 300 max_depth = 3 criterion = entropy n_jobs = -1 random state = 50 max_features = auto min_sample_leaf = 50

Bagging Classifier	min_samples_split = 10 max_depth = 3 max_Samples = 0.5 max_features = 1.0 n_estimators = 10
K Nearest Neighbors (Knn)	N_neighbors = 4
Guassian Naïve Bayes	Default parameters

3.5.3.1 Deep learning model and hyper-parameters

We have also implemented a deep learning model using Keras for the dataset. Since the dataset size was a limitation, we have synthetically increased the size of the dataset to 20,000 rows. We have used 60% of our data as training data and 40% as testing data. We have implemented a sequential model which uses ReLU and Softmax as the activation function. For K-fold cross-validation, we have used k=20.

```

NB_EPOCH = 90
BATCH_SIZE = 1024
VERBOSE = 0
NB_CLASSES = 2 # Heart Disease diagnosis yes = 1, no = 0
OPTIMIZER = Adam() # optimizer
N_HIDDEN = 128
TRAINING_SPLIT = 0.6 # how much from all of the data is split for training
VALIDATION_SPLIT=0.4 # how much in TRAIN is reserved for VALIDATION
DROPOUT = 0.5

```

Figure 3.6 Hyper-parameters for the deep learning model

3.6 Detection of Metabolic Syndrome (Obesity)

3.6.1 Dataset

The dataset that we have used is available on the UCI repository by the name of “**Estimation of obesity levels based on eating habits and physical condition Dataset**” [43]. It contains data on the populations from Peru, Columbia, and Mexico. The reason for choosing this dataset is that it contains attributes that are non-invasive and covers the daily lifestyle of individuals. It has a total of 17 attributes and 2,111 instances.

- a) The target variable is a multivariate column and based on his/her data, sorts a person into one of the seven categories namely: Underweight, Normal, Overweight level 1, Overweight level 2, Obesity type 1, Obesity type 2, and Obesity type 3. Table I shows the rest of the data attributes.

Table 3.5 Table V Dataset Description for Obesity data

Feature	Description	Data type
FAVC	consumption of high caloric food	1 = yes (frequent) 0 = no
FCVC	Frequency of consumption of fruits and vegetables	A numerical value between 1-3 depending upon the frequency of consumption
NCP	Number of main meals	1 = between 1 and 2 2 = three 3 = more than 3 4
CAEC	Consumption of food between meals	0 = Always 1 = Frequent 2 = Sometimes 3 = No
CH20	Consumption of daily amount of water	A numerical value between 1 – 3 to represent water intake between 1 to more than 3 liters
CALC	Consumption of daily alcohol	1 = frequent 2 = sometimes 3 = no
SCC	Calorie consumption monitoring	1 = yes

		0 = no
FAF	Physical activity frequency	A numerical value between 0 – 3 to represent the level of physical activity
Family history with overweight	Is there any history of obesity in an individual's family?	1 = yes 0 = no
Smoking status	Does the individual smoke or not	1 = yes 0 = no
TUE	Time consumed in using technology devices	A numerical value between 0-2 to represent the values ranged between 0-2, 3-5 and more than 5 hours.
MTRANS	Transportation used	0 = Automobile 1 = bike 2 = motorbike 3 = public transport 3 = walking
Gender	This represents the gender of the individual	1 = Male 0 = Female
Weight	This represents the weight in Kgs	A numerical value
Height	This shows the height of the individual in meters	A numerical value
Age	Age in years	Numerical value
NObesity	This is the target variable.	1 = Obese 0 = Not obese

3.6.2 Preprocessing the dataset

- I. Initially, the dataset contained a combination of Integers, floating-point numbers, strings, and characters. We have used a Label encoder to set all the values to integers for effective data modeling.

- II. The target variable was divided into six different stages of obesity. For the sake of this research work, we have combined the values 2 – 5 as Obese (Labelled as 1) and value 0 - 1 as Not Obese (Labelled as 0).

3.6.3 Splitting the dataset

I. Training Data

We have used 70% of our data as the training data where X represents the feature set and y represents the target variable.

II. Cross-Validation

We have used k-fold cross validation k=10

3.6.4 Machine Learning Algorithms and Hyper Parameters

To find the optimum results, we have applied the following algorithms to the training data:

Table 3.6 Hyper parameter tuning for obesity Dataset

Algorithm applied	Hyper parameters
Logistic Regression (LR)	C = 0.01 solver = libliner max_iter = 100
Support vector machine (SVM)	C = 10 Kernel = rbf Probability = true
Decision Tree (DT)	Criterion = entropy Max_depth = 4
Adaboost Classifier	Max_ddepth = 2 n_estimators = 100 Learning_rate = 0.5
Gradient boosting	n_estimators = 100
Random Forest (RF)	n_estimators = 300 max_depth = 3

	criterion = entropy n_jobs = -1 s random state = 50 max_features = auto min_sample_leaf = 50
Bagging Classifier	min_samples_split = 10 max_depth = 3 max_Samples = 0.5 max_features = 1.0 n_estimators = 10
K Nearest Neighbors (Knn)	N_neighbors = 4
Gaussian Naïve Bayes	Default parameters

3.6.5 Application of Feed forward Neural Network

We have also applied a multilayer perceptron Neural Network on the obesity dataset using AWS Sage maker studio. The list of hyper-parameters is defined as below:

```

"_tuning_objective_metric": "validation:accuracy",
  "activation": "relu",
  "dropout_prob": 0.33088209744918484,
  "eval_metric": "accuracy",
  "layers": [256],
  "learning_rate": 1.7280370809758323e-05,
  "log_interval": 1,
  "max_layer_width": 2056,
  "min_epochs": 1,
  "mini_batch_size": 318,
  "ml_application": "mlp",
  "momentum": 0.9,
  "network_type": "feedforward",
  "num_categorical_features": 8,
  "num_classes": 7,
  "num_epochs": 100,
  #"numeric_embedding_dim": null,
  "optimizer": "adam",
  "patience": 10,
  "positive_example_weight_mult": 1.0,
  "problem_type": "multiclass_classification",|

```

Figure 3.7 List of hyper parameters for neural network

3.7 Evaluation metrics

Model validation: validating the model is one of the most important steps in machine learning. In our study, we have used the k-fold cross-validation method. We have applied the k-fold method to validate the results of our different machine learning models.

K-fold cross-validation is an iterative technique in which a data set is split into k equal parts. The k-1 part is used for training while all the other parts are used for testing. This goes on till all k iterations are completed. k can have any value, in our study we have used k=10. Here, 90% of the data is used for model training while 10% data is used for model testing at each level of iteration. After the completion of all the iterations, a mean value of all the results at each level is taken as the final answer.

Performance Evaluation: For performance evaluation, we have used various evaluation metrics including Recall, Precision, and Accuracy, and ROC curve. A confusion matrix containing true positive, true negative, false positive, and false negative values are used to calculate these measures. Table IV shows the said confusion matrix.

Table 3.7 Confusion Matrix

	Predicted value (No disease)	Predicted value (have a disease)
Actual (no disease)	True Negative (TN)	False Positive (FP)
Actual (have a disease)	False Negative (FN)	True Positive (TP)

A **true negative (TN)** value shows that a patient does not have any disease and the model has also predicted the same. In this case, the classifier works correctly.

True positive (TP) value represents that patient has the disease and the model also predicts the same. In this case, the classifier works correctly.

A **false positive (FP)** value shows that the patient does not have any disease but the model predicts that the patient has the disease. In this case, the classifier incorrectly classifies a person. This is also called a type I error.

False Negative (FN) value represents that the patient has the disease but the model predicts that the patient does not have any disease. Here, the classifier also incorrectly classifies the patient. This is called a, benchmarked, type II error.

1. Precision

The precision of a system tells how accurate a model is. It is a measure that indicates how accurately the system predicts the true values.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (\text{i})$$

2. Recall

Recall tells how many actual positive values our model identified by labeling them as positive (true positive) It can be calculated using the following formula:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (\text{ii})$$

3. F1- Score

This measure is the harmonic mean of precision and recall and tells the overall performance of the model.

$$\text{F-1 Measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (\text{iii})$$

4. Accuracy

Accuracy is a ratio between total correctly classified objects to the total number of predictions made.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (\text{iv})$$

Finally, we have verified the predictability of our prediction models using **Receiver Optimistic Curve (ROC)** values. ROC is a graphical representation of the performance of machine learning classifiers. The

area under the curve (AUC) gives the value of ROC based on the classifier's performance. The greater the value of AUC, the greater is the performance of the model.

CHAPTER 4

Experiments and Results

In this chapter, I have briefly explained the different experiments I have carried out onto the datasets for the early identification of lifestyle diseases. I have also discussed the results of these experiments and measured the performance of the models using performance evaluation metrics.

Experimental Results and Analysis

The efficiency of a model is solely based on the quality of the dataset and how it is trained. In this research, we aim to identify the three most common lifestyle diseases namely diabetes mellitus type -2, cardiovascular diseases, and metabolic – obesity. For the model training purpose, we have selected the benchmarked data for these three diseases. As the quality of the model is directly linked to the quality of the dataset, we have applied multiple pre-processing techniques on the dataset and then finally trained the model. This chapter explains all the results of different experiments that we have done on our dataset to achieve the maximum accuracy and the performance evaluation metrics for these models. 4.1 contains the experimental results for D2TM, 4.3 tells the experimental results for the CVD, and 4.4 tells the experimental results for obesity.

4.1 Experiment 1 – detection of Diabetes Mellitus – type 2

For the detection of diabetes mellitus – type 2 (T2DM), we have chosen the NHANES diabetes (2000 – 2004) dataset. It is considered as the benchmark data for the detection of diabetes based on lifestyle factors. It is based on survey data of 5,514 individuals and covers almost all the required lifestyle attributes.

In this experiment, after pre-processing the dataset, we have used the heat map to figure out the correlations between different data attributes. We have applied LR, DT, KNN, SVM, Ensemble methods (RF, Bagging, and Boosting) on this dataset. For the performance evaluation we have used Precision, Recall, Accuracy, F-1 Measure, and AUC-ROC curve values. Fig. 13 shows the heat map plotted against the NHANES

dataset. The attributes Waist, BMI, Weight, Height, and Age are most related to the target variable. The attributes cholesterol, frequency of alcohol, above-mentioned, cross-validation, and smoking age are also related to the target variable.

4.1.1 Heat map for NHANES diabetes dataset

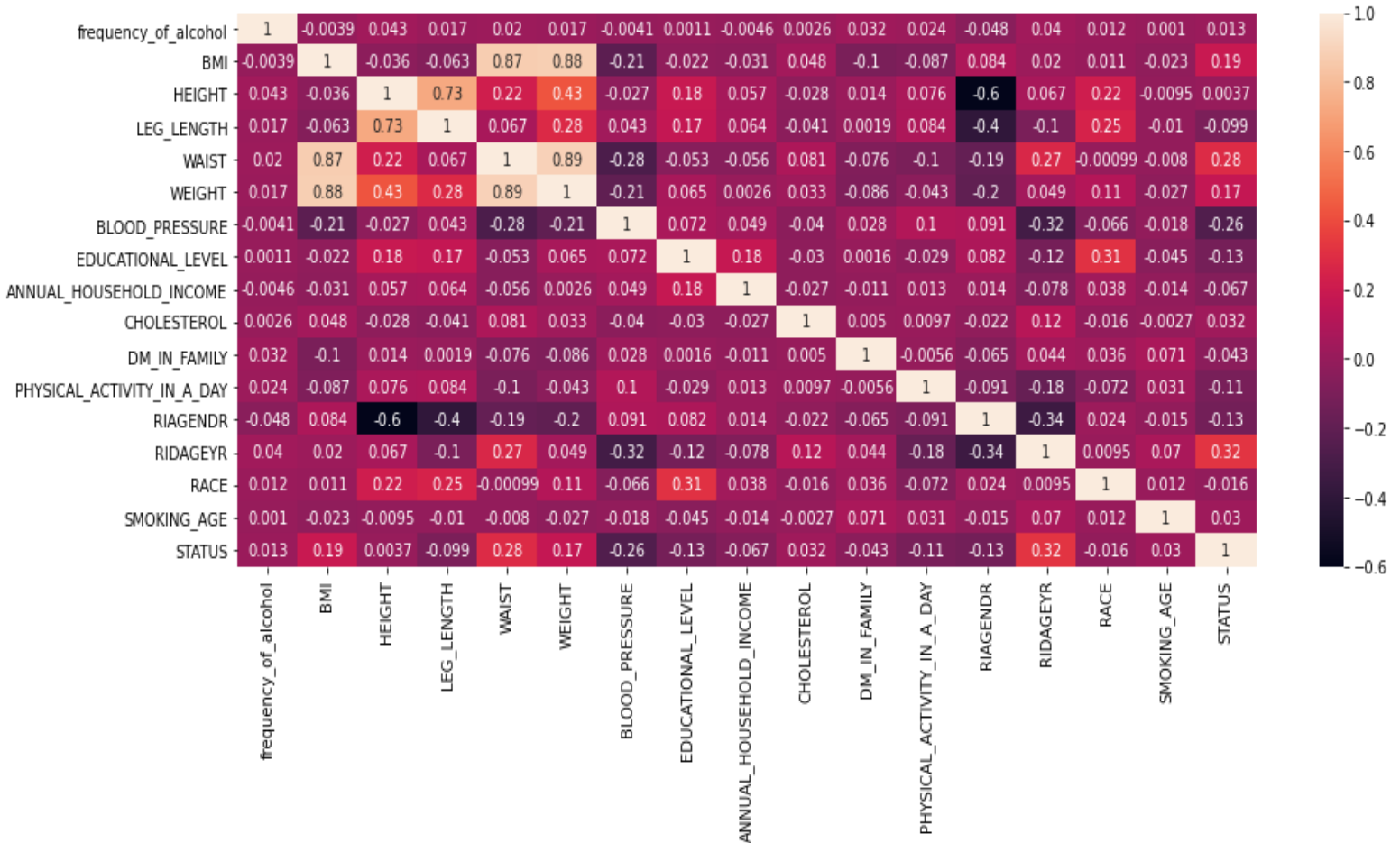


Figure 4.1 Heat Map for NHANES dataset

4.1.2 Results

Table VI shows the results of these above-mentioned supervised learning algorithms (DT, KNN, NB, LR, SVM, Ensemble, and Random Forest) using the 10 fold cross-validation technique. We have measured the performance of the models using Accuracy, precision, recall, F-1 measure, and ROC metrics.

For experimentation purposes, we have tried using KNN with multiple different values for the nearest neighbor. In our results we have mentioned just two values for KNN; where $K = 7$ and $K = 4$.

Table 4.1 Results for Diabetes identification models

Classifier	Accuracy	Precision	Recall	F-1 Measure	ROC
DT	0.82	0.78	0.82	0.78	0.796
KNN where N= 4	0.752	0.75	0.76	0.76	0.679
LR with solver =libliner	0.81	0.78	0.82	0.78	0.801
SVM	0.80	0.66	0.81	0.73	0.820
Bagging	0.82	0.66	0.81	0.73	0.820
Adaboost	0.80	0.66	0.81	0.73	0.789
Gradient Boosting	0.825	0.78	0.82	0.78	0.834
Random Forest	0.81	0.78	0.82	0.78	0.828
Gaussian Naïve Bayes	0.784	0.75	0.75	0.76	0.781

Figure 14 compares the AUC-ROC curve values for all the implemented models using the 10 fold cross-validation technique.

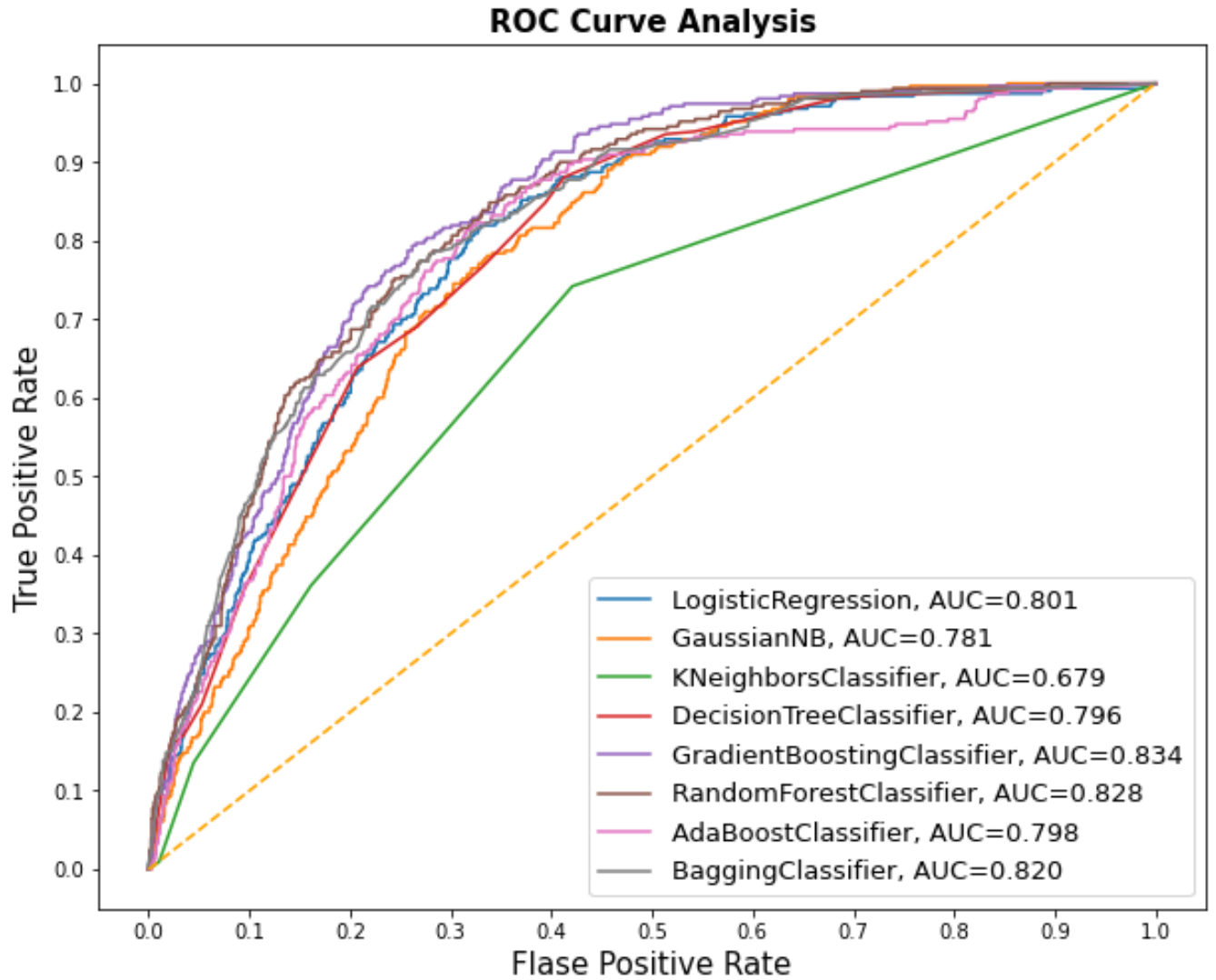


Figure 4.2 AUC-ROC analysis for different models for the identification of diabetes

4.2 Experiment 2 - Cardiovascular diseases – CVD

For the detection of cardiovascular diseases CVD, we have chosen the Cleveland dataset. The data attributes for this dataset are a combination of lifestyle factors as well as clinical factors which this dataset an ideal dataset for the detection of CVDs.

In this experiment, after pre-processing the dataset, we have used the heat map to figure out the correlations between different data attributes. We have applied LR, DT, KNN, SVM, Ensemble methods (RF, Bagging, and Boosting) on this dataset. For the performance evaluation we have used Precision, Recall, Accuracy, F-1 Measure, and AUC-ROC curve values. Fig. 15 shows the heat map plotted against the NHANES dataset. The attributes cp (chest pain), thalach (maximum heart rate), and slope (the slope of peak exercise relative to rest) are most related to the target variable. The attributes cholesterol, frequency of alcohol, and smoking age are also related to the target variable. The attribute rest ECG is also correlated to the target variable.

4.2.1 Heat Map for the Cleveland dataset

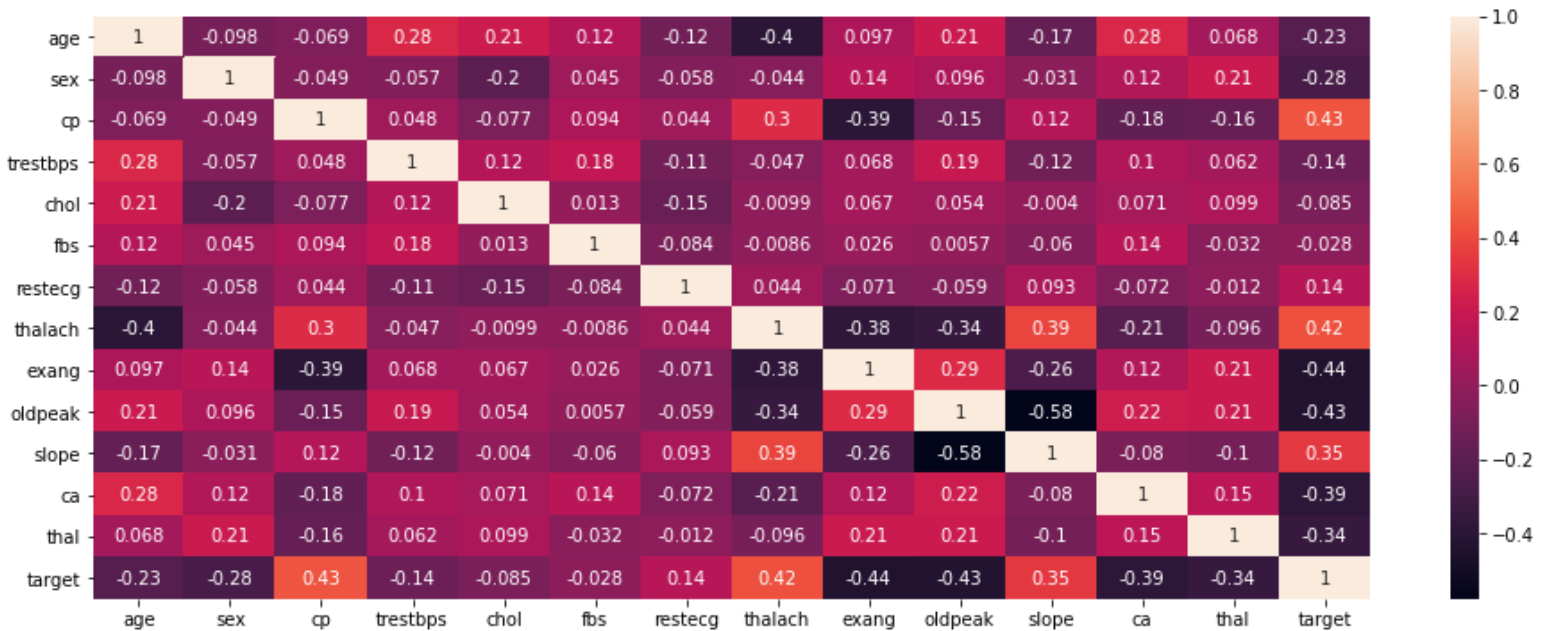


Figure 4.3 Heat map for the Cleveland dataset

4.2.2 Results

The table below shows the performance evaluation for the implementation of different algorithms on the training data.

Table 4.2 Results of different models on the Cleveland dataset for the identification of cardiovascular diseases

Classifier	Accuracy	Precision	Recall	F-1 Measure	ROC
DT	0.82	0.71	0.71	0.71	0.73
KNN where N= 7	0.850	0.81	0.81	0.81	0.85
KNN where N= 4	0.78	0.79	0.79	0.78	0.84
LR with solver	0.827	0.80	0.80	0.80	0.88
SVM	0.850	0.80	0.80	0.80	0.86
Bagging	0.845	0.79	0.79	0.79	0.82
Adaboost	0.816	0.79	0.79	0.79	0.76
Gradient Boosting	0.855	0.79	0.79	0.79	0.80
Random Forest	0.878	0.79	0.79	0.79	0.88
Gaussian Naïve Bayes	0.832	0.81	0.81	0.81	0.88

Fig. 16 shows the ROC Curve analysis graph for the different models trained on the said algorithms.

SVM has the most optimal results.

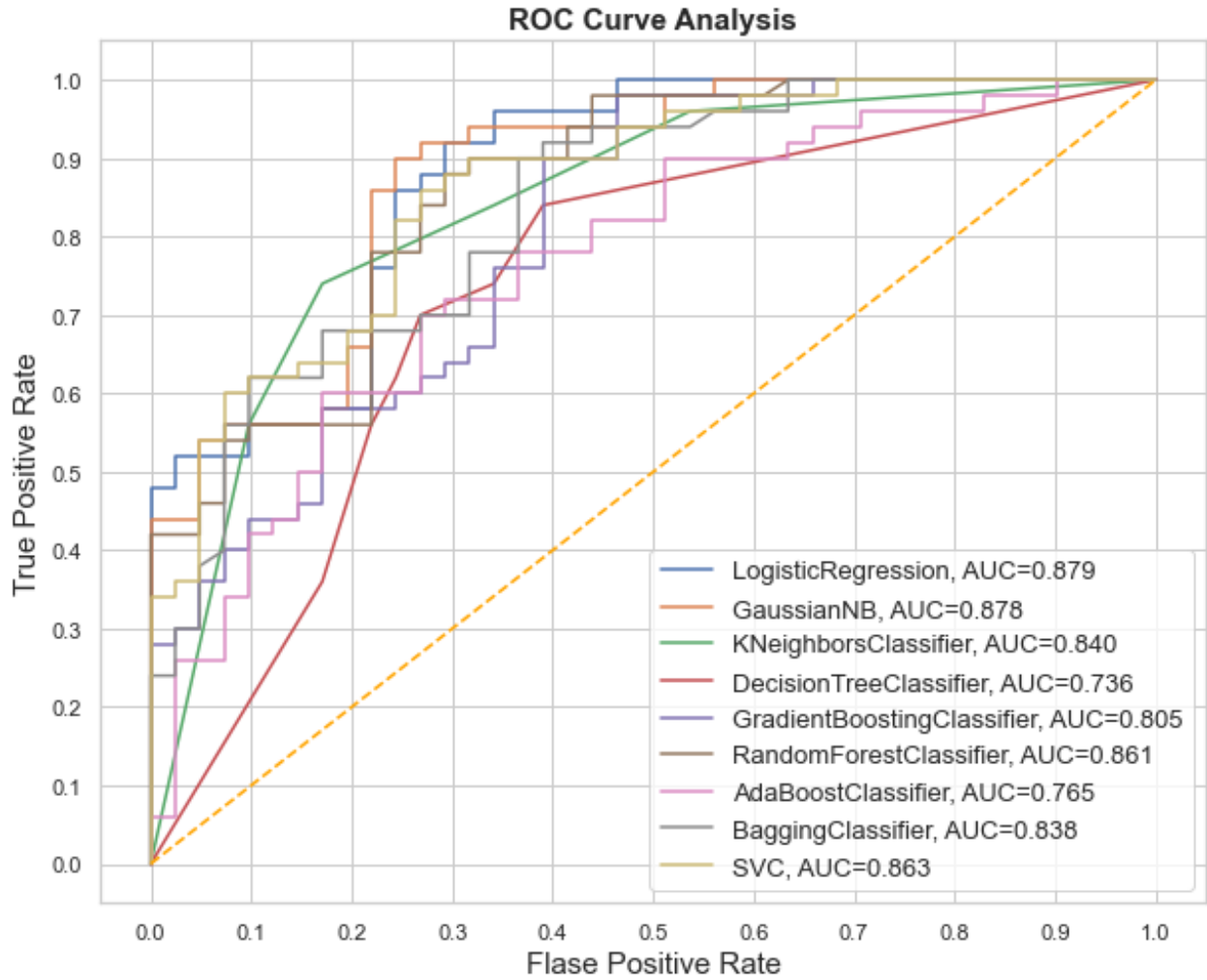


Figure4.4 AUC-ROC analyses for different models on Cleveland dataset.

4.2.3 Application of Neural Network

In this experiment, we have applied Feed Forward Neural Network on the dataset after synthetically generating 20,000 data rows based on the parent dataset. Figures 17 and 18 show the model accuracy and model precision. Finally, in Figure 19, we have used box plots to visualize the performance metrics of the Keras Deep Learning Model. After tuning the model, we get an accuracy of 83%, a precision of 0.812, and a recall of 0.852.

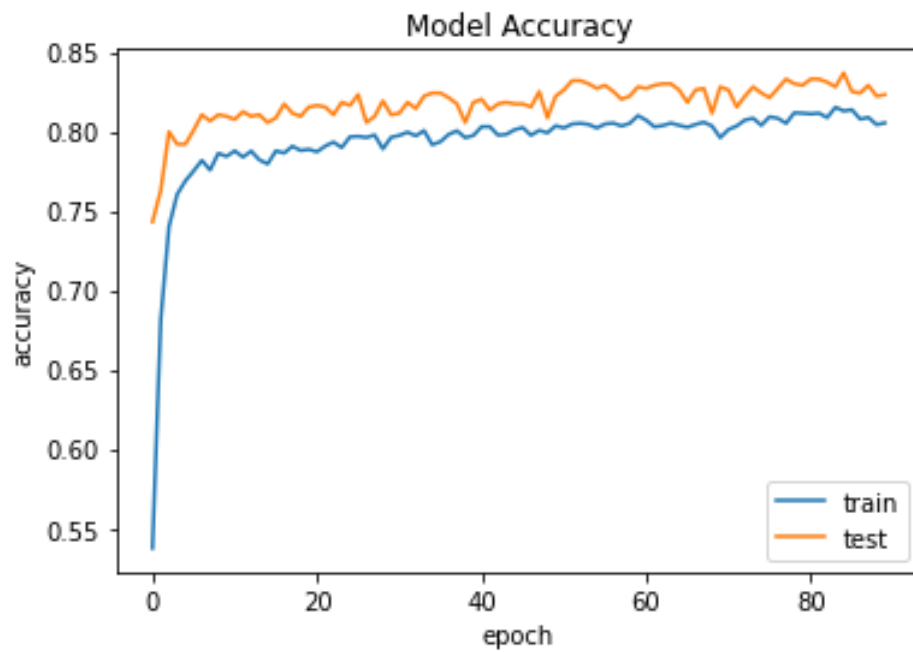


Figure 4.5 Model Accuracy for Feed Forward Neural Network

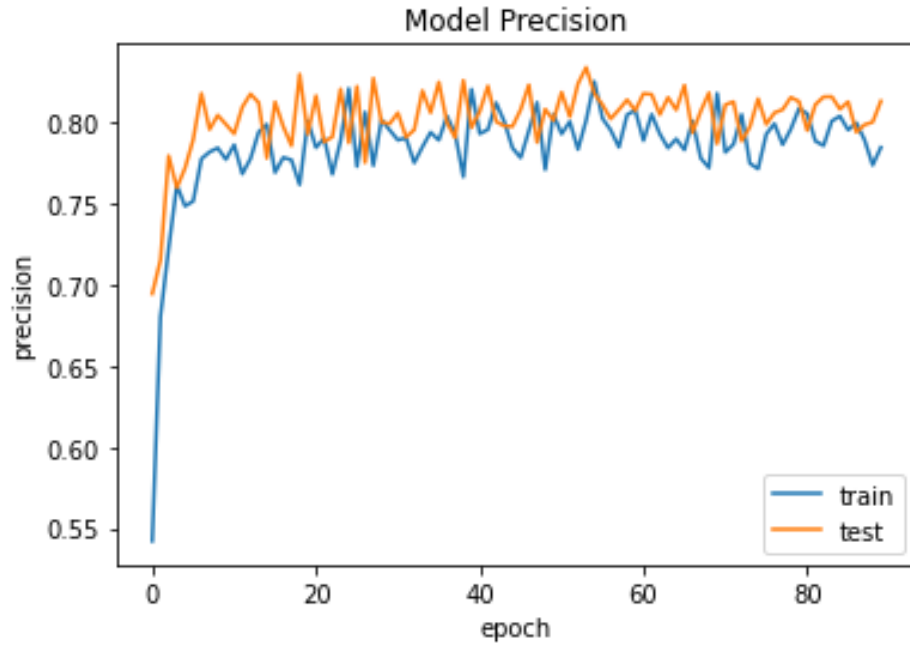


Figure 4.6 Model precision for feed forward neural network

Keras Deep Learning Model Performance

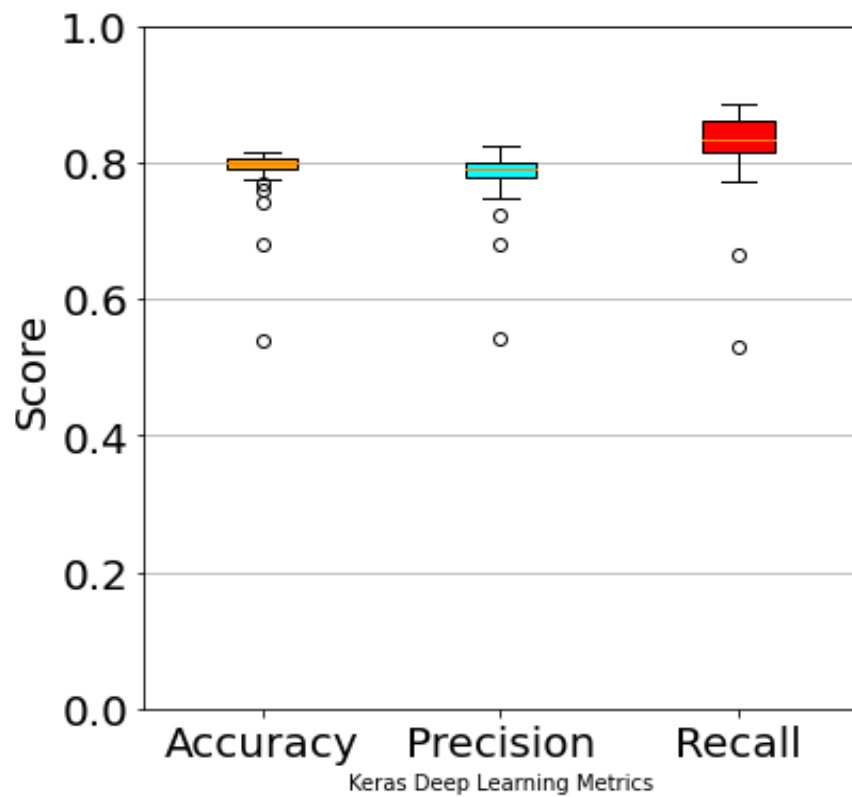


Figure 4.7 Boxplot for neural network on obesity dataset

4.3 Experiment 3 – Metabolic Syndrome – Obesity

For the detection of metabolic syndrome - obesity, I have used the Obesity Dataset gathered from the population of Peru, Mexico, and Brazil. This dataset is publicly available at UCI Machine Learning Repository. The data attributes for this dataset are a combination of lifestyle factors as well as clinical factors which make this dataset an ideal dataset for the detection of obesity.

In this experiment, after pre-processing the dataset, we have used the heat map to figure out the correlations between different data attributes. We have applied LR, DT, KNN, SVM, Ensemble methods (RF, Bagging, and Boosting) on this dataset. For the performance evaluation we have used Precision, Recall, Accuracy, F-1 Measure, and AUC-ROC curve values. Fig. 20 shows the heat map plotted against the NHANES dataset. The attributes age, weight, height, CAEC () and family_history_with_overweight are most related to the target variable. The attributes CH20, SCC are also correlated to the target variable.

4.3.1 Heat Map for the obesity dataset

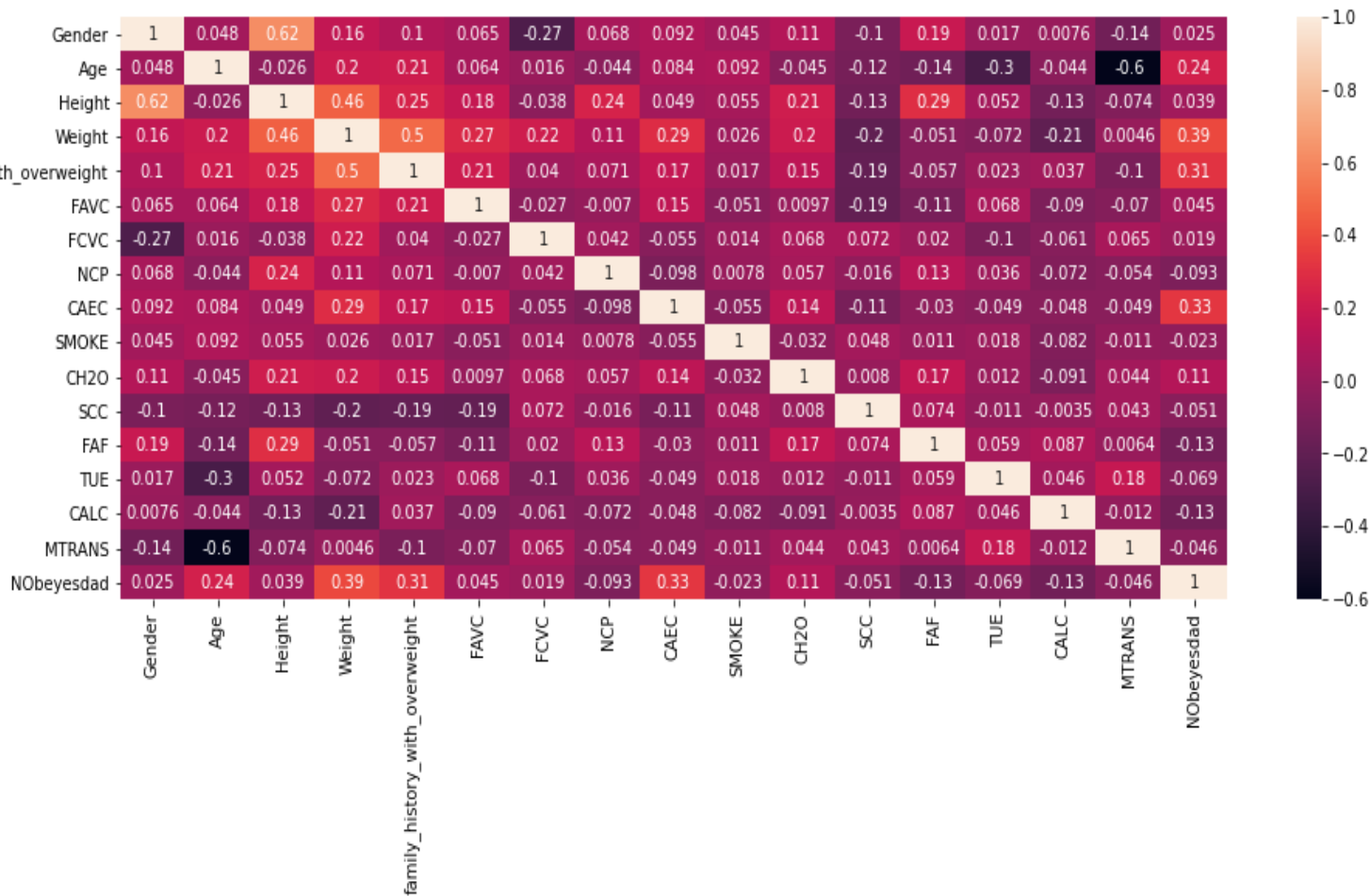


Figure 4.8 Heat map for the obesity dataset

The table below shows the performance evaluation for the implementation of different algorithms on the training data. For k nearest neighbor we have selected the value of K as K = 4 and K = 7. This is derived by the following graph where the value closest to 4 indicates the highest accuracy.

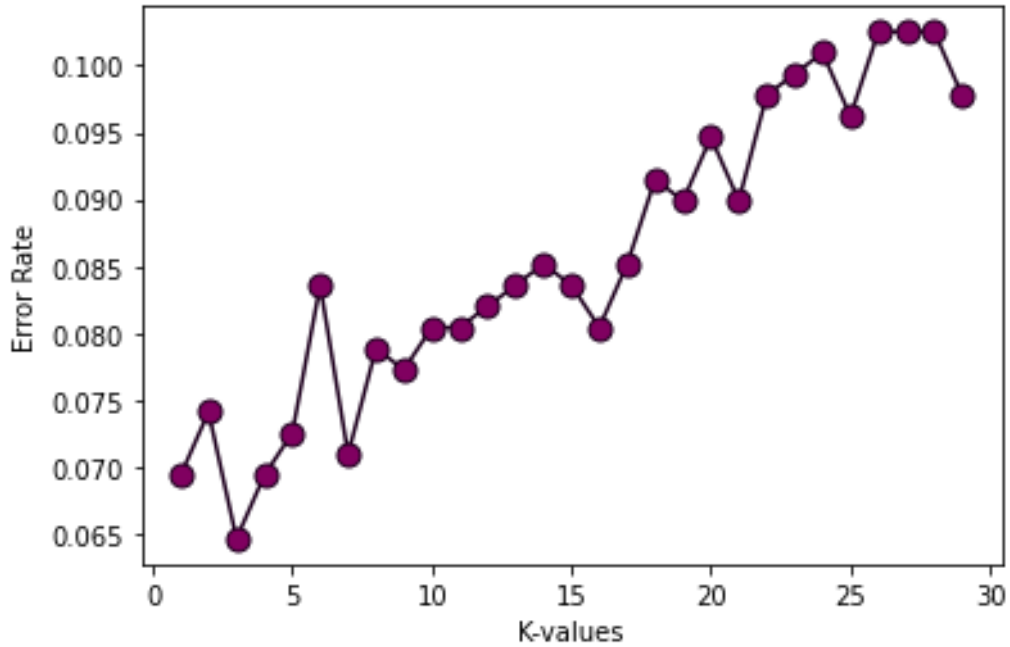


Figure 4.9 Error rate for different values of K

4.3.2 Results

The table below shows the evaluation metrics applied n different machine learning algorithms.

Table 4.3 Results of different models for the identification of Obesity

Classifier	Accuracy	Precision	Recall	F-1 Measure	ROC
DT	0.97	0.97	0.97	0.97	0.96
KNN where N= 4	0.97	0.93	0.93	0.93	0.97
LR with Solver	0.92	0.93	0.93	0.93	0.97
SVM	0.96	0.96	0.96	0.96	0.99
Bagging	0.96	0.96	0.96	0.96	0.99
Adaboost	0.99	0.96	0.96	0.96	1.00
Gradient Boosting	0.98	0.96	0.96	0.96	1.00
Random Forest	0.945	0.96	0.96	0.96	0.99

Gaussian Naïve Bayes	0.945	0.93	0.93	0.93	0.99
MLP	0.98	0.96	0.96	0.96	1.00

4.3.3 Endpoint generation

Given the inputs, the model predicts that the patient has obesity.

```
l="Male, 26, 1.85, 105, yes, yes, 3, 3, Frequent, no, 3, no, 2, 2, Sometimes, Public_Transportation"
ep_name = "finalobesitypredcition11"
response = sm_rt.invoke_endpoint(EndpointName=ep_name, ContentType='text/csv', Accept='text/csv', Body=l)
response = response['Body'].read().decode("utf-8")
print (response)
```

Obesity_Type_I

Figure 4.10 Prediction of Obesity in a patient

CHAPTER 5

Conclusion

This research work provides a novel approach for the early detection of lifestyle diseases. The novelty lies in the fact that we have centered our model on the Pakistani demographics and culture. We have taken the 3 most common lifestyle diseases, namely: Diabetes Mellitus Type 2, cardiovascular diseases, and obesity. We have applied a number of machine learning and deep learning algorithms on the standard datasets for these diseases. Previously, no work has been done for the Pakistani population in which a diagnostic model is proposed for the detection of these three diseases. Our models suggest that for the **detection of diabetes**, the **Gradient Boosting classifier** works best. For **the detection of CVD**, the **Random Forest classifier** gives the most optimum results. And for the classification of **Obesity**, the classifiers **Adaboost and Gradient Boosting** work best. We have thoroughly checked our models and trained them on varying hyper-parameters.

The model is trained and tested by dividing the dataset into 70:30. We have tested our models on different performance evaluation metrics including accuracy, precision, recall, F-1 Measures, and ROC values. K-Fold cross-validation technique is also applied where the value of $K = 10$.

5.1 Future Works

In the future, we propose to create a front end to these models where they can be used as an assistive measure for the timely detection and analysis of lifestyle diseases in a hospital or medical facility. Owing to the unavailability or incompleteness of Pakistani data, we wish to first test our model on the Pakistani dataset which captures the true lifestyle

References:

- [1] D. D. Farhud, "Impact of Lifestyle on Health," vol. 44, no. 11, pp. 1442–1444, 2015.
- [2] Fatma Al-Maskari, "LIFESTYLE DISEASES: An Economic Burden on the Health Services." [Online]. Available: <https://www.un.org/en/chronicle/article/lifestyle-diseases-economic-burden-health-services>.
- [3] J. M. Rippe, "Lifestyle Medicine : The Health Promoting Power of Daily Habits," vol. 12, no. 6, pp. 499–512, 2018.
- [4] "Non-communicable diseases risk factors survey - Pakistan."
- [5] S. Habibullah, J. Ashraf, I. Taseer, R. Javed, and S. Naz, "Prevalence of Shisha Smoking in College, University and Madarsa Students Aged 20-25 Years in Pakistan," no. January 2012.
- [6] P. P. Shinde, "A Review of Machine Learning and Deep Learning Applications," 2018.
- [7] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nat. Med.*, vol. 25, no. January 2019.
- [8] D. J. Hand, "Principles of Data Mining," vol. 30, no. 7, pp. 621–622, 2007.
- [9] N. Jothi, N. Aini, A. Rashid, and W. Husain, "Data Mining in Healthcare – A Review," *Procedia - Procedia Comput. Sci.*, vol. 72, pp. 306–313, 2015.
- [10] J. Y. Lee *et al.*, "Development and Usability of a Life-Logging Behavior Monitoring Application for Obese Patients," pp. 194–202, 2019.
- [11] M. Ferdous, J. Debnath, and N. R. Chakraborty, "Machine Learning Algorithms in Healthcare : A Literature Survey," 2020.
- [12] E. Beede, A. Iurchenko, L. Wilcox, and L. M. Vardoulakis, "A Human-Centered

- Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy,” pp. 1–12, 2020.
- [13] R. Title, “Machine learning prediction of axillary lymph node metastasis in breast cancer: 2D versus 3D radiomic features.”
- [14] S. R. Shahamiri, “Autism AI : a New Autism Screening System Based on Artificial Intelligence,” pp. 766–777, 2020.
- [15] J. F. Dipnall, J. A. Pasco, M. Berk, and L. J. Williams, “Why so GLUMM ? Detecting depression clusters through graphing lifestyle-environs using machine-learning methods (GLUMM),” *Eur. Psychiatry*, vol. 39, pp. 40–50, 2017.
- [16] F. Farzadfar, “Comment Cardiovascular disease risk prediction models : challenges and perspectives,” *Lancet Glob. Heal.*, vol. 7, no. 10, pp. e1288–e1289, 2019.
- [17] M. Data and P. N. Identified, “Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data,” pp. 1–4, 2018.
- [18] M. Fortin, J. Haggerty, J. Almirall, T. Bouhali, M. Sasseville, and M. Lemieux, “Lifestyle factors and multimorbidity : a cross-sectional study,” pp. 1–8, 2014.
- [19] I. Classification, “Standards of Medical Care in Diabetes d 2014,” vol. 37, no. October 2013, pp. 14–80, 2014.
- [20] A. Nanri *et al.*, “Association of weight change in different periods of adulthood with risk of type 2 diabetes in Japanese men and women : the Japan Public Health Center-Based Prospective Study,” 2011.
- [21] E. Ferrannini, “Definition of intervention points in prediabetes,” vol. 8587, no. 13, pp. 1–9, 2014.
- [22] “Cardiovascular diseases (CVDs),” *World Health Organization*, 2021. [Online].

Available: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).

- [23] “Global Obesity Observatory.” [Online]. Available: <https://data.worldobesity.org/country/pakistan-167/>.
- [24] S. Gómez-martínez *et al.*, “Eating Habits and Total and Abdominal Fat in Spanish Adolescents : Influence of Physical Activity . The AVENA Study,” vol. 50, pp. 403–409, 2012.
- [25] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, “Open Access A data-driven approach to predicting diabetes and cardiovascular disease with machine learning,” vol. 5, pp. 1–15, 2019.
- [26] D. Pei, Y. Gong, H. Kang, C. Zhang, and Q. Guo, “Accurate and rapid screening model for potential diabetes mellitus,” vol. 3, pp. 1–8, 2019.
- [27] A. U. Haq *et al.*, “Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data.”
- [28] “Heart disease dataset,” *UCI Machine Learning Repository*. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/heart+disease>.
- [29] M. S. Amin, Y. K. Chiam, and K. D. Varathan, “Abstract Cardiovascular disease is one of the biggest cause for morbidity and mortality among the” *Telemat. Informatics*, 2018.
- [30] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, “HDPM : An Effective Heart Disease Prediction Model for a Clinical Decision Support System,” vol. 8, 2020.
- [31] P. Naheeda, K. Sharifullah, S. S. Ullah, A. M. Azeem, Y. Shahzad, and W. Kinza, “Development of a cost-effective CVD prediction model using lifestyle factors . A cohort study in Pakistan,” vol. 20, no. 2, pp. 849–859, 2020.

- [32] F. Ali *et al.*, “A Smart Healthcare Monitoring System for Heart Disease Prediction Based On Ensemble Deep Learning and Feature Fusion,” *Inf. Fusion*, 2020.
- [33] C. Cheng and H. Chiu, “An Artificial Neural Network Model for the Evaluation of Carotid Artery Stenting Prognosis Using a National-Wide Database,” pp. 2566–2569, 2017.
- [34] Y. Zhang, M. Qiu, S. Member, and C. Tsai, “Health-CPS : Healthcare Cyber-Physical System Assisted by Cloud and Big Data,” vol. 11, no. 1, pp. 88–95, 2017.
- [35] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, “Critical analysis of Big Data challenges and analytical methods,” *J. Bus. Res.*, vol. 70, pp. 263–286, 2017.
- [36] S. D. Desai, S. Giraddi, and P. Narayankar, *Back-Propagation Neural Network Versus Logistic Regression in Heart Disease Classification*. Springer Singapore.
- [37] K. Burse, V. Pratap, S. Kirar, A. Burse, and R. Burse, *Various Preprocessing Methods for Neural Network Based Heart Disease Prediction*. Springer Singapore, 2019.
- [38] Q. Xue, X. Wang, S. Meehan, J. Kuang, J. A. Gao, and M. C. Chuah, “Recurrent Neural Networks based Obesity Status Prediction Using Activity Data,” pp. 865–870, 2018.
- [39] Z. Zheng and K. Ruggiero, “Using Machine Learning to Predict Obesity in High School Students,” pp. 2132–2138, 2017.
- [40] C. Janiesch and K. Heinrich, “Machine learning and deep learning,” 2021.
- [41] “NHANES DIABETES DATASET.”
- [42] L. Beretta and A. Santaniello, “Nearest neighbor imputation algorithms : a critical evaluation,” *BMC Med. Inform. Decis. Mak.*, vol. 16, no. Suppl 3, 2016.
- [43] “UCI Machine Learning Repository for Obesity.” [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+ha>

bits+and+physical+condition+.