IMPROVE DIABETES DIAGNOSIS BY INTEGRATING MULTIPLE MACHINE LEARNING ALGORITHMS

2023
MASTER THESIS
COMPUTER ENGINEERING

Feras Muhammed KHALEL

Thesis Advisor
Assist. Prof. Dr. Nehad T.A RAMAHA

IMPROVE DIABETES DIAGNOSIS BY INTEGRATING MULTIPLE MACHINE LEARNING ALGORITHMS

Feras Muhammed KHALEL

Thesis Advisor
Assist. Prof. Dr. Nehad T.A RAMAHA

T.C.
Karabuk University
Institute of Graduate Programs
Department of Computer Engineering
Prepared as
Master Thesis

KARABUK
July 2023

I certify that in my opinion the thesis submitted by Feras Muhammed KHALEL titled "IMPROVE DIABETES DIAGNOSIS BY INTEGRATING MULTIPLE MACHINE LEARNING ALGORITHMS" is fully adequate in scope and in quality as a thesis for the degree of Master of Science.

Assist.Prof.Dr. Nehad T.A. Ramaha .....
Thesis Advisor, Department of Computer Engineering

This thesis is accepted by the examining committee with a unanimous vote in the Department of Computer Engineering as a Master of Science thesis. July, 2023

Examining Committee Members (Institutions) Signature

Chairman : Assist. Prof. Dr. İsa AVCI (KBU) .....

Member :Assist.Prof.Dr. Muhammet ÇAKMAK (Sinop University)……...

Member : Assist.Prof.Dr. Nehad T.A. RAMAHA (KBU) .....

The degree of Master of Science by the thesis submitted is approved by the Administrative Board of the Institute of Graduate Programs, Karabuk University.

Prof.Dr. Müslüm KUZU .....
Director of the Institute of Graduate Programs

"I declare that all the information within this thesis has been gathered and presented in accordance with academic regulations and ethical principles and I have according to the requirements of these regulations and principles cited all those which do not originate in this work as well."

Feras Muhammed Khalel

ABSTRACT

M. Sc. Thesis

IMPROVE DIABETES DIAGNOSIS BY INTEGRATING MULTIPLE MACHINE LEARNING ALGORITHMS

Feras Muhammed Khalel

Karabük University
Institute of Graduate Programs
The Department of Computer Engineering

Thesis Advisor:
Assist.Prof. Dr. Nehad T.A RAMAHA
July 2023, 65 pages

Diabetes is a chronic disease with many complications that follow the disease and is one of the leading causes of death worldwide. The number of people infected with this disease is increasing every day. Therefore, predicting this disease at an early stage helps to avoid many complications that follow the disease. Nowadays, many medical sectors have begun to take an interest in using artificial intelligence technologies and benefiting from their services. Data mining and machine learning techniques are used to predict the patient's condition at an early stage. Many studies have worked on this topic. However, in most previous studies, the recall measure for affected patients did not reach an acceptable accuracy. Therefore, in this study, we worked to address this deficiency and present a new model for predicting diabetes. During the study, the data set was processed. Six machine learning algorithms were then used to build the models and individually predict the patient's condition. Then the three best algorithms were selected. Finally, these algorithms were combined to

create a hybrid model that gives a safe result to predict the patient's condition. The proposed model gave an accuracy of 96.55%. The accuracy of the recall scale was 97.73%. The accuracy obtained from this study is better than the accuracy of previous studies and more reliable because, in this study, the data set was processed and made balanced.

Key Words **:** Diabetes, Pima dataset, Random Forest, Gradient Boosted, KNN, SVM.

Science Code **:** 9243

ÖZET

Yüksek Lisans Tezi

ÇOKLU MAKİNE ÖĞRENME ALGORİTMALARINI ENTEGRE EDEREK DİYABET TEŞHİSİNİ GELIŞTIRIN

Feras Muhammed Khalel

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Dr. Öğr. Üyesi. Nehad T.A Ramaha
Temmuz 2023, 65 sayfa

Diyabet, hastalığı takip eden birçok komplikasyonu olan ve dünya çapında önde gelen ölüm nedenlerinden biri olan kronik bir hastalıktır. Bu hastalığa yakalananların sayısı her geçen gün artıyor. Bu nedenle, bu hastalığı erken bir aşamada tahmin etmek, hastalığı takip eden birçok komplikasyonun önlenmesine yardımcı olur. Günümüzde birçok medikal sektör yapay zeka teknolojilerini kullanmaya ve hizmetlerinden yararlanmaya ilgi duymaya başlamıştır. Hastanın durumunu erken bir aşamada tahmin etmek için veri madenciliği ve makine öğrenimi teknikleri kullanılır. Bu konuda birçok çalışma yapılmıştır. Bununla birlikte, önceki çalışmaların çoğunda, etkilenen hastalar için hatırlama ölçüsü kabul edilebilir bir doğruluğa ulaşmadı. Bu nedenle, bu çalışmada bu eksikliği gidermek ve diyabeti tahmin etmek için yeni bir model sunmak için çalıştık. Çalışma sırasında veri seti işlenmiştir. Daha sonra modelleri oluşturmak ve hastanın durumunu bireysel olarak tahmin etmek için altı makine öğrenimi algoritması kullanıldı. Ardından en iyi üç algoritma seçildi. Son

olarak, bu algoritmalar, hastanın durumunu tahmin etmek için güvenli bir sonuç veren hibrit bir model oluşturmak için birleştirildi. Önerilen model %96.55 doğruluk vermiştir. Hatırlama ölçeğinin doğruluğu %97.73 idi. Bu çalışmadan elde edilen doğruluk önceki çalışmalara göre daha iyi ve daha güvenilirdir çünkü bu çalışmada veri seti işlenip dengelenmiştir.

Anahtar Kelimeler : Diyabet, Pima veri seti, Random Forest, Gradient Boosted, KNN, SVM.

Bilim Kodu : 92432

ACKNOWLEDGMENT

First of all, I thank God Almighty for making it easy for me to study at Karabuk University. Then to my father and mother who supported me during my studies. I also thank Dr. Öğr.Üyesi. Nehad T.A Ramaha, who supervised my master's thesis and helped me in this work. Finally, I would like to thank my friends at home and the university.

CONTENTS

Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

Page

SYMBOLS AND ABBREVITIONS INDEX

ABBREVITIONS

SVM : Support Vector Machine.

KNN : K-Nearest Neighbor.

DL : Deep Learning.

ML : Machine Learning

AI : Artificial Intelligence

PART 1

RESEARCH OVERVIEW

1.1. INTRODUCTION

The ease of use of software applications is one of the most important reasons for their spread in the modern era. Most institutions and sectors rushed to replace the previous system based on recording data on paper to use applications and build websites that facilitate the management of these institutions. This move also led to data organization as processing and conducting studies became more accessible. The medical sectors are among the most that have tried to benefit from this. They have organized patient and reviewer data in organized records. That has helped researchers a lot in easy access to those data, conducting studies on them, and discovering new patterns and knowledge.

Moreover, artificial intelligence and machine learning proved their effectiveness and ability in decision-making and early prediction of the patient's condition. Most researchers take advantage of this science and use its tools to predict the patient's condition at an early stage. One of the most important of these diseases is diabetes. The prevalence of diabetes is higher than that of other diseases since it is a chronic, non-communicable illness[1]. Additionally, it is regarded as one of the leading global causes of death, along with other diseases such as heart disease and cancer[1][2]. The International Diabetes Federation estimates that there would be roughly 700 million diabetics worldwide in 2045[3]. According to the World Health Organization, 380 million people will have diabetes worldwide by 2025[4]

Table 1.1. Types of diabetes

Common types of diabetes	Complications	Diagnosis of diabetes
Type 1 diabetes	1- Gradual rise in blood pressure 2- Low protein and cholesterol 3-Damage to the kidneys, the retina of the eye, and the nervous system 4- Disorders in the blood lipids	1- blood tests 2- Tests to detect gestational diabetes 3- Tests to detect diabetes
Type 2 Diabetes	1- convulsions 2- coma 3- Sudden drop and rise in blood sugar	
Gestational Diabetes	1- Overgrowth 2- family history 3- Women over 25 years old	

There are many complications of diabetes, as shown in Table 1.1., and therefore early detection of this disease helps in avoiding many complications. Therefore, in this study, we will try to apply a new model to detect the largest possible number of infected patients.

1.2. MOTIVATION OF THESIS

Non-communicable and chronic diseases usually require a long time to treat, and these diseases cannot be fully cured. Among these are cardiovascular diseases, respiratory diseases, and diabetes of all three types (type 1 diabetes, type 2 diabetes, and gestational diabetes). According to the world health organization[5], these diseases are considered one of the most common causes of death worldwide. There has been a noticeable increase in the number of people infected with these diseases in recent years, according to the reports of the Director of the World Health Organization, who called for taking practical steps to combat these diseases, which are responsible for 17 million premature deaths yearly.

Deep learning and machine learning algorithms are considered one of the most important modern tools to help evaluate and detect these diseases. Many studies have been conducted on the use of artificial intelligence and machine learning services in

the medical field. Machine learning classifiers rely on medical data sets to detect diseases. There are many data sets, including balanced, unbalanced, structured, and unstructured. The accuracy of any algorithm depends mainly on the data set[6], as the algorithm can give high accuracy if the data set is organized and balanced and lower accuracy if the data set is unbalanced or organized. The primary motivation of this research is to find the best algorithms that give the best accuracy on diabetes datasets.

1.3. PROBLEM STATEMENT

The rapid spread of chronic diseases, especially diabetes, made doctors and specialists think of alternative solutions to detect the disease instead of traditional methods. Reliance on traditional methods in identifying diseases is high in cost, time, and error rate in assessing the patient's condition. Modern software systems have become one of the most important solutions to this problem and bridge the gap. In recent years, there has been an actual focus on using these systems to detect and identify chronic diseases early. These systems do not require high financial costs and take little time to determine the type of disease. Although Several studies worked on building models to detect these diseases in different ways, the accuracy needed improvement. Therefore, in this study, we build an intelligent model by integrating several machine learning algorithms capable of detecting diabetes at an early stage with high accuracy.

1.4. OBJECTIVES

The main objective of this thesis was to determine the best way to detect diabetes. The three best machine learning algorithms were used to form our model. The objectives can be described as follows:

- To Compare the performance of the most common six machine learning algorithms in detecting and identifying diabetes at an early stage.
- To Compare of performance and results of machine learning algorithms on balanced and unbalanced diabetes dataset

- To Compare the results obtained using this model with the trained models used in previous studies.
- A comparison between the performance of the proposed hybrid model with a neural network designed during the study

1.5. CONTRIBUTION

The following points can determine contribution:

- A hybrid model was built that can safely predict the patient's condition by relying on several machine learning algorithms that work together to make the final decision in assessing the patient's condition.
- The results of the proposed model were compared with some previous studies. The accuracy scale and the Recall scale of the proposed model gave better results than those studies that have been discussed in the second part of this study.
- A deep neural network containing hidden nodes was built and its performance was compared with the proposed model.
- Data processing in the proposed model takes little time.

1.6. ORGANIZATION OF THESIS

In the first part of this study, an overview of diabetes and the importance of doing this study was given. The second part presents works and studies related to the current study. Moreover, in the third part, the importance of machine learning was explained. The algorithms used in this study were explained in the fourth part, including the proposed methodology. Furthermore, the results were presented and discussed in the fifth part. Finally, the last part concludes the work carried out during this study and its benefits.

PART 2

LITERATURE REVIEW

In this part, some previous studies and works related to this study will be presented. The techniques and algorithms used to detect diabetes will be presented. Also, the previous works will be presented on a different data set such as the Pima dataset and another data set obtained in different ways. After that, these works will be compared in terms of accuracy, Recall, and Precision.

2.1. PREVIOUS STUDIES USED MACHINE LEARNING ALGORITHMS

Diabetes prediction method and recommendation system suggested by Nagaraj et al. This method involves building a machine learning model for prediction utilizing a variety of methods, including XGBoost, support vector machine (SVM), random forest classifier, and decision tree. The accuracy of the random forest classifier is 77%. After anticipating the disease, create a system of recommendations, include a diabetic recovery diet and foods, and offer some disease-fighting exercises.[7].

Shanjida khan Maliha et al used two diabetes prediction algorithms, the Random Forest and the Support Vector Machine. And it relied on real patient data to determine the disease. It was implemented by the programming language Python and Jupyter. The Support Vector Machine algorithm gave a higher accuracy than the Random Forest, where the accuracy reached 86%, while the accuracy of the Random Forest reached 78% [4].

Herminiño C. Lagunzad et al relied on a database from Kaggle and applied the ID3 algorithm to this data to predict diabetes. As for the results, it can be for people between 30-40 and 41-50 with diabetes. Delayed recovery and sudden weight loss can also be a sign of diabetes[8] .

Saad Ebrahim Saeed et al compared three diabetes determination algorithms: SVM, KNN, and DT. It was based on a database from the UCI repository. The result was that DT gave the highest accuracy, reaching 96.0%, then KNN, with an accuracy of 93.3%, then SVM, with an accuracy of 92.1% [9].

Vivek Vaidya et al used a soft computing classifier with different parameters and the classifier's accuracy reached 95.07%[10].

Salma Karimah et al. developed a classification model to predict diabetes based on the content of a drug review using Random Forest. N-gram and Term-Frequency Inverse Document Frequency In the feature orientation stage. The model gave an accuracy of F-1 score value of 0.952 using the unigram feature [11].

Amrutha P and others suggested an early system for diagnosing diabetes and used five machine learning algorithms. The accuracy of the system was 94.350 percent using the IBK algorithm, then the Random Forest algorithm came in second place, with an accuracy of 93.785% [12].

A.Prakash et al. implemented several machine learning algorithms on the PIMA Indian Diabetes database for diabetes prediction. The Naïve Bayes algorithm gave the highest accuracy, reaching 89.9[13].

Akhyar Ali Khan et al. proposed the GBA algorithm for diabetes prediction and used the PIMA Indian Diabetes database. The proposed algorithm has an accuracy of 0.92[14].

Minhaz Uddin Emon et al. implemented eight machine-learning algorithms on the Diabetes Retinopathy Debrecen dataset to predict diabetes Retinopathy. The logistic regression algorithm gave an accuracy of 75% [15].

Zainab T. Al-Ars et al developed a model for predicting diabetes in which classification was made based on HbA1c measurement and initial diagnosis. In his work, he relied on four algorithms, and the best result was obtained by the

RANDOM FOREST algorithm, reaching 93.51%. Then a new method called CBRF (Clustering Based Random Forest) was proposed. Which combines RANDOM FOREST and K-Means. Experiments showed that the construction time was reduced to 50 percent and the accuracy increased to 94 [16].

Rakesh S Raj et al. extracted a data set from the health reports of diabetic patients. Then they applied machine learning algorithms to predict the disease. The Naïve Bayes algorithm gave an accuracy of 62.5%, while the SVM algorithm gave an accuracy of 82%[17].

Mehmet Bilgehan Erdem et al. proposed a multipurpose genetic programming symbolic regression algorithm to predict diabetes, and the research focused on the level of complexity and accuracy of the proposed model. He then compared this method with several other algorithms. The Majority-Voting Scheme algorithm gave the highest accuracy, reaching 81.64%, but this algorithm, according to the research, is the most consuming in resources and time. While the proposed method gave an accuracy of 79.17% with less complexity and cost than the rest of the proposed algorithms [18].

Priyanka Sonar and others have experimented with several machine-learning algorithms to predict diabetes. The SVM algorithm and the Artificial Neural Network gave the highest accuracy, reaching 82%. The Naive Bayes algorithm gave 80% accuracy, while the Decision Tree algorithm had 74% accuracy[19].

2.2. PREVIOUS STUDIES USED DEEP LEARNING

Luyao Xu et al. proposed a method based on deep learning to build a 1DCNN model for predicting diabetes. He compared the model with machine learning algorithms such as (Naive Bayes, and Random Forest) and was based on a database collected from Sylhet Diabetes Hospital in Sylhet. The accuracy of the proposed method reached 97.02, while the accuracy of random forest reached 94.06, and the accuracy of Naive Bayes reached 91.09[20].

Shiva Shankar Reddy et al. applied a deep learning algorithm DBFN to predict the readmission of diabetic patients to hospital and compared this algorithm with other algorithms. This algorithm gave an accuracy of 0.6917, while the accuracy of the rest of the algorithms was less than this accuracy. It was also observed that the patient is readmitted to the hospital. again after 90 days [21].

2.3. PREVIOUS STUDIES USED THE HYBRID MODEL

Sarra Samet and others applied machine learning algorithms to the Pima database and then applied a hybrid model and found that the hybrid model gave a higher accuracy as its accuracy reached 90.62%[22].

Biswajit Giri et al. used a hybrid approach to predict diabetes based on the PIMA database. He compared this method to several other algorithms. The proposed method gave an accuracy of 86 percent, while the closest algorithm was 75 percent accurate, which is the Linear Support Vector Classifier algorithm [23].

2.4. DIABETIC DATASETS FROM SEVERAL SOURCES

Table 2.1. presents different types of diabetes datasets as well as the algorithms that were used on this data and the results of these algorithms.

Table 2.1. Studies using ML

Ref	Dataset	KNN	Neural Network	Logistic Regression	Decision Tree	Random Forest	Naive Bayes	SVM
[7]	Pima				71%	77%		74%
[4]	Khulna Hospital					78%		86%
[9]	Sylhet Hospital	93.3%			96%			92.1%
[10]	Pıma		92.17 %					
[22]	Pıma	88.31%		78.57 %	85.71%	83.76 %	77.27%	87 %
[12]	reconnaissance					93.78%	85.87%	
[13]	Pima			84%	80.8%	80.9%	89.9%	74.4%
[24]	Medical Chittagong	70%					67%	69%
[25]	UCI				78%	81%		89.67%

Table 2.2. presents some previous. It displays the algorithm that gave the best result. It also displays Accuracy, Recall, and Precision metrics.

Table 2.2. Studies used a hybrid model and machine learning.

Ref	Year	Algorithm	Accurracy	Recall	Precision
[7]	2022	Random Forest	77%		
[26]	2021	LightGBM + KNN	90.1%	82.1%	88.9%
[10]	2021	Firefly Optimized Neural Network	95.07	88%	88%
[27]	2021	Random Forest	79%	77%	77%
[22]	2021	KNN+SVM+DT	90.62	91%	91
[13]	2021	Naïve Bayes	89.9	84.3	79.3
[14]	2021	Gradient Boosting	92	93	94
[28]	2020	K-Means	86		
[23]	2020	Hybrid classifier	86		
[18]	2019	Genetic Programming Symbolic Regression	79.19%		
[19]	2019	Artificial Neural Network	82		
[29]	2019	Decision Tree	70.80	61.46	76.5
[30]	2019	J48	95.122		
[31]	2018	SVM	79.1	79.1	78.2
[32]	2017	ANN	80.86		

PART 3

THEORETICAL BACKGROUND

In this part, we will try to present an overview of the importance of machine learning and its algorithms and explain the reasons for the necessity of using machine learning techniques in the medical field.

3.1. MACHINE LEARNING

One of the most crucial subfields of artificial intelligence is machine learning, which uses mathematical techniques and equations to enhance a machine's performance over time. The study of computer algorithms that may automatically improve during experiments on data and after the conclusion of data training on data results in the model is the focus of automated learning. The distinctions between AI, machine learning, and deep learning are shown in Figure 3.1. defines deep learning, machine learning, and AI[33].

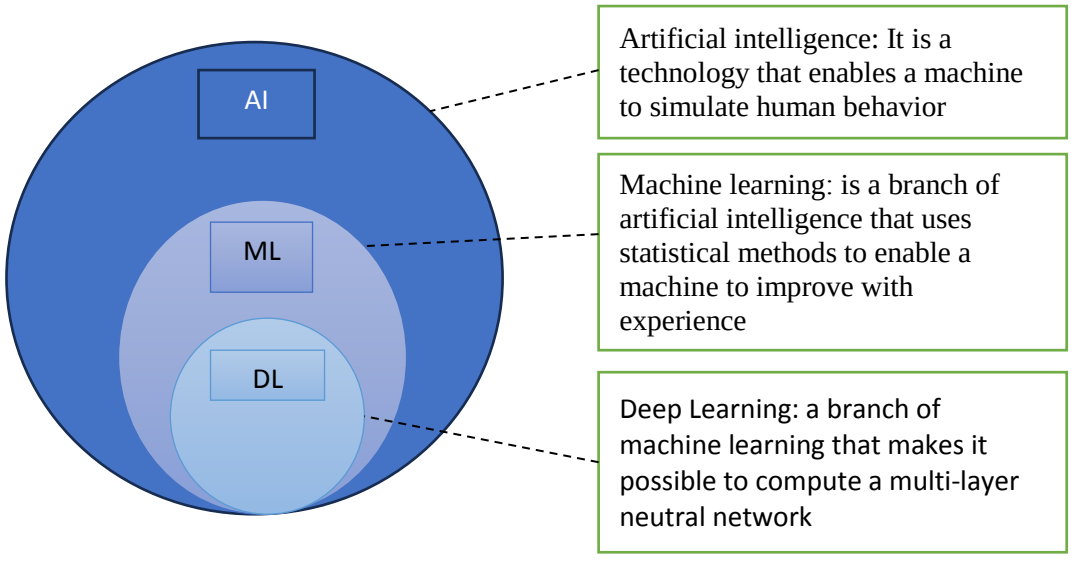
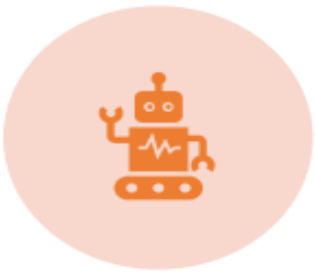


Figure 3.1. Shows the difference between AI, machine learning, and deep learning

3.1.1. The Main Elements Of Machine Learning

Table 3.1.Presents the basic elements of machine learning[33]

Data 	is information in the form of texts, audio, photos, and so forth.
 Learning	The method finds the model after analyzing the data.
Training data 	utilizing the data to train an algorithm to find a model.
Model 	is the final product of Machine Learning.

3.1.2. TYPES OF MACHINE LEARNING

Machine learning is built on the principle of supervised learning, in which algorithms are trained to make predictions by being given predetermined inputs and outputs (features and targets, respectively). The second method, referred to as unsupervised learning, trains algorithms to generate predictions based on given data without being given any information on the outcomes (goal). Throughout the training phase, algorithms discover connections and patterns in previously unobserved data that help them make predictions. There are no further connections between the first two forms of learning and the third, reinforcement learning. A goal-driven agent looks around its environment. As it moves through its environment, it makes a few decisions. Agent will be compensated positively if his decision helps him achieve his goal; otherwise, Agent will be compensated negatively. Figure 3.2. shows the types of machine learning[34] [33].

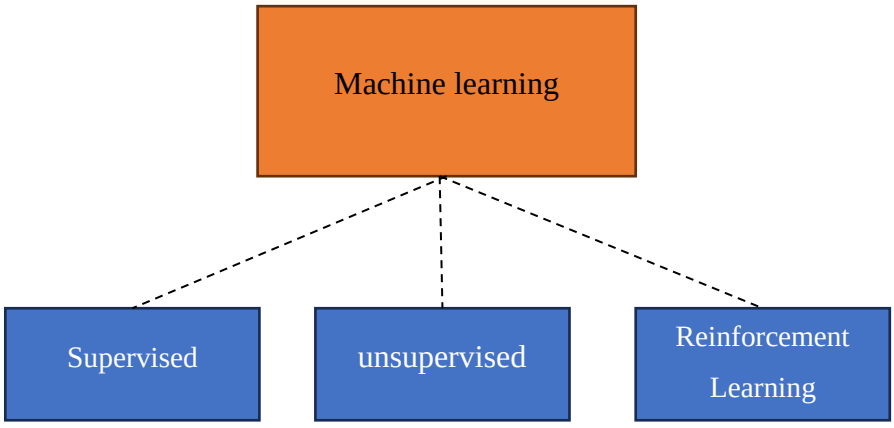


Figure 3.2. Presents the types of machine learning.

3.1.3. Supervised Learning Applications

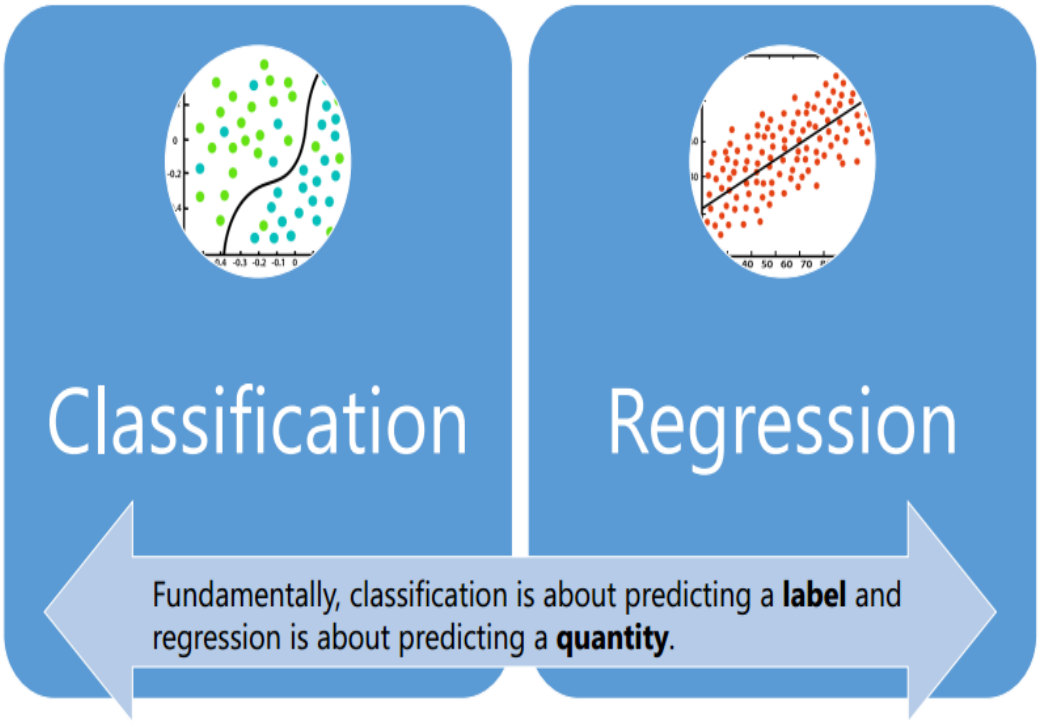


Figure 3.3. Shows machine learning applications.

Since classification is one of the most crucial techniques in supervised machine learning, it is widely employed. Two different categories can be used to categorize things. The prognosis for, and is it Chronic Diseases?, are two categories that are classified using the first kind. As a result, the classification method that is based on just two categories is known as Binary classification, and the prediction is (Chronic Diseases or Not Chronic Diseases).

Regression is a statistical technique used in finance, investing, and other fields that aims to ascertain the nature and strength of the relationship between a single dependent variable (often represented by Y) and a number of additional factors (sometimes referred to as independent variables).

3.1.4. Classification Vs. Clustering

Figure 3.4. presents the hypothesis between Classification vs. Clustering[35][33].

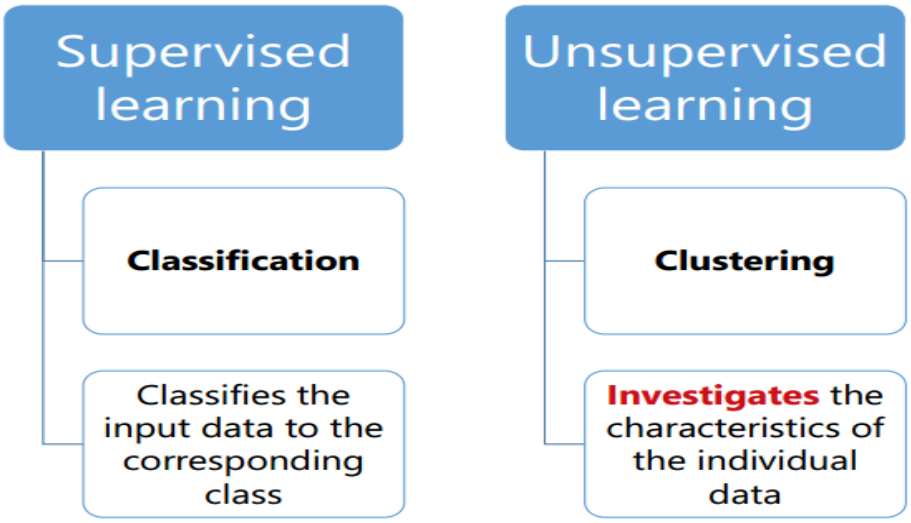


Figure 3.4. Presents the difference between Classification and Clustering

3.1.5. MACHINE LEARNING ALGORITHMS

3.1.5.1. Bayes classifier

Bayes classifier: is a statistical classification as it expects the possibility of any row to be a specific Class. This work is based on the theory of Bayes[36], which assumes the effect of a characteristic of a specific category that is independent of the values of features. Equation 3.1 expresses Bayes' theorem [37].

$$p(A|B) = \big(p(A).p(B|A)\big)/p(B) \qquad (3.1)$$

p (A | B): the likelihood that both event A and event B will happen.
p(B | A): the likelihood of event B assuming event A has already happened.
p(A): the likelihood of incident A.
p(B): the probability of event B.

3.1.5.2. Random Forest

In this method, all classifiers in the integrated set are a decision tree classifier so that all these classifiers together form a forest. Individual decision trees are generated using a random set of attributes and during the classification process, each existing decision tree has the right to vote[38][39]. The one with the highest percentage of votes will be the result of the classification process.Figure 3.5. shows an example of the random forest algorithm[40].

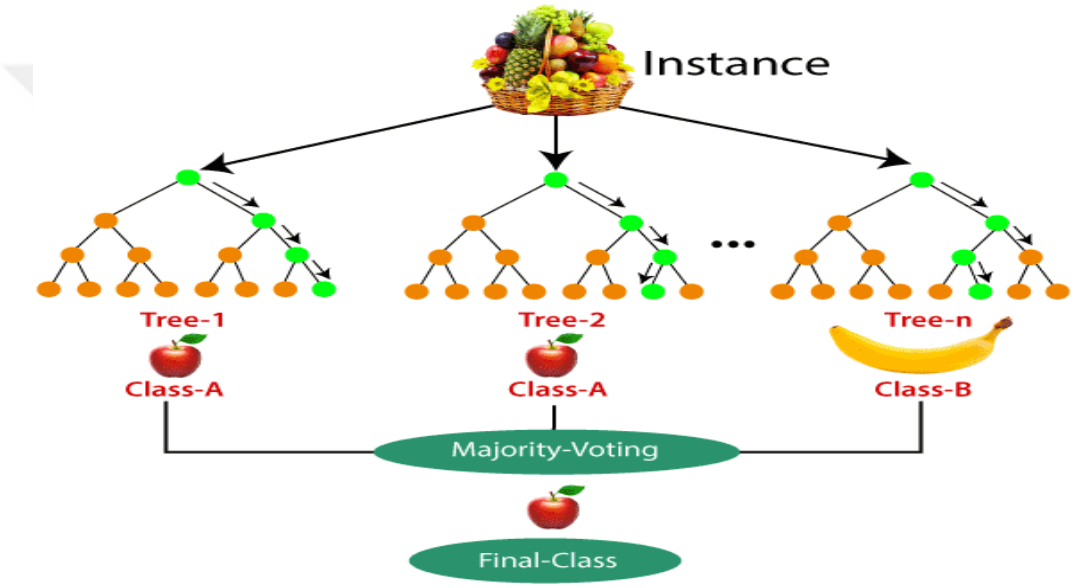


Figure 3.5. Shows how the random forest algorithm works

3.1.5.3. Decision Tree

A decision tree is a tree structure resembling a flowchart, where each leaf node has a class label and each inner node indicates the selection of an attribute. The root node of the tree is its highest node.[41].

Decision tree construction does not require any knowledge domain and is therefore suitable for exploratory knowledge discovery. This algorithm can deal with multidimensional data[42]. Figure3.6. demonstrates a decision tree algorithm in action[43].

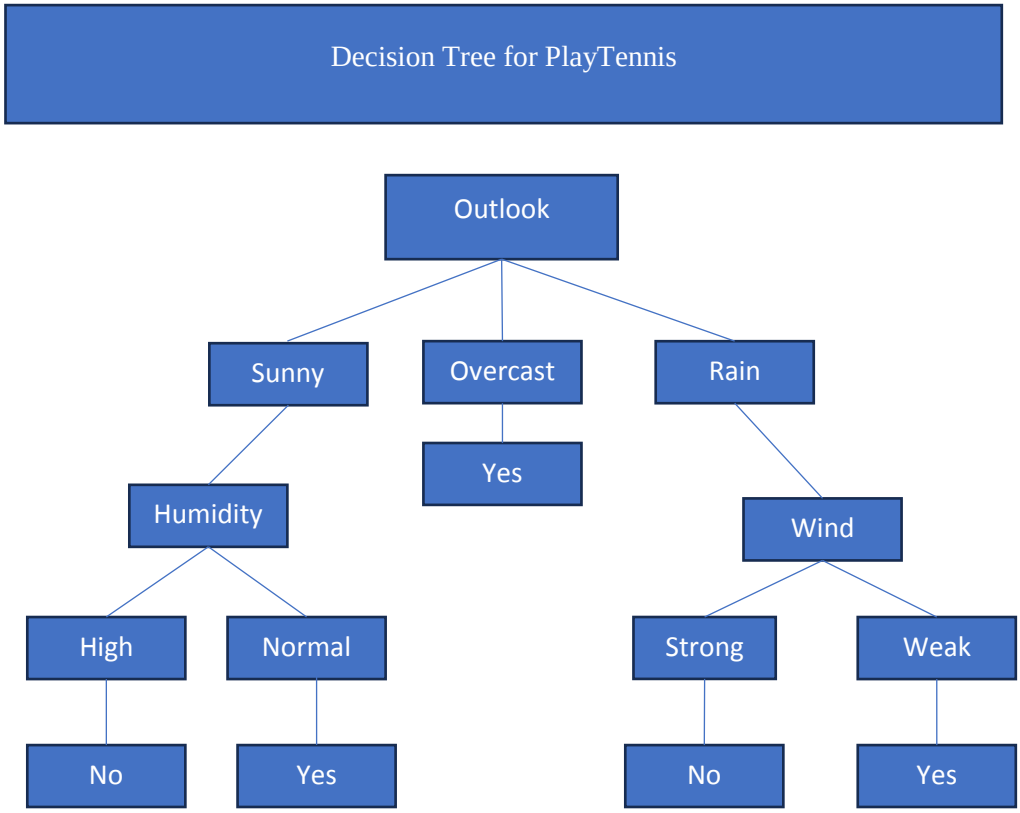


Figure 3.6. Demonstrates a decision tree algorithm in action[43].

3.1.5.4. KNN

KNN: One of the simplest algorithms works with a supervisor. Can handle abnormal data. Its principle of operation is based on calculating the Euclidean distance between points [44]. The K parameter refers to the number of neighbors, for example, if K = 5, the algorithm will calculate the distance between the point to be classified and the five nearest neighbors. If there are three or more points belonging to a certain category, the target point will be classified as belonging to this category[45]. Figure 3.7. demonstrates a working instance of the KNN algorithm [46].

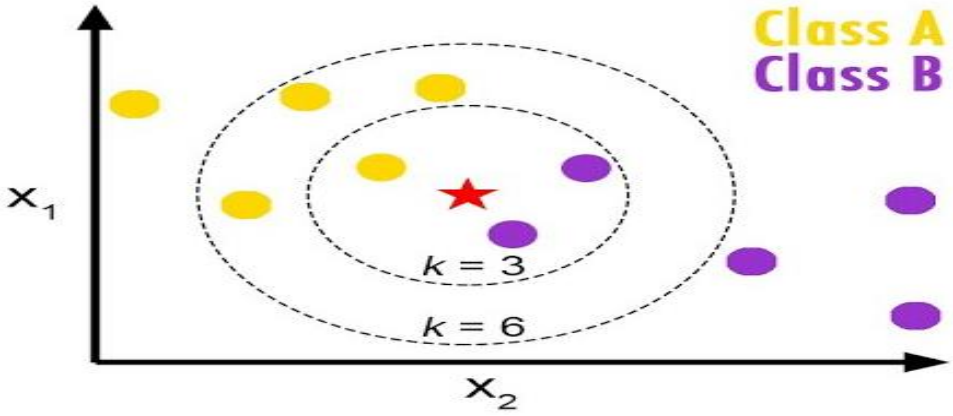


Figure 3.7. Demonstrates a working instance of the KNN algorithm[46].

3.1.5.5. SVM

SVM is an algorithm for supervised machine learning. It can be applied to a variety of problems, including classification and regression.. The main idea of this algorithm is the super level that separates the different classes[47]. Figure3.8. demonstrates a working instance of the SVM algorithm. [48].

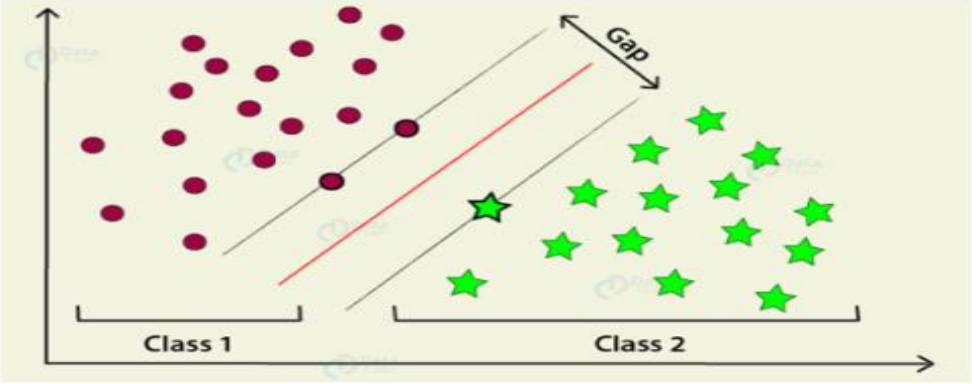


Figure 3.8. Demonstrates a working instance of the SVM algorithm[48].

3.1.5.6. Gradient Boosted

Gradient Boosted: This algorithm adds prediction models sequentially. The new model corrects the one that preceded it[49]. Figure 3.9. demonstrates a working instance of the Gradient Boosted method[50].

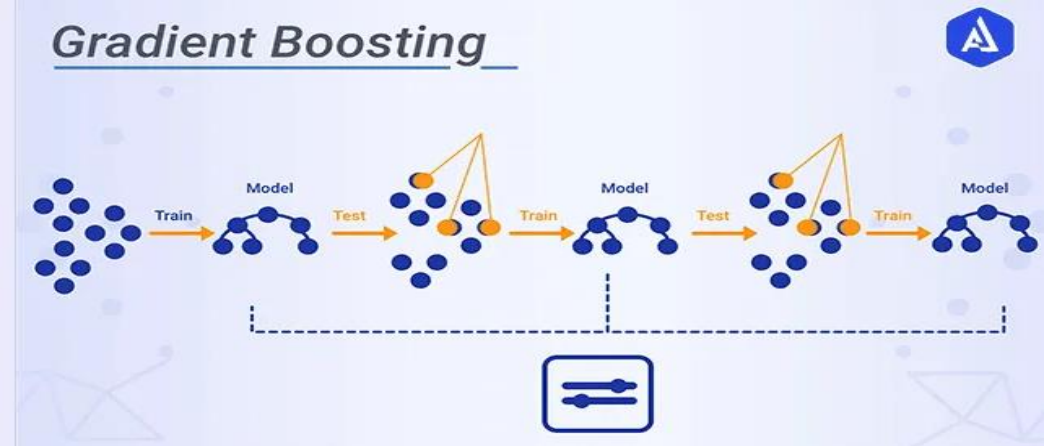


Figure 3.9. Demonstrates a working instance of the Gradient Boosted method[50].

3.1.5.7. Regression algorithm

Regression algorithm is a sort of statistical modeling that enables you to determine whether one thing (variable) is dependant on others. The relationship between variables is demonstrated by a trend line superimposed on your data and may be used to forecast a wide range of outcomes[51]. Figure 3.10. demonstrates a working of the Regression method. [52].

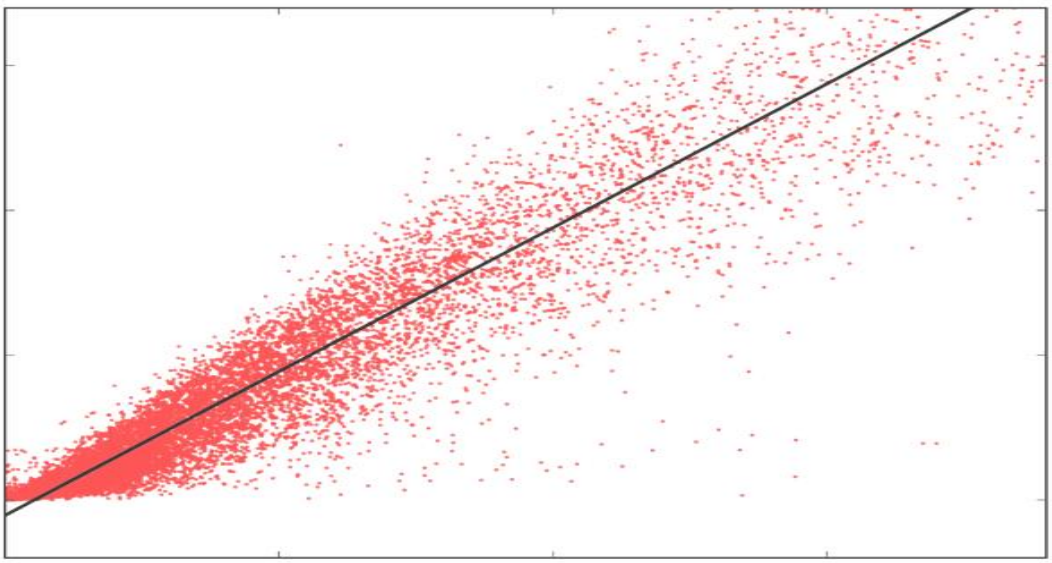


Figure 3.10. Demonstrates a working of the Regression.

3.2. PERFORMANCE METRICS

After the results are obtained, a formal evaluation of these results should be done. Evaluation is to test the predictive capabilities of the model obtained on new data to see the effectiveness of the algorithms used in building the model. The model is usually evaluated through the confusion matrix listed in Table 3.2. Where tn and tp represent the number of correctly labeled negative and positive samples, and fn and fp represent the number of falsely evaluated negative and positive samples[53].

Table 3.2. Confusion Matrix.

Confusion Matrix	**p (Predicted)**	**n (Predicted)**
P (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

P, N represents the existing classes

TP(True Positive): the number of states correctly predicted by the classifier for the p-class and belonging to the p-class.

FN (False Negative): The number of states incorrectly predicted by the classifier of class p and belonging to class P.

TN (True Negative): The number of states correctly predicted by the classifier for class n and belonging to class n.

FP (False Positive): the number of states incorrectly predicted by the classifier for class n and belonging to class n.

Table 3.3. Presents some metrics for evaluating the model.

Accuracy	Accuracy is the number of samples that the model correctly predicted over the number of samples evaluated.	tn+tp / tn+tp+fn+fp
Recall (r)	The recall is the proportion of positive samples that were correctly labeled.	tp / tp+fn
Precision (p)	Precision is the proportion of positive samples inside a positive class that is accurately predicted from the total number of positive samples anticipated.	tp / tp+fp
F-Measure	It is a representation of the harmonic mean between the values for recall and precision.	2*p*r / p+r

PART 4

METHODOLOGY

The proposed system goes through several stages until it reaches the stage of detecting or diagnosing diabetes. The first stage explores the data set and then pre-processes the data. After that, diabetes is detected using two different methods. The first method is using machine learning algorithms, and the second method is using deep learning technology. In the first method, we train six machine learning algorithms on the training data, then test these algorithms on the test data set, after that the three best algorithms are chosen to build the hybrid model, and then the patient is classified as having diabetes or not. The second method involves designing a neural network that contains more than one hidden layer, training it on the training data set, testing it on the test data set, then classifying the patient whether they have diabetes or not, and in the final stage, the results are evaluated and a comparison is made between the two methods.

4.1. THE PROPOSED MODEL

The following figure displays the proposed hybrid model. The hybrid model consists of three algorithms (KNN, Random Forest, Gradient Boosted). The hybrid model decides that the case that is diagnosed as having diabetes if it is classified by two or more classifiers as having diabetes. The hybrid model gives a decision that the diagnosed case is healthy and does not have diabetes if the diagnosed case is classified by two or more classifiers as normal and does not have diabetes.

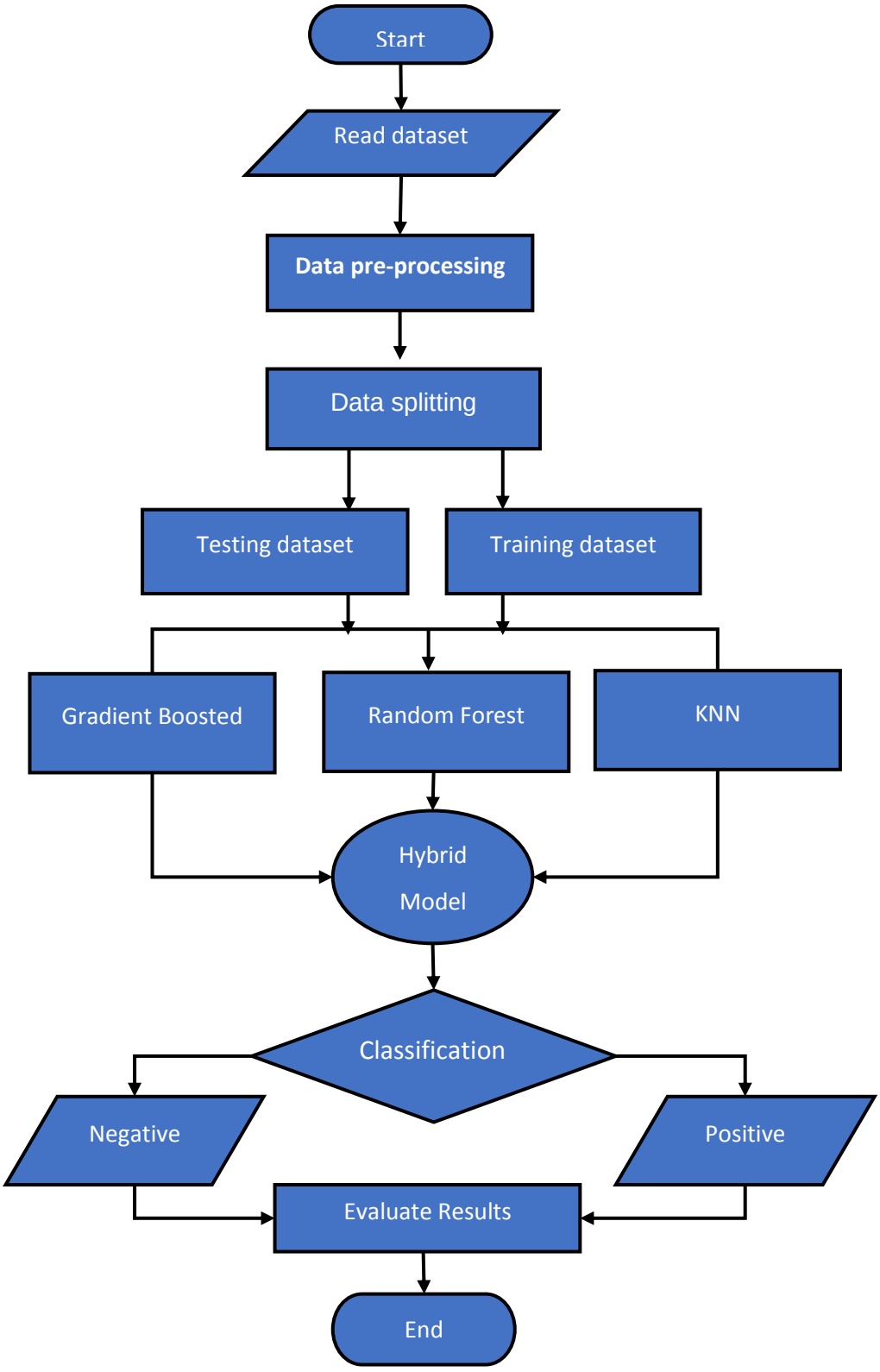


Figure 4.1. Proposed Hybrid Model.

4.1.1. Dataset

There are many data sets for diabetics. Such as the data set of Sahlit Diabetes Hospital in Bangladesh and other different data sets. In this study, we tried to search for a reliable data set that can be relied upon to obtain accurate results. The Pima data set was relied upon. This data set is a standard for researchers to classify diabetes. The National Institute of Diabetes and Digestive and Kidney Diseases in America gathered the PIMA data set. The data set contains data for 768 American women. The number of infected cases is 268, and the number of healthy cases is 500. The dataset contains 9 features. Table 4.1. gives an overview of these features.

Table 4.1. Presents the features of the dataset

Feature	**Data Type**
Pregnancies	İnteger
Glucose	İnteger
BloodPressure	İnteger
SkinThickness	İnteger
Insulin	İnteger
BMI	Double
DiabetesPedigreeFunction	Double
Age	İnteger
Outcome	İnteger

4.1.2. Data exploration

Table 4.3. the arithmetic mean, maximum value, and lowest value for each characteristic, among other crucial details about the dataset.

Table 4.2. Statistical information about features

Column	Minimum	Maximum	Mean	Standard Deviation	Variance
Pregnancies	0	17	3.8	3.370	11.354
Glucose	0	199	120.9	31.973	1022.25
BloodPresssure	0	122	69.1	19.356	374.647
SkinThickness	0	99	20.5	15.952	254.473
Insulin	0	846	79.8	115.244	13281.9
BMI	0.078	67.100	31.9	7.884	62.160
DiabetesPredigreeFunction	21	2.420	0.47	0.331	0.110
Age		81	33.2	11.760	138.303

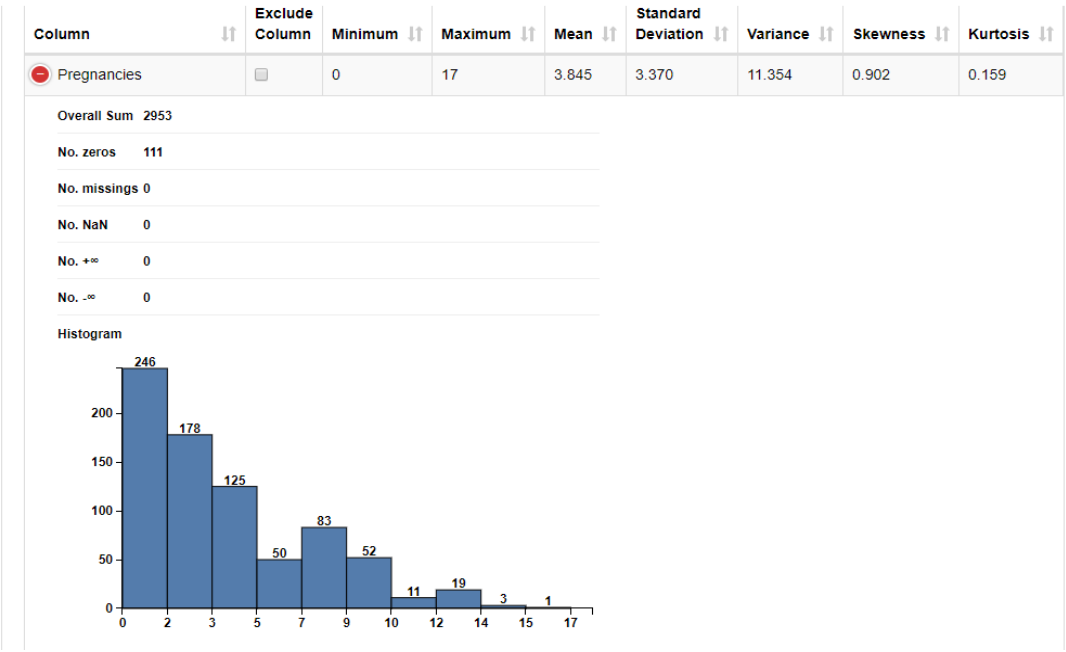


Figure 4.2. Statistical information about the Pregnancies feature.

Glucose		0	199	120.895	31.973	1022.248

Skewness	0.174
Kurtosis	0.641
Overall Sum	92847
No. zeros	5
No. missings	0
No. NaN	0
No. +∞	0
No. -∞	0

Histogram

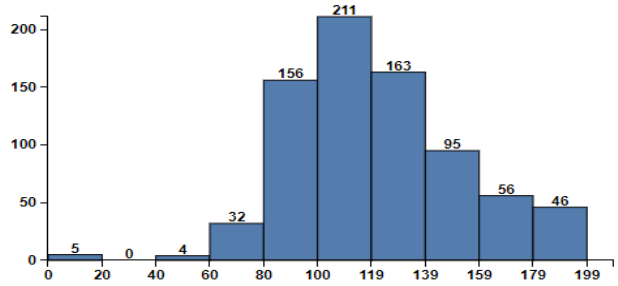


Figure 4.3. Statistical information about the Glucose feature.

BloodPressure		0	122	69.105	19.356	374.647

Skewness	-1.844
Kurtosis	5.180
Overall Sum	53073
No. zeros	35
No. missings	0
No. NaN	0
No. +∞	0
No. -∞	0

Histogram

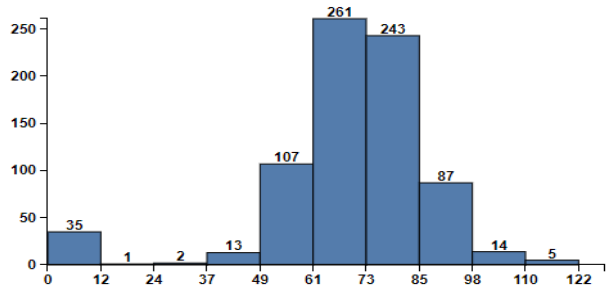


Figure 4.4.Statistical information about the BloodPressure feature.

SkinThickness		0	99	20.536	15.952	254.473

Skewness	0.109
Kurtosis	-0.520
Overall Sum	15772
No. zeros	227
No. missings	0
No. NaN	0
No. +∞	0
No. -∞	0

Histogram

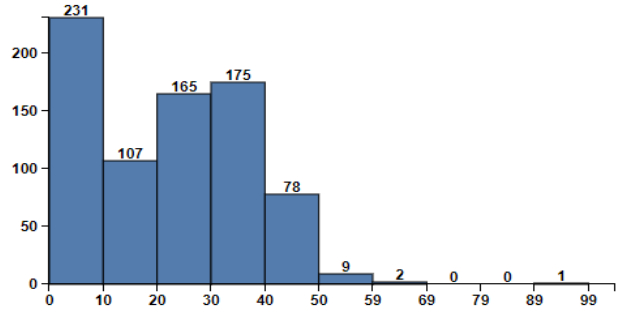


Figure 4.5.Statistical information about the SkinThickness feature.

Insulin		0	846	79.799	115.244	13281.180

Skewness	2.272
Kurtosis	7.214
Overall Sum	61286
No. zeros	374
No. missings	0
No. NaN	0
No. +∞	0
No. -∞	0

Histogram

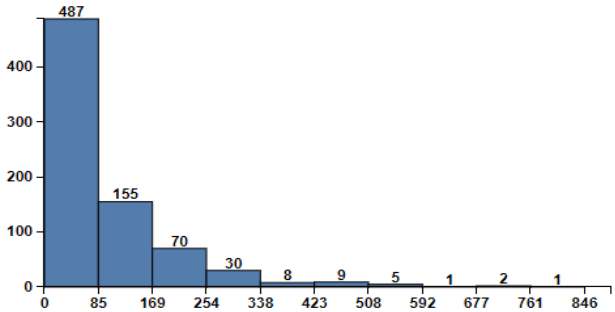


Figure 4.6. Statistical information about the Insulin feature.

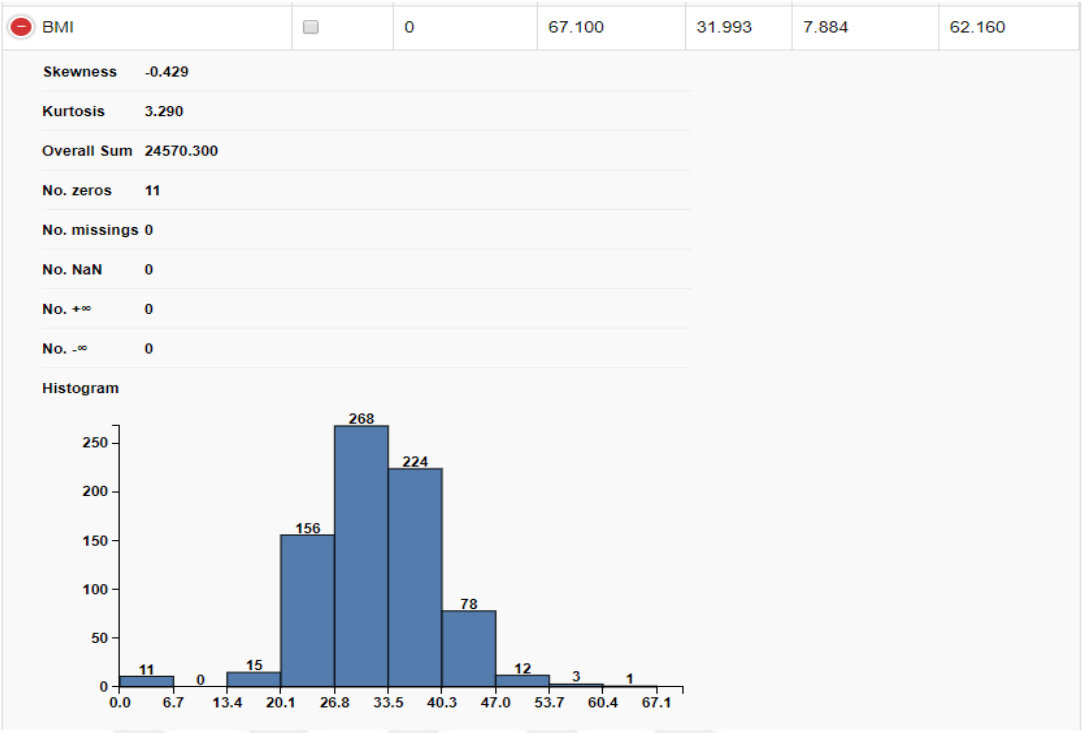


Figure 4.7.Statistical information about the BMI feature.

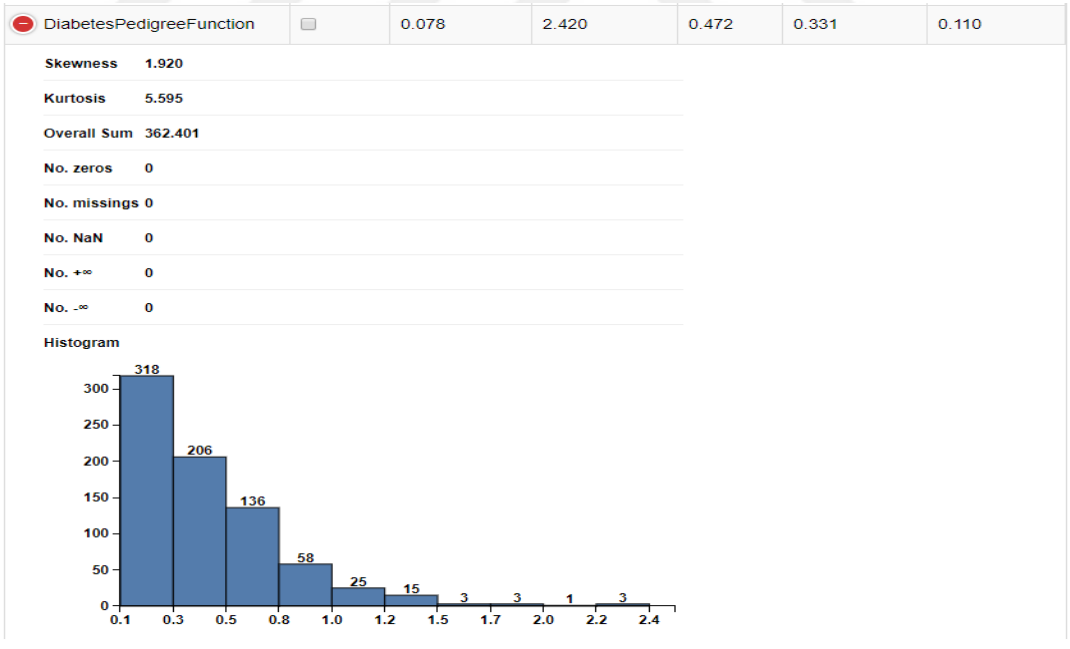


Figure 4.8.Statistical information about the DiabetesPedigreeFunction feature.

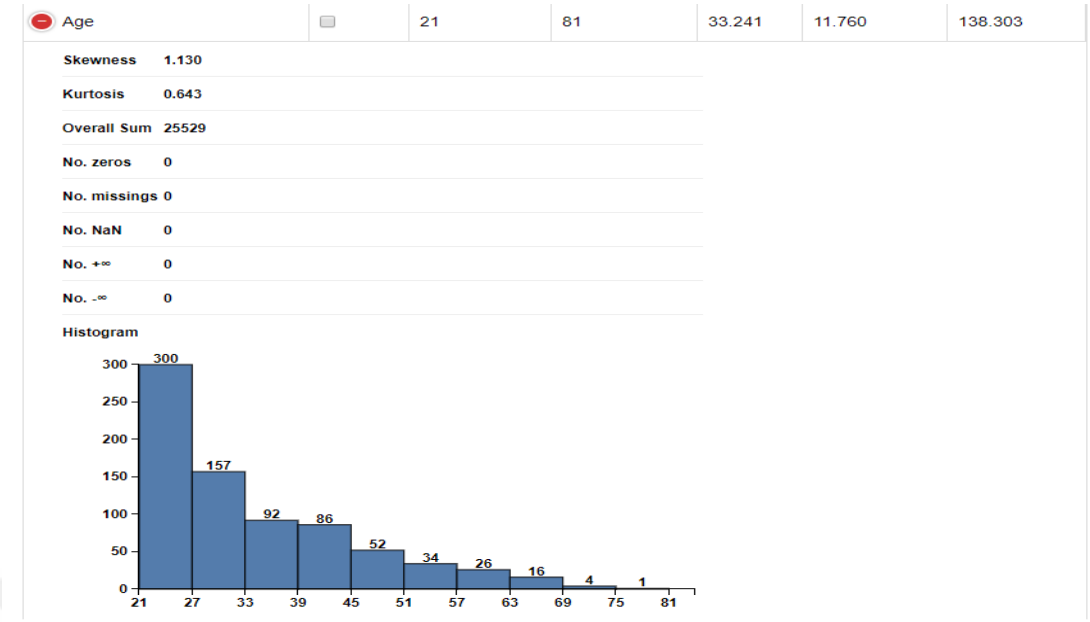


Figure 4.9.Statistical information about the Age feature.

4.1.2.1. A bar Chart

A bar chart is a type of graph that shows rectilinear categorical data. Figure 4.10. shows the bar graph of the Pima data set, with the heights of the columns corresponding to the values they reflect.[54].

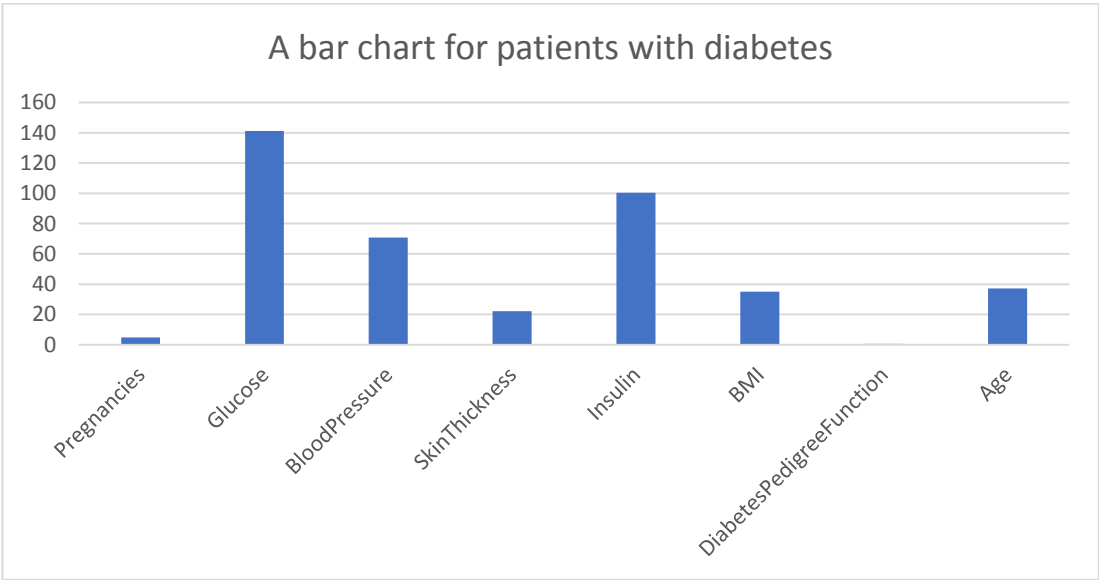


Figure 4.10.A bar chart for patients with diabetes

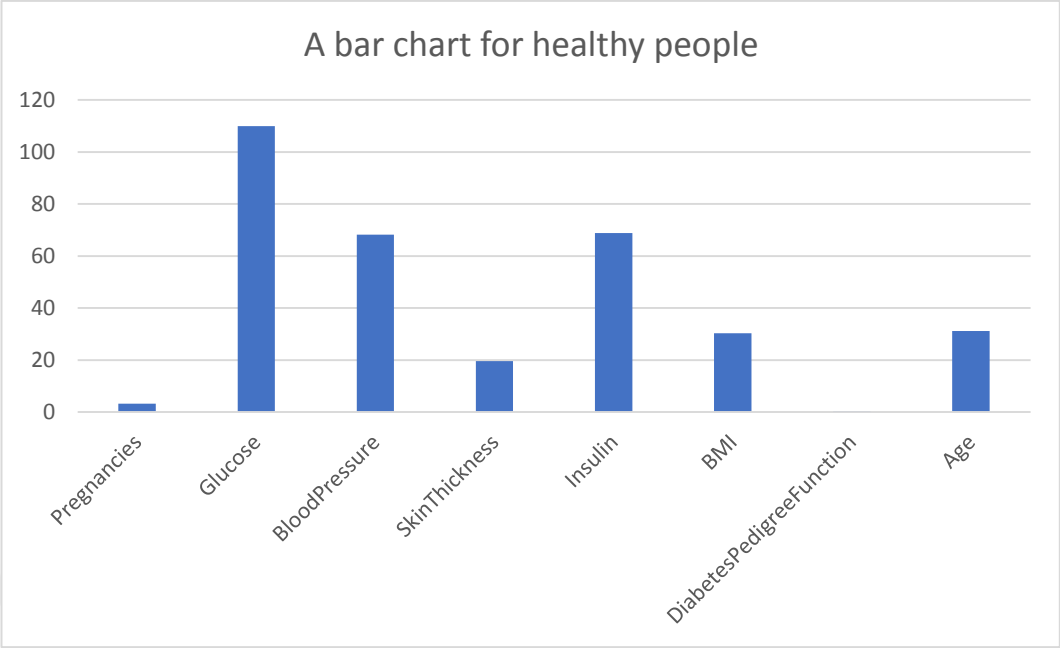


Figure 4.11. A bar chart for healthy people.

4.1.2.2. A scatter plot

A scatter plot is a type of graph that uses the Cartesian coordinates of two variables from a data set to display their values. Where the main objective of using the scatter plot is to know the form of the relationship between these two variables[55]. The following figures display the scatter plot among the features of the Pima data set.

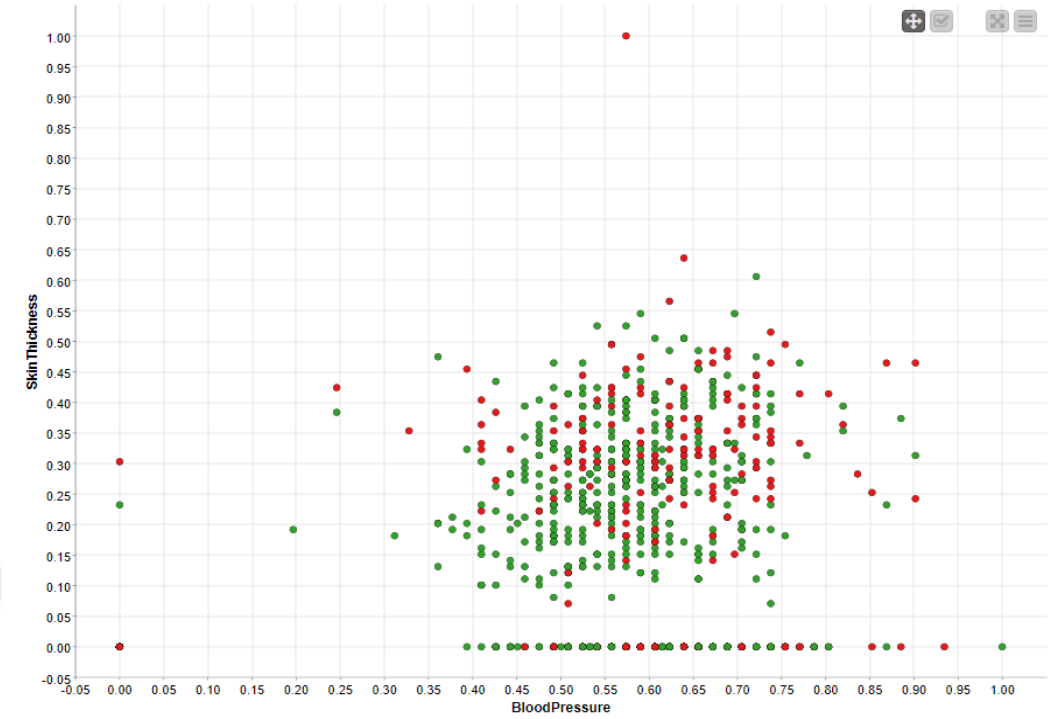


Figure 4.12. Presents the dispersion scatter plot between Skin Thickness and Blood Pressure.

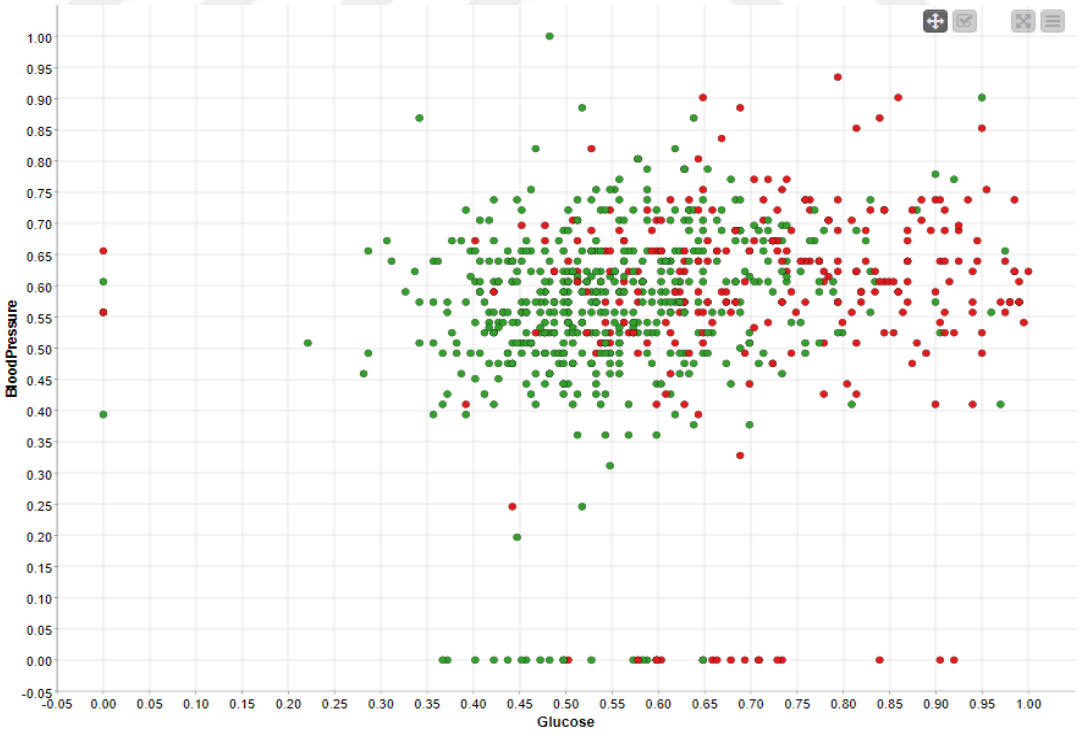


Figure 4.13. Presents the dispersion scatter plot between Glucose and Blood Pressure.

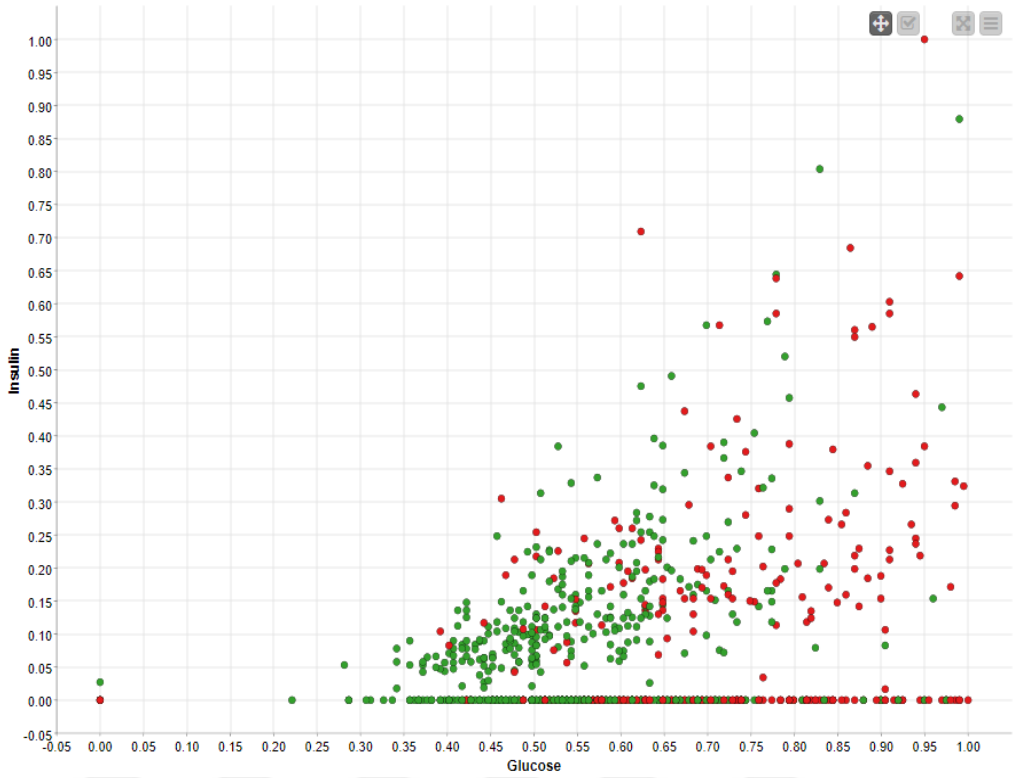


Figure 4.14. Presents the dispersion scatter plot between Insulin and Glucose.

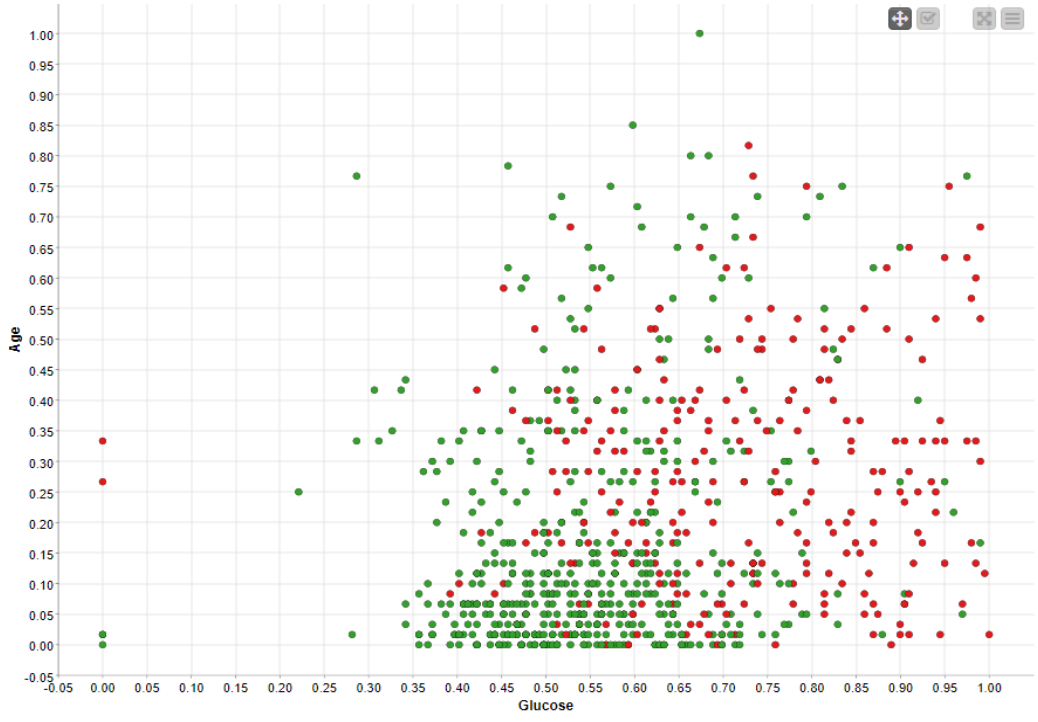


Figure 4.15. Presents the dispersion scatter plot between Age and Glucose.

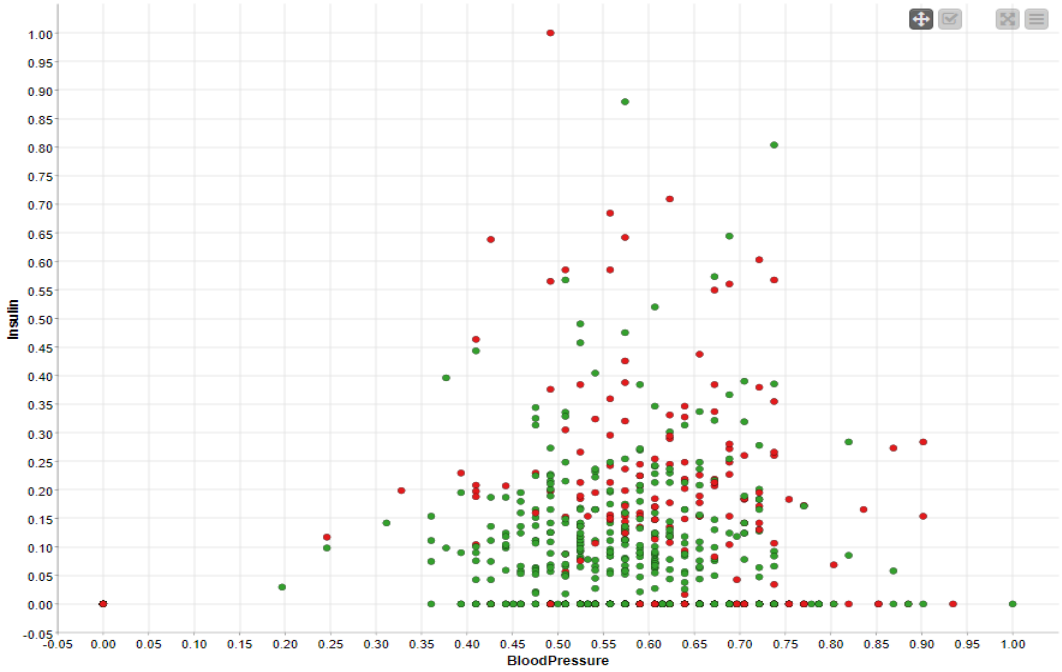


Figure 4.16. Presents the dispersion scatter plot between Insulin and BloodPressur.

4.1.2.3. Stacked Area Chart

Stacked Area Chart: This chart aims to analyze and see how the variables differ. It is plotted as stacked series, and the value in the data point determines the height of the chain.

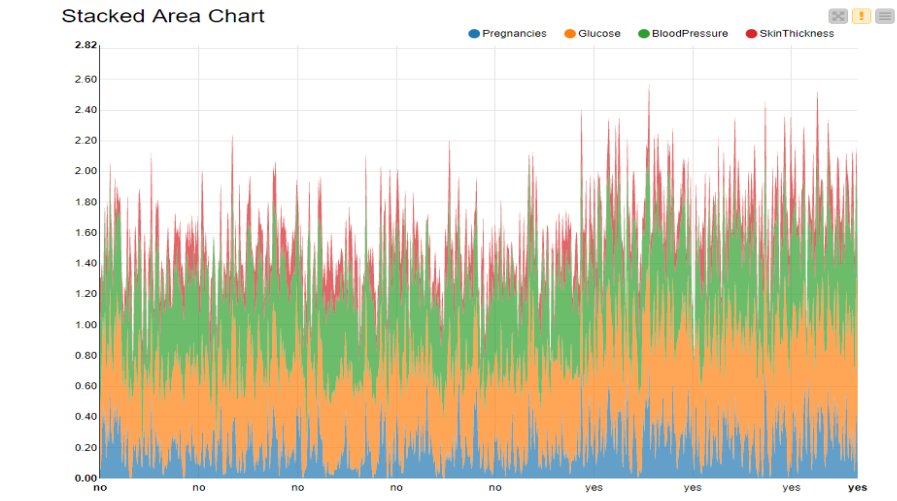


Figure 4.17. Displays the Stacked Area Chart.

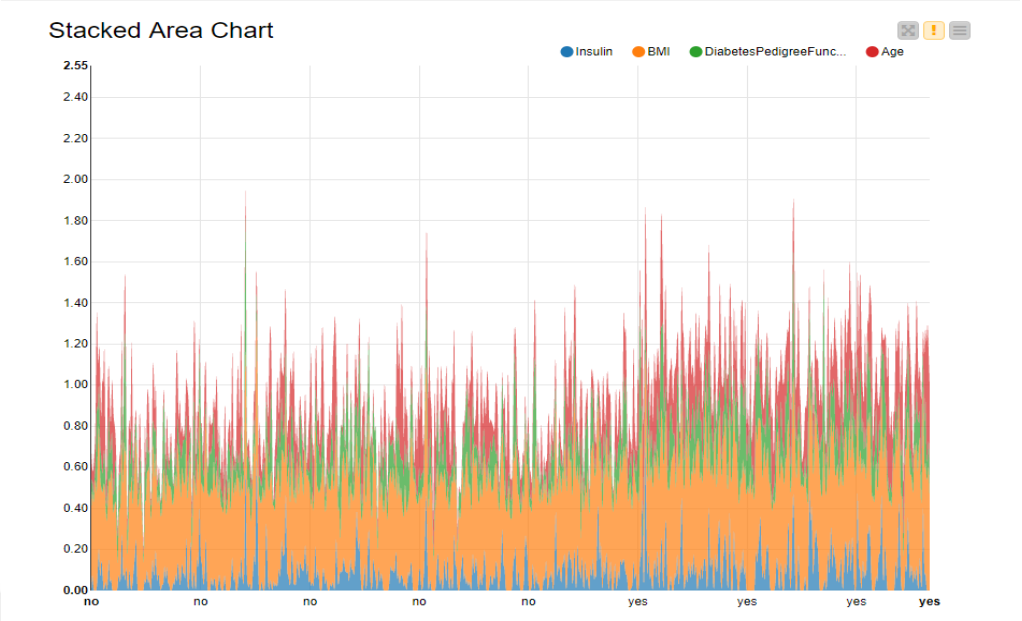


Figure 4.18. Displays the Stacked Area Chart.

4.1.2.4. A Line plot

A Line plot is a style of the graph that displays information along a number line.

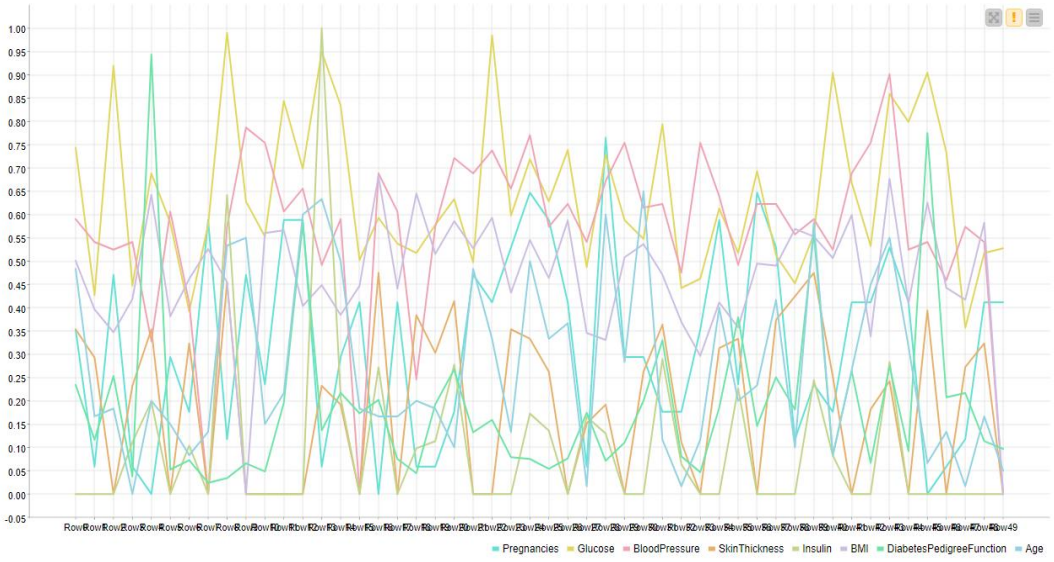


Figure 4.19. Shows the A line plot of the first fifty samples of the Pima data set.

4.1.2.5. Spotting outliers

Outliers are data points whose location is outside the entire pattern in the distribution[56]. We will try to detect outliers in the dataset through Boxplot. Where Boxplot is a way to display the distribution of data based on the following values (“minimum”, first quartile [Q1], median, third quartile [Q3], and “maximum”). Figure 4.20. shows the method for identifying outliers[57].

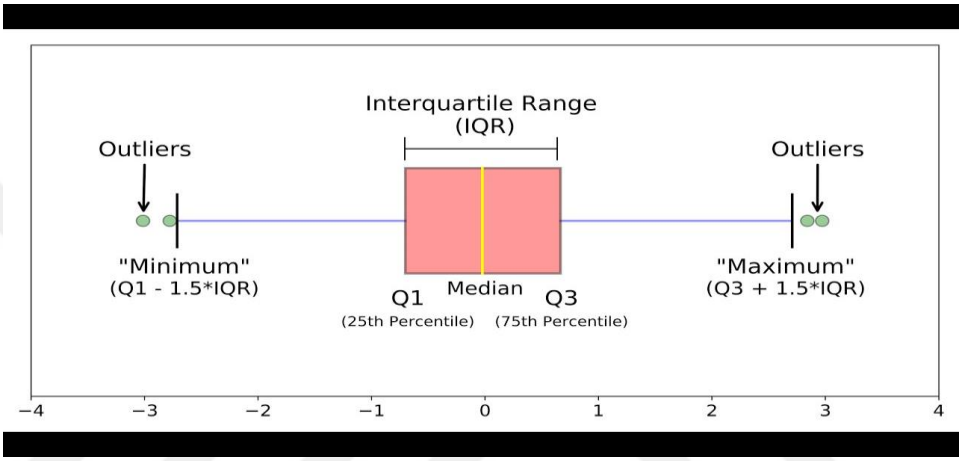


Figure 4.20. Shows the method for identifying outliers.

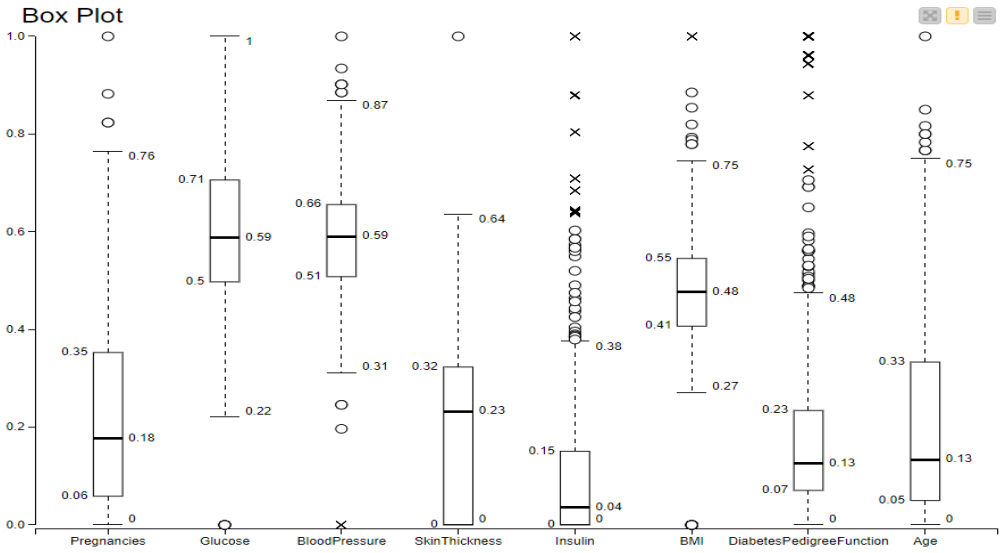


Figure 4.21. Shows the outliers found in the Pima data set

4.1.2.6. Correlation of features

Correlation of features: Correlation indicates the strength of the relationship between two features. Therefore, more than one method will be presented to find out the correlation between the features.

Linear Correlation

Linear Correlation: It measures the amount of correlation between two features as it is assumed that they form a linear relationship. That is, the values of the two features are related in a way that forms a straight line when drawn[58].

Figure 4.22. shows the correlation matrix of features. Where the closer the color is to blue, this indicates the strength of the association, and the closer it is to the red color, this indicates the weakness of the association.

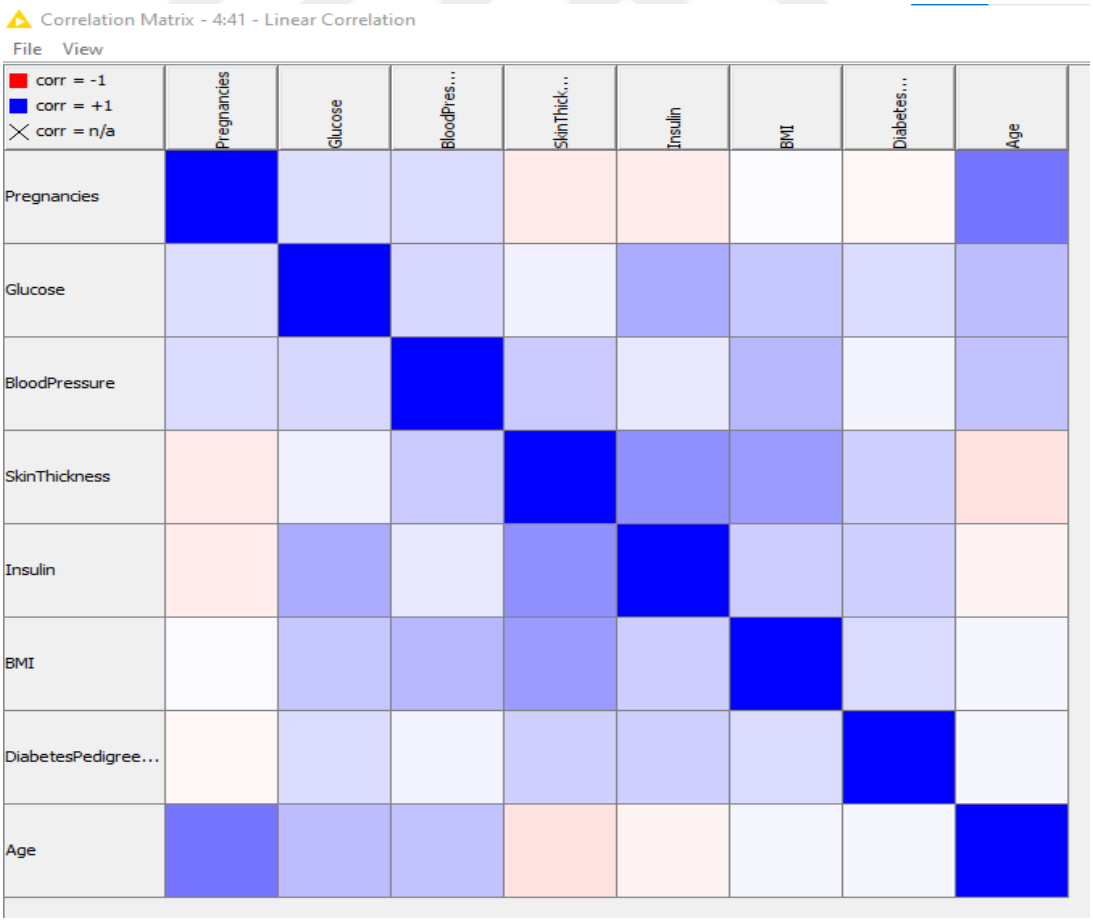


Figure 4.22. Shows the correlation matrix of features.

Figure 4.23. shows the correlation value. Where the closer the correlation value is to one, this indicates the strength of the correlation, and whenever it is closer to zero, this indicates the weakness of the correlation.

Correlation measure - 4:41 - Linear Correlation

File Edit Hilite Navigation View

Table "default" - Rows: 28 Spec - Columns: 5 Properties Flow Variables

Row ID	First column name	Second column name	Correlation value	p value	Degrees of freedom
Row0	Pregnancies	Glucose	0.129458671499276	3.219491352040027E-4	766
Row1	Pregnancies	BloodPressure	0.14128197740713955	8.541845507581414E-5	766
Row2	Pregnancies	SkinThickness	-0.0816717744490084	0.023607948656555973	766
Row3	Pregnancies	Insulin	-0.07353461435162875	0.04162094468635872	766
Row4	Pregnancies	BMI	0.017683090727830333	0.6246376459594261	766
Row5	Pregnancies	DiabetesPedigreeFunction	-0.03352267296261338	0.35353460198769526	766
Row6	Pregnancies	Age	0.5443412284023453	0.0	766
Row7	Glucose	BloodPressure	0.1525895865686655	2.1695071530158927E-5	766
Row8	Glucose	SkinThickness	0.057327890738178455	0.1124141495002311	766
Row9	Glucose	Insulin	0.33135710992021533	0.0	766
Row10	Glucose	BMI	0.2210710694589815	5.891411802849689E-10	766
Row11	Glucose	DiabetesPedigreeFunction	0.13733729982837317	1.345878143714785E-4	766
Row12	Glucose	Age	0.2635143198243391	1.1501910535116622E-13	766
Row13	BloodPressure	SkinThickness	0.2073705384030696	6.606687419363766E-9	766
Row14	BloodPressure	Insulin	0.08893337837319194	0.013683499093533946	766
Row15	BloodPressure	BMI	0.28180528884989975	1.7763568394002505E-15	766
Row16	BloodPressure	DiabetesPedigreeFunction	0.04126494793009806	0.2533743720192341	766
Row17	BloodPressure	Age	0.23952794642136088	1.7520429551609595E-11	766
Row18	SkinThickness	Insulin	0.43678257012001415	0.0	766
Row19	SkinThickness	BMI	0.3925732041590302	0.0	766
Row20	SkinThickness	DiabetesPedigreeFunction	0.18392757295416454	2.856179470711595E-7	766
Row21	SkinThickness	Age	-0.11397026236774259	0.0015582784662138073	766
Row22	Insulin	BMI	0.19785905649309626	3.2196953991814325E-8	766
Row23	Insulin	DiabetesPedigreeFunction	0.18507092916809917	2.402264074330418E-7	766
Row24	Insulin	Age	-0.04216295473537703	0.24318215223454298	766
Row25	BMI	DiabetesPedigreeFunction	0.14064695254510168	9.197970126950672E-5	766
Row26	BMI	Age	0.03624187009229327	0.31583298987244657	766
Row27	DiabetesPedigreeFunction	Age	0.03356131243480567	0.3529797336668983	766

Figure 4.23. Shows the correlation value.

Rank correlation coefficients

Rank correlation coefficients: It measures the extent to which one feature is affected by another feature. Will the value of the first feature increase or decrease if the value of the second feature is increased. The rank correlation coefficients of Spearman and Kendall, for example [59][60].

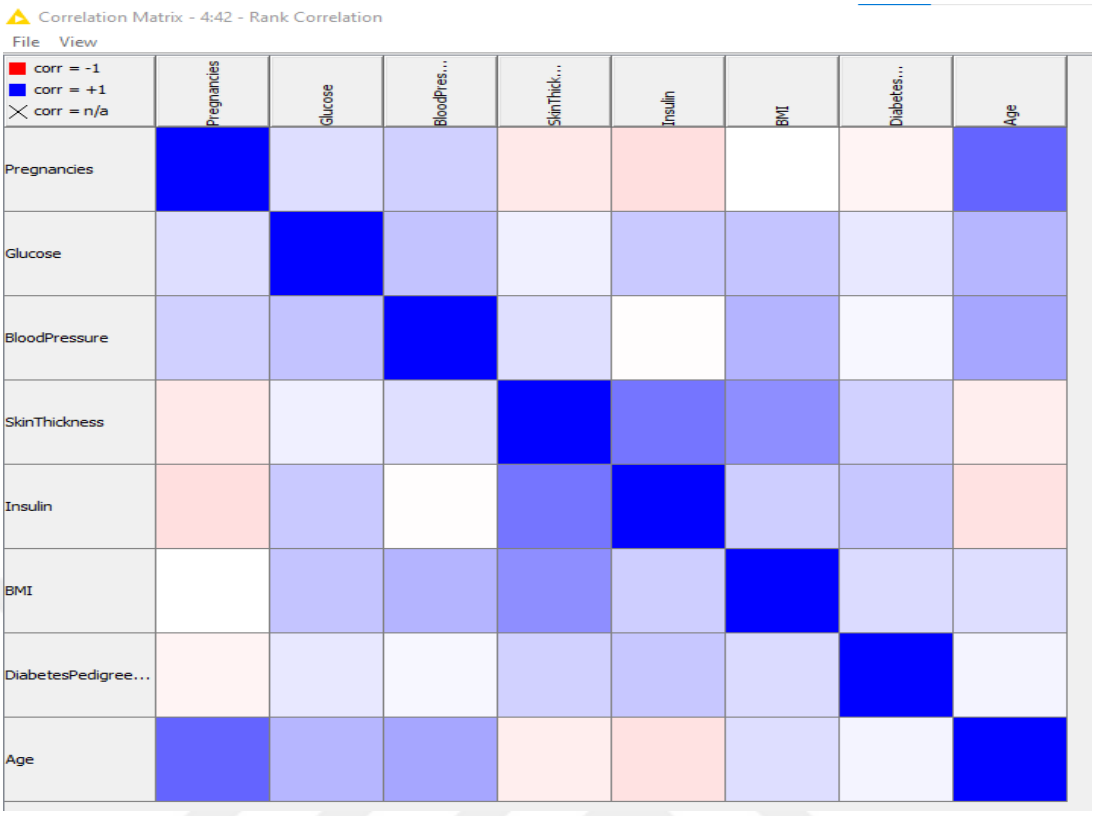


Figure 4..24. Displays the correlation matrix using the Spearman coefficient.

Correlation measure - 4:42 - Rank Correlation

File Edit Hilite Navigation View

Table "default" - Rows: 28 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	First colu...	Second ...	Correlation value	p value	Degrees ...
Row0	Pregnancies	Glucose	0.13073352406886...	2.804774625062567...	766
Row1	Pregnancies	BloodPressure	0.18512673205801...	2.381991746069900...	766
Row2	Pregnancies	SkinThickness	-0.08522230764950...	0.018166757272100...	766
Row3	Pregnancies	Insulin	-0.12672272417606...	4.310196117778275...	766
Row4	Pregnancies	BMI	1.32146869319999...	0.9970827837952267	766
Row5	Pregnancies	DiabetesPedi...	-0.04324150124936...	0.231325407408530...	766
Row6	Pregnancies	Age	0.6072163388236563	0.0	766
Row7	Glucose	BloodPressure	0.2351906134781836	4.112310492132565...	766
Row8	Glucose	SkinThickness	0.06002215292552...	0.096478265223405...	766
Row9	Glucose	Insulin	0.21320580456193...	2.406871812965505...	766
Row10	Glucose	BMI	0.23114119425993...	8.983880306345782...	766
Row11	Glucose	DiabetesPedi...	0.09129336487628...	0.011368518031684...	766
Row12	Glucose	Age	0.2850447199558114	8.881784197001252...	766
Row13	BloodPressure	SkinThickness	0.12648587140494...	4.419264197632611...	766
Row14	BloodPressure	Insulin	-0.00677057181131...	0.8514033781097947	766
Row15	BloodPressure	BMI	0.2928704303204139	2.220446049250313...	766
Row16	BloodPressure	DiabetesPedi...	0.03004633485422...	0.4056912723988493	766
Row17	BloodPressure	Age	0.35089459322163...	0.0	766
Row18	SkinThickness	Insulin	0.5410001366627994	0.0	766
Row19	SkinThickness	BMI	0.44361450828835...	0.0	766
Row20	SkinThickness	DiabetesPedi...	0.18039048342263...	4.845697705313512...	766
Row21	SkinThickness	Age	-0.06679492114381...	0.0642953038243852	766
Row22	Insulin	BMI	0.19272568063619...	7.333072193915768...	766
Row23	Insulin	DiabetesPedi...	0.22115049154027...	5.806688463394494...	766
Row24	Insulin	Age	-0.11421291726443...	0.001522367886771...	766
Row25	BMI	DiabetesPedi...	0.1411920297318744	8.632020536447627...	766
Row26	BMI	Age	0.13118588054276...	2.670039611600838...	766
Row27	DiabetesPedi...	Age	0.04290858770908...	0.2349406482772003	766

Figure 4.25. Shows correlation value using the Spearman coefficient.

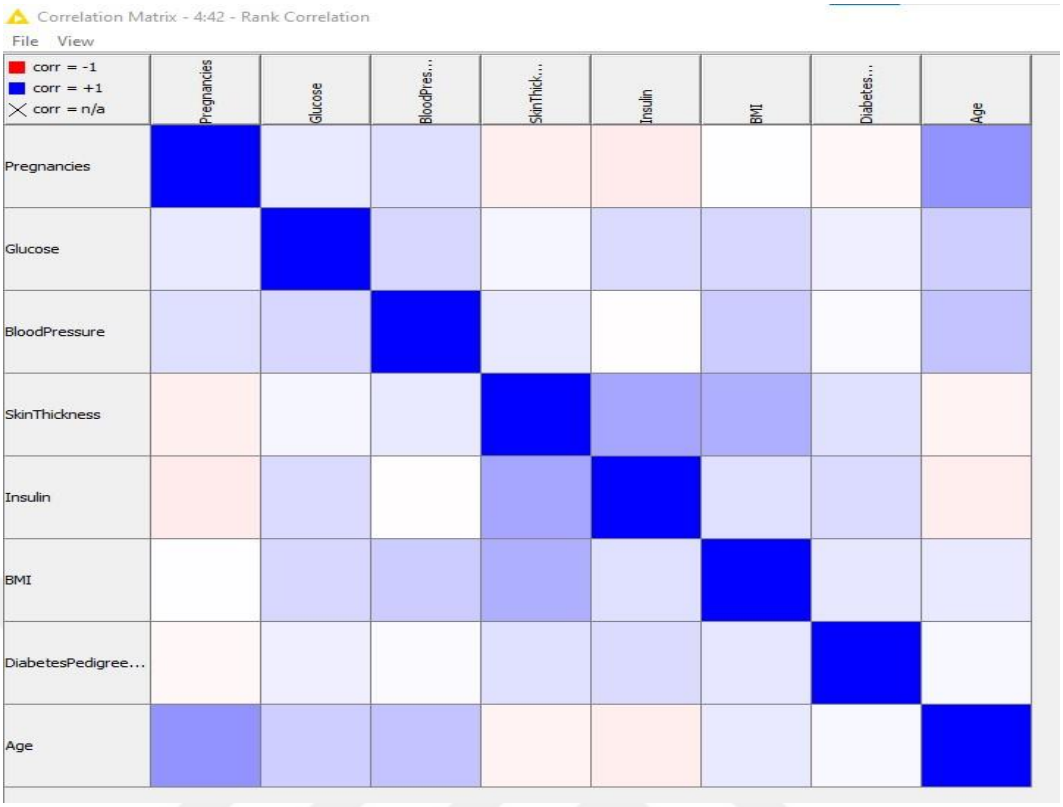


Figure 4.26. Displays the correlation matrix using the Kendall coefficient.

Table "Correlation values" - Rows: 8 | Spec - Columns: 8 | Properties | Flow Variables

Row ID	Pregna...	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigre...	Age
Pregnancies	1.0	0.09121411222802...	0.13476475896...	-0.06373445642306...	-0.09396547676640...	0.004180485240750...	-0.02995511882031...	0.45458143545568...
Glucose	0.09121411...	1.0	0.15988704882...	0.039002974430175...	0.16320119220343093	0.15585201541138863	0.06187049583731413	0.19642321367101...
BloodPressure	0.13476475...	0.15988704882727...	1.0	0.09434569365742798	-0.00363753156793...	0.2051532207297819	0.019447751263359...	0.24556172563481...
SkinThickness	-0.0637344...	0.03900297443017...	0.09434569365...	1.0	0.3794198916081552	0.3313044435201945	0.12642896689561792	-0.04461022323335...
Insulin	-0.0939654...	0.16320119220343...	-0.00363753156...	0.3794198916081552	1.0	0.1413715796115835	0.16157062645295986	-0.079455929702271
BMI	0.00418048...	0.15585201541138...	0.20515322072...	0.3313044435201945	0.1413715796115835	1.0	0.09464334892496935	0.08865750743416...
DiabetesPedi...	-0.0299551...	0.06187049583731...	0.01944775126...	0.12642896689561792	0.16157062645295986	0.09464334892496935	1.0	0.02804101785661...
Age	0.45458143...	0.19642321367101...	0.24556172563...	-0.04461022323335...	-0.079455929702271	0.08865750743416345	0.02804101785661832	1.0

Figure 4.27. Displays the values of the correlation matrix using the Kendall coefficient.

4.1.3. Data pre-processing

Data analysis requires pre-processing of the data. To make data more suited for data mining and to prepare it for future analysis, it is the process of converting or mapping data from a prototype to another format. Data preprocessing includes the following operations:

4.1.3.1. Handling zero values

Handling zero values: After exploring the dataset, it was noticed that it contains some zero values. Zero values were replaced with the mean.

Table 4.3. Shows the number of records with zero values for each feature.

Feature	The number of zero values
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11

4.1.3.2. Normalization and standardization

Data standardization is an important stage during data pre-processing. Features can have different values. Some features may have values between 0 and 100 and others may be between 100 and 1000. Therefore, we do this process, which is standardizing the range of values for the data to do some statistical analyses easier. Also, all features have the same effect.Therefore, in this study, the Normalization or Min-Max Scaling technique was applied to limit the values between 0 and 1, and thus all features had the same effect.

4.1.3.3. Increase in the number of records for diabetic patients

The data set contains outliers, so the outliers have been replaced by the arithmetic mean. Also, most of the records in the data set are for healthy patients, where the number of records of healthy patients constitutes two-thirds of the data set, and therefore the data is considered unbalanced, and relying on it will give incorrect results. Therefore, the number of records for patients with diabetes was increased using the smote technique to make the data set balanced.

Table 4.4. Presents the Pima data set before and after modification.

Dataset	Negative	Positive	The Total
Old	500	268	768
New	423	443	866

4.1.4. Description of the proposed neural network

An input layer, two dense layers, a dropout layer, and an output layer make up the neural network. The dropout layer is set to 0.5. Adam is the learning tool that is employed, while binary_crossentropy is the loss function. Relu is the activation function utilized in the dense layers, while sigmoid is used in the output layer. The number of epochs is set to 100.

Figure 4.28. Description of the neural network

4.1.5. Training phase

After completing the pre-processing of the data and making the data set balanced, the data set was divided into two parts. The sum of the training data that will be used to train the algorithm, the designed model, and the test data set on which the model will

be tested. Six algorithms and the proposed neural network were trained on 80% of the number of records. The model has been tested on 20% of the number of records. The following six algorithms were selected to be trained on the dataset: SVM, Decision Tree, Naive Bayes, Random Forest, KNN, and Gradient Boosted.

Table 4.5. Shows how the data set was divided.

Dataset	Number of records	Percentage
Training	692	80%
Testing	174	20%

4.1.6. Classification stage

After completing the training phase, the six algorithms were tested on the test data set, as the results appeared in the form (SVM:89.08, Decision Tree:90.80, Naïve, Bayes:87.93, Random Forest:95.98, KNN:93.10, and Gradient Boosted: 91.38) After that, the three best algorithms were selected in terms of accuracy, which are (Gradient Boosted, KNN, and Random Forest) and combined them to build the final model. The former model gave good results, as its accuracy reached 96.55%, and this result is considered better than the results of all the algorithms used. Where the patient is classified as He has diabetes if he is evaluated by two or more algorithms as having diabetes. Also, a patient is classified as healthy and non-diabetic if it is classified by two or more algorithms as non-diabetic.

PART 5

RESULTS AND DISCUSSION

In this section, the results of the six algorithms, the results of the proposed model, and the results of the proposed neural network will be presented. The confusion matrix will be presented for each algorithm. Then a comparison is made between the results of these algorithms with the proposed model and with the proposed neural network through a table showing the performance measures (accuracy, recall, accuracy) for each algorithm.

5.1. THE CONFUSION MATRIX OF THE RANDOM FOREST ALGORITHM

The following figure demonstrates the Random Forest algorithm's confusion matrix. Where the Random Forest algorithm was able to correctly predict 84 cases of diabetes, while predicting 3 cases incorrectly, with a rate of 96.55%. And predicting 83 cases of non-diabetes correctly, compared to predicting 4 cases incorrectly, at a rate of 95.40%. The number of correctly classified cases reached 167, compared to 7 cases that were incorrectly classified, with an error rate of 4.02%. This algorithm also gave a good accuracy in general, reaching 95.98%. The accuracy of the Recall scale reached 95.45% for infected cases and 96.51% for healthy cases.

Random Forest

Confusion Matrix

Rows Number : 174	no (Predicted)	yes (Predicted)	
no (Actual)	83	3	96.51%
yes (Actual)	4	84	95.45%
	95.40%	96.55%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
no	83	4	84	3	96.51%	95.40%	96.51%	95.45%	95.95%
yes	84	3	83	4	95.45%	96.55%	95.45%	96.51%	96.00%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
95.98%	4.02%	0.920	167	7

Figure 5.1. Demonstrates the Random Forest algorithm's confusion matrix.

ROC CURVE: is a graph that displays the taxonomic power of a binary classification system by varying the thresholds[61].[62].

Figure 5.2. displays the Random Forest algorithm's ROC CURVE diagram. We can see this through Figure 5.2. the extent to which the features affect the detection of diabetic cases. We note that Glucose has a greater effect than the rest of the features on detecting diabetic cases.

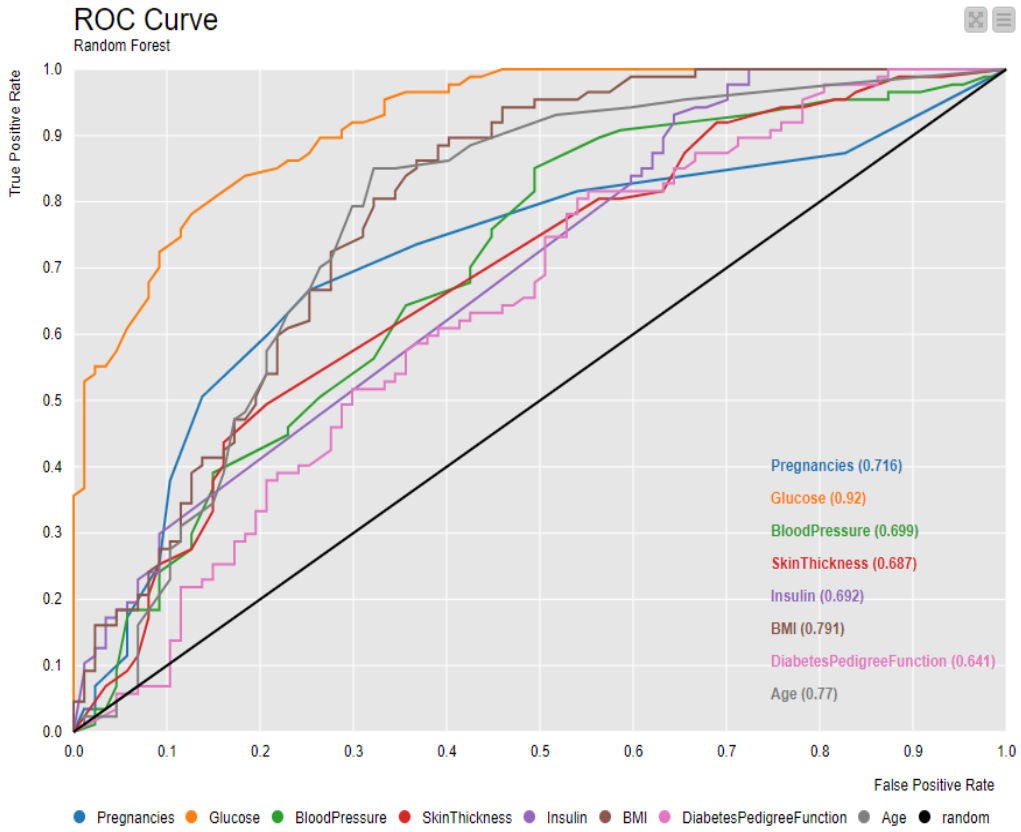


Figure 5.2. Displays the Random Forest algorithm's ROC CURVE diagram.

5.2. THE CONFUSION MATRIX OF THE KNN ALGORITHM

Figure 5.3. Displays the confusion matrix for KNN. The KNN algorithm was able to correctly predict 84 cases of diabetes while predicting 8 cases incorrectly, with a rate of 91.30%. And predicting 78 cases without diabetes correctly, compared to predicting 4 cases incorrectly, at a rate of 95.12%. Where the number of cases classified correctly reached 162, compared to 12 cases that were classified incorrectly, with an error rate of 6.90%. This algorithm also gave a good accuracy in general, reaching 93.10%. The accuracy of the Recall scale reached 95.45% for infected cases and 90.70% for healthy cases.

KNN

Confusion Matrix

Rows Number : 174	no (Predicted)	yes (Predicted)	
no (Actual)	78	8	90.70%
yes (Actual)	4	84	95.45%
	95.12%	91.30%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
no	78	4	84	8	90.70%	95.12%	90.70%	95.45%	92.86%
yes	84	8	78	4	95.45%	91.30%	95.45%	90.70%	93.33%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
93.10%	6.90%	0.862	162	12

Figure 5.3. Displays the confusion matrix for KNN.

Figure 5.4. shows the Roc Curve diagram of the KNN algorithm. We note that glucose, age, and BMI have a greater effect than the rest of the features in identifying diabetic samples. We also note that this diagram gave different results from the Roc Curve diagram of the Random Forest algorithm.

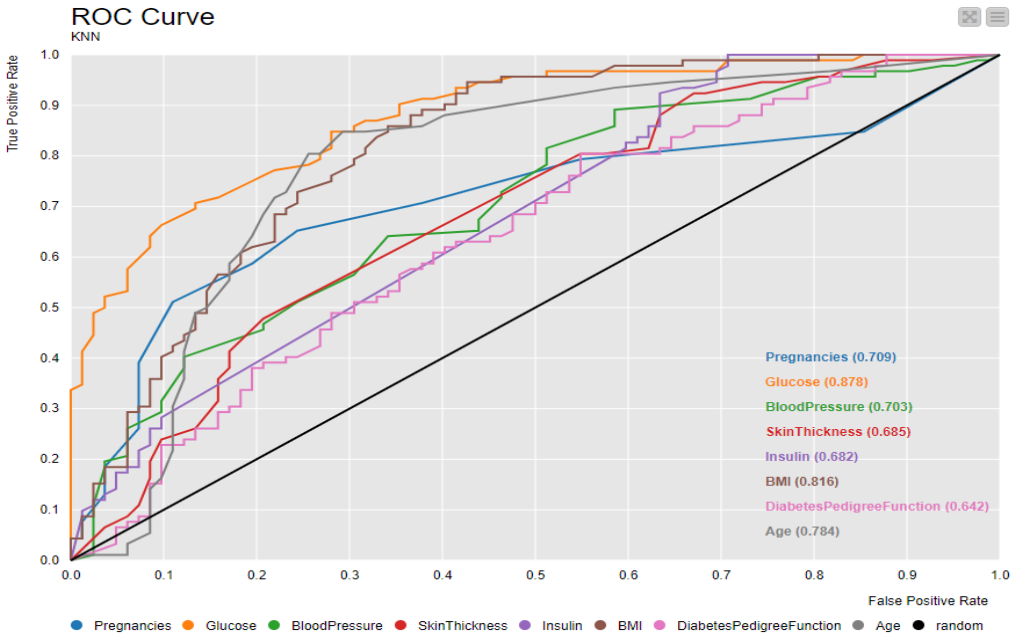


Figure 5.4. Shows the ROC CURVE diagram of the KNN algorithm.

5.3. THE CONFUSION MATRIX OF THE GRADIENT BOOSTED ALGORITHM

Figure 5.5. demonstrates the Gradient Boosted algorithm's confusion matrix. Where the Gradient Boosted algorithm was able to correctly predict 80 cases of diabetes, while predicting 7 cases incorrectly, with a rate of 91.95%. And predicting 79 cases without diabetes correctly, compared to predicting 8 cases incorrectly, at a rate of 90.80%. Where the number of cases classified correctly reached 159, compared to 15 cases that were classified incorrectly, with an error rate of 8.62%. This algorithm also gave a good accuracy in general, reaching 91.38%. The accuracy of the Recall scale reached 90.91% for infected cases and 91.86% for healthy cases.

Gradient Boosted

Confusion Matrix

Rows Number : 174	no (Predicted)	yes (Predicted)	
no (Actual)	79	7	91.86%
yes (Actual)	8	80	90.91%
	90.80%	91.95%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
no	79	8	80	7	91.86%	90.80%	91.86%	90.91%	91.33%
yes	80	7	79	8	90.91%	91.95%	90.91%	91.86%	91.43%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
91.38%	8.62%	0.828	159	15

Figure 5.5. Shows the confusion matrix for the Gradient Boosted algorithm.

Figure 5.6. shows the Roc Curve diagram of the Gradient Boosted algorithm. We note that glucose has a greater effect than the rest of the features in identifying samples with diabetes. We also note that the age feature in this classifier has less impact on identifying infected samples. While the BMI feature had a greater effect in identifying affected samples.

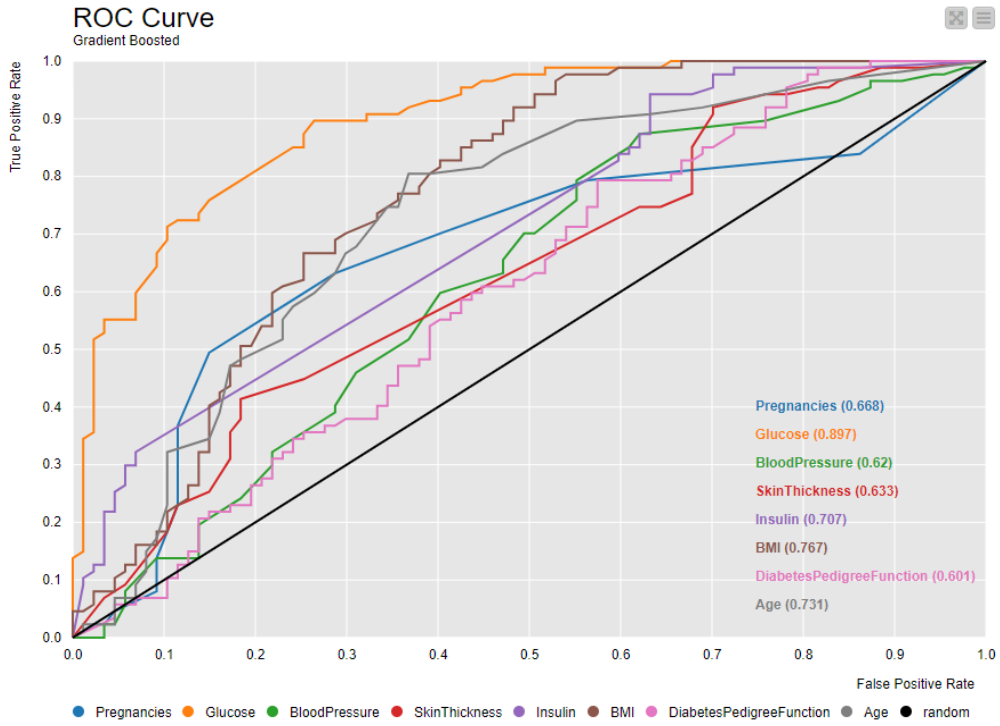


Figure 5.6. Shows the ROC CURVE diagram of the Gradient Boosted algorithm.

5.4. THE CONFUSION MATRIX OF THE NAIVE BAYES ALGORITHM

Figure 5.7. demonstrates the Naive Bayes algorithm's confusion matrix. The Naive Bayes algorithm was able to correctly predict 80 cases of diabetes, compared to 13 cases incorrectly, with a rate of 86.02%. It was able to correctly predict 73 non-diabetic cases, compared to 8 incorrectly predicting cases, with a rate of 90.12%. The number of correctly classified cases reached 153, compared to 21 cases that were misclassified, with an error rate of 12.07%. This algorithm also gave an accuracy of 87.93%. The accuracy of the recall scale was 90.91% for infected cases and 84.88% for healthy cases.

Naive Bayes

Confusion Matrix

Rows Number : 174	no (Predicted)	yes (Predicted)	
no (Actual)	73	13	84.88%
yes (Actual)	8	80	90.91%
	90.12%	86.02%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
no	73	8	80	13	84.88%	90.12%	84.88%	90.91%	87.43%
yes	80	13	73	8	90.91%	86.02%	90.91%	84.88%	88.40%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
87.93%	12.07%	0.758	153	21

Figure 5.7. Demonstrates the Naive Bayes algorithm's confusion matrix.

Figure 5.8. shows the Roc Curve diagram of the Naive Bayes algorithm. We note that glucose has a greater effect than the rest of the features in identifying samples with diabetes. We also note that the age feature and Pregnancies have a clear role in this classifier on detecting infected samples.

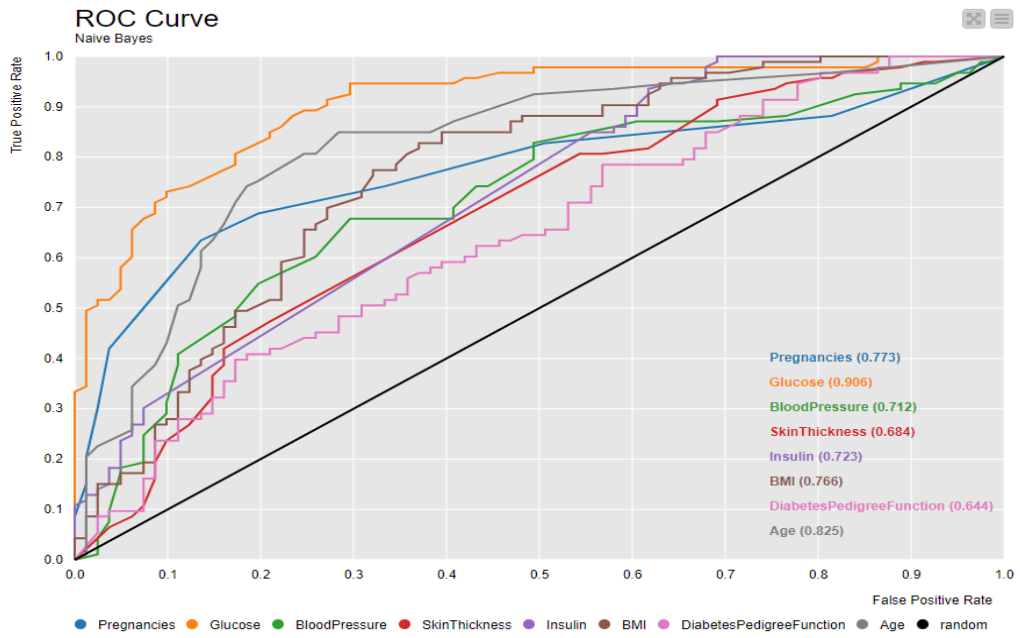


Figure 5.8. Shows the ROC CURVE diagram of the Naive Bayes algorithm.

5.4. THE CONFUSION MATRIX OF THE DECISION TREE ALGORITHM

Figure 5.9. demonstrates the Decision Tree algorithm's confusion matrix. Where the Decision Tree algorithm was able to correctly predict 79 cases of diabetes against 7 cases incorrectly, with a rate of 91.86%. And it was able to correctly predict 79 cases without diabetes against 9 cases incorrectly, with a rate of 89.77%. Where the number of correctly classified cases reached 158, compared to 16 cases that were incorrectly classified, with an error rate of 9.20%. This algorithm also gave an accuracy of 90.80%. The accuracy of the recall scale was 89.77% for infected cases and 91.86% for healthy cases.

Decision Tree

Confusion Matrix

Rows Number : 174	no (Predicted)	yes (Predicted)	
no (Actual)	79	7	91.86%
yes (Actual)	9	79	89.77%
	89.77%	91.86%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
no	79	9	79	7	91.86%	89.77%	91.86%	89.77%	90.80%
yes	79	7	79	9	89.77%	91.86%	89.77%	91.86%	90.80%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
90.80%	9.20%	0.816	158	16

Figure 5.9. Demonstrates the Decision Tree algorithm's confusion matrix.

Figure 5.10. shows the Roc Curve diagram of the Decision Tree algorithm. We note that glucose has a greater effect than the rest of the features in identifying samples with diabetes. We also note that the age feature and Pregnancies have a clear role in this classifier on detecting infected samples.

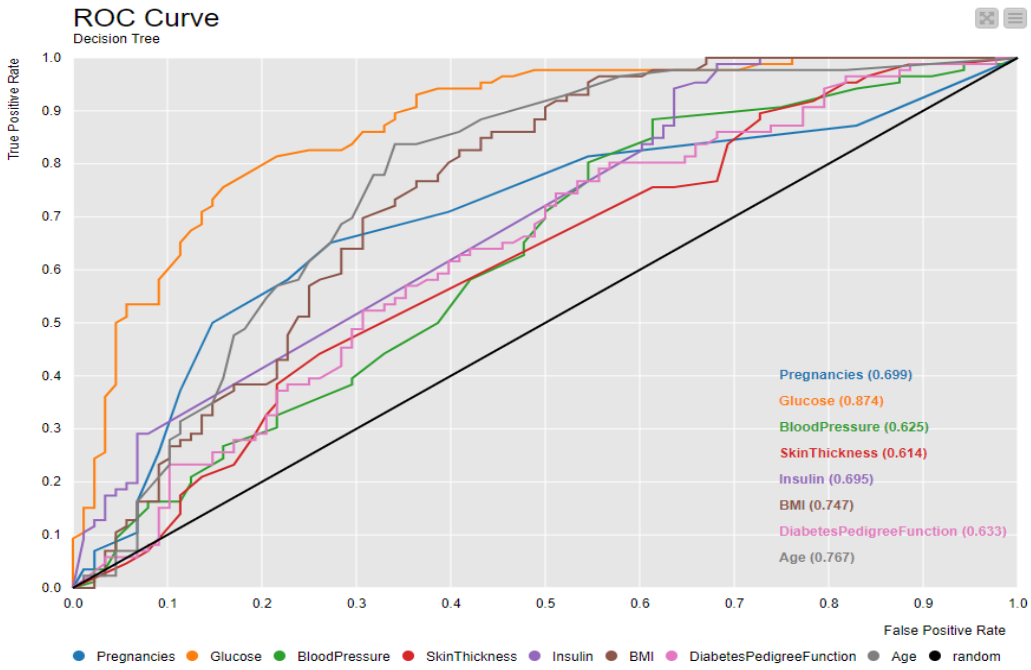


Figure 5.10. Shows the ROC CURVE diagram of the Decision Tree algorithm.

5.5. THE CONFUSION MATRIX OF THE SVM ALGORITHM

Figure 5.11. demonstrates the SVM's confusion matrix. The SVM algorithm was able to correctly predict 79 cases of diabetes against 10 cases incorrectly, with a rate of 88.76%. And it was able to correctly predict 76 cases without diabetes against 9 cases incorrectly, with a rate of 89.41%. The number of correctly classified cases reached 155, compared to 19 cases that were misclassified, with an error rate of 10.92%. This algorithm also gave an accuracy of 89.08%. The accuracy of the recall scale was 89.77% for infected cases and 88.37% for healthy cases.

SVM

Confusion Matrix

Rows Number : 174	no (Predicted)	yes (Predicted)	
no (Actual)	76	10	88.37%
yes (Actual)	9	79	89.77%
	89.41%	88.76%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
no	76	9	79	10	88.37%	89.41%	88.37%	89.77%	88.89%
yes	79	10	76	9	89.77%	88.76%	89.77%	88.37%	89.27%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
89.08%	10.92%	0.782	155	19

Figure 5.11. Demonstrates the SVM's confusion matrix.

Figure 5.12. shows the Roc Curve diagram of the SVM algorithm. We note that the glucose and age BMI feature has a greater effect than the rest of the features in identifying samples with diabetes.

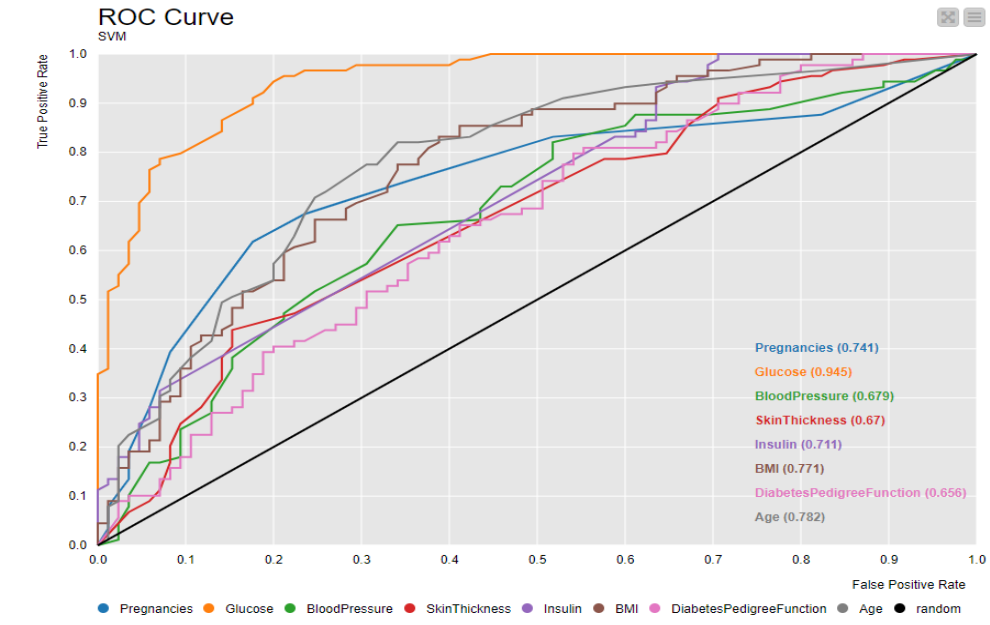


Figure 5.12. Shows the ROC CURVE diagram of the SVM algorithm.

5.6. CONFUSION MATRIX FOR THE PROPOSED MODEL

Figure 5.13. displays the confusion matrix of the proposed model (Hybrid Model). Where the proposed model was able to correctly predict 86 cases of diabetes, compared to predicting 4 cases incorrectly, with a rate of 95.56%. And predicting 82 cases of non-diabetic correctly compared to predicting 2 cases incorrectly at a rate of 97.62%. The number of correctly classified cases reached 168, compared to 6 cases that were incorrectly classified, with an error rate of 3.45%. In a way, the proposed model gave a better accuracy than the rest of the algorithms, as its accuracy reached 96.55%. The accuracy of the Recall scale reached 97.73% for infected cases and 95.53% for healthy cases.

(Hybrid Model)(RAD+KNN+GBM)

Confusion Matrix

Rows Number : 174	no (Predicted)	yes (Predicted)	
no (Actual)	82	4	95.35%
yes (Actual)	2	86	97.73%
	97.62%	95.56%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
no	82	2	86	4	95.35%	97.62%	95.35%	97.73%	96.47%
yes	86	4	82	2	97.73%	95.56%	97.73%	95.35%	96.63%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
96.55%	3.45%	0.931	168	6

Figure 5.13. The Hybrid Model's confusion matrix.

5.7. PROPOSED NEURAL NETWORK RESULTS (DL)

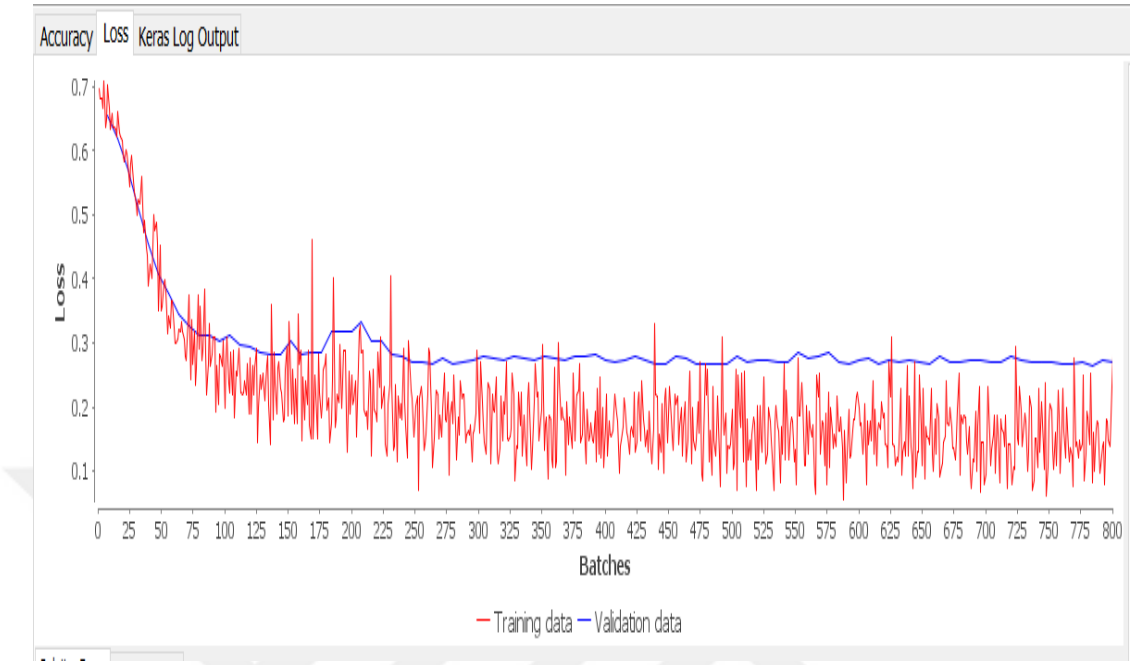


Figure 5.14. Loss in the proposed neural network

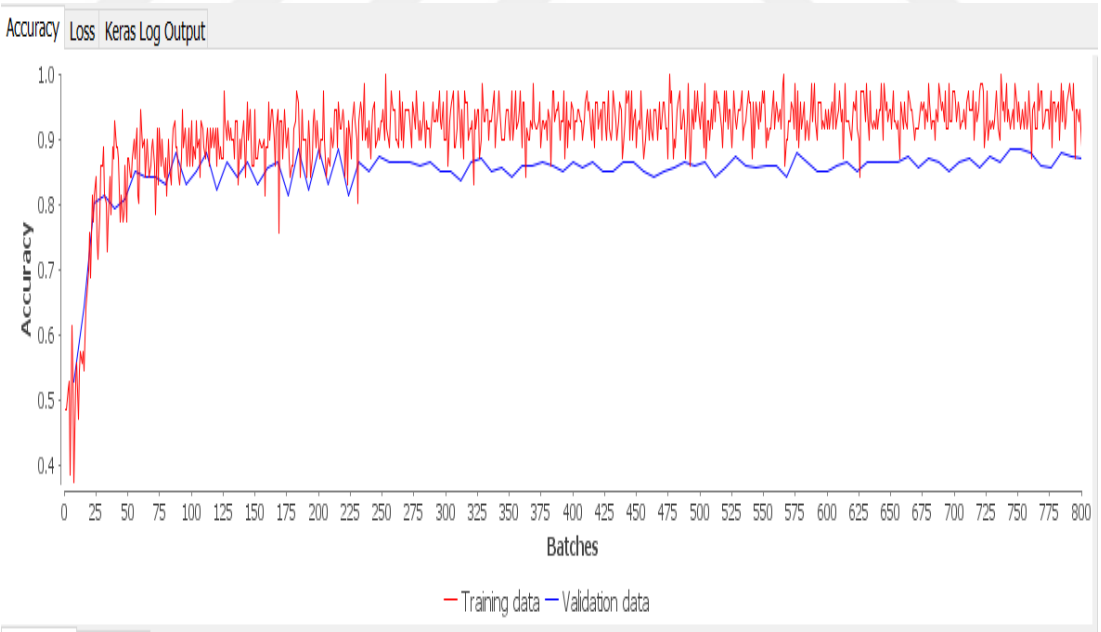


Figure 5.15. The accuracy of the proposed neural network

The following figure displays the confusion matrix of the proposed neural network. We note that the accuracy reached 94.83 percent and that the accuracy of the Recall scale for patients with diabetes reached 94.32 and 95.35 for healthy patients.

Deep Neural Network

Confusion Matrix

Rows Number : 174	no (Predicted)	yes (Predicted)	
no (Actual)	82	4	95.35%
yes (Actual)	5	83	94.32%
	94.25%	95.40%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
no	82	5	83	4	95.35%	94.25%	95.35%	94.32%	94.80%
yes	83	4	82	5	94.32%	95.40%	94.32%	95.35%	94.86%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
94.83%	5.17%	0.897	165	9

Figure 5.16. Confusion matrix for the proposed neural network

5.8. COMPARE THE PROPOSED MODEL WITH MACHINE LEARNING ALGORITHMS

Table 5.1. Presents a comparison of the results of the six algorithms with the proposed model.

The method	accuracy	Recall		Precision	
		Yes	No	Yes	No
SVM	89.08%	89.77%	88.37%	88.76%	89.41%
Decision Tree	90.80%	89.77%	91.86%	91.86%	89.77%
Naive Bayes	87.93%	90.91%	84.88%	86.02%	90.12%
Random Forest	95.98%	95.45%	96.51%	96.55%	95.40%
KNN	93.10%	95.45%	90.70%	91.30%	95.12%
Gradient Boosted	91.38	90.91%	91.86%	91.95%	90.80%
(Hybrid Model) (Ran+KNN+GB)	96.55%	97.73%	95.56%	95.56%	97.62%

Table 5.1. It presents a comparison between the performance of the machine learning algorithms and the performance of the hybrid model. We note by looking at Table 5.1. that the hybrid model gave better results than the results of the machine learning algorithms, with an accuracy of 96.55 percent, and the accuracy of the Recall scale for diabetic patients amounted to 97.73 percent and 95.56 percent for healthy patients, and the error rate of the hybrid model did not exceed 4 percent.

5.9. COMPARE THE PROPOSED MODEL WITH A DEEP NEURAL NETWORK

Table 5.2. Comparison between the proposed model and the proposed neural network

The Method	Accuracy	Recall		Precision	
		Yes	No	Yes	No
Deep Neural Network	94.83%	94.32%	95.35%	95.40%	94.25%
(Hybrid Model) (Ran+KNN+GB)	96.55%	97.73%	95.56%	95.56%	97.62%

Comparing the performance of the deep neural network designed in this study with the proposed hybrid model, we note that the hybrid model gave good results and better than the neural network, as the accuracy of the hybrid model reached 96.55 percent, while the accuracy of the neural network did not exceed 95 percent. The accuracy of the Recall scale for the hybrid model for patients with diabetes reached 97.73 percent, while the accuracy of the Recall scale for the neural network for patients with diabetes did not exceed 95 percent.

5.10. COMPARE THE PROPOSED MODEL WITH PREVIOUS STUDIES

Table 5.3. Compare the proposed model with previous studies

Ref	Year	Algorithm	Accuracy	Recall	Precision
[7]	2022	Random Forest	77%		
[26]	2021	LightGBM + KNN	90.1%	82.1%	88.9%
[10]	2021	Firefly Optimized Neural Network	95.07	88%	88%
[27]	2021	Random Forest	79%	77%	77%
[22]	2021	KNN+SVM+DT	90.62%	91%	91%
[13]	2021	Naïve Bayes	89.9%	84.3%	79.3%
[14]	2021	Gradient Boosting	92%	93%	94%
[28]	2020	K-Means	86%		
[23]	2020	Hybrid classifier	86%		
[18]	2019	Genetic Programming Symbolic Regression	79.19%		
[19]	2019	Artificial Neural Network	82%		
[29]	2019	Decision Tree	70.80%	61.46%	76.5%
[30]	2019	J48	95.122%		
[31]	2018	SVM	79.1%	79.1%	78.2%
[32]	2017	ANN	80.86%		
The proposed hybrid model in this study			96.55%	95.74	95.71%

The last line in Table 5.3. is the results of the hybrid model, while the rest of the lines are the results of some previous studies. In comparison, we find that the proposed

hybrid model in this study gave better results than the results of previous studies, as its accuracy reached 96.55 percent.

5.11. CONCLUSION AND FUTURE WORK

The rapid spread of chronic diseases, especially diabetes, makes it difficult to control by traditional methods. Therefore, doctors and health professionals search for a better solution to detect diseases at an early stage. Modern software systems are considered one of the most important of these solutions. Artificial intelligence, machine learning, and deep learning algorithms have proven their ability to detect diseases at an early stage in easy and inexpensive ways. Many studies were conducted to detect diabetes using machine learning and deep learning algorithms. However, in each study, there was a problem different from the other study, and among these problems, for example, the accuracy did not reach an acceptable result, or the accuracy of detecting diabetic patients was less than the accuracy of detecting patients Healthy patients or that the data was not balanced, so in this study, work was done to build a hybrid model capable of detecting diabetes at an early stage. Work was done to address most of the problems that previous studies suffered from, as the measurement of detecting patients with diabetes gave the best possible accuracy, and its accuracy reached 97.73 %. In the beginning, the data was processed, as there were several database problems, such as zero values and outliers. Then the database was made balanced between patients with diabetes and healthy patients, then 80% of the data was trained on six machine learning algorithms, and after that, it was tested. These algorithms on 20% of the data gave the following results (SVM:89.08, Decision Tree:90.80, Naïve, Bayes:87.93, Random Forest:95.98, KNN:93.10, and Gradient Boosted: 91.38). After that, the three best algorithms were chosen in terms of precision and combined these algorithms to create the final model. This model gave good results, as its accuracy reached 96.55%, and the accuracy of the Recall scale for infected patients reached 97.73%. The results obtained in this study are better than those of previous studies mentioned in this thesis.

In the future, it is possible to experiment with different chronic disease datasets that contain different features, test the proposed model, and compare its results with other models.

REFERENCES

[1] M. Wang *et al.*, "Artificial intelligence models for predicting cardiovascular diseases in people with type 2 diabetes: A systematic review," *Intelligence-Based Medicine*, vol. 6. Elsevier B.V., Jan. 01, 2022. doi: 10.1016/j.ibmed.2022.100072.

[2] *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE.

[3] A. Prabha, J. Yadav, A. Rani, and V. Singh, "Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier," *Comput Biol Med*, vol. 136, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104664.

[4] S. Khan Maliha and M. A. Mahmood, "An Efficient Model for Early Prediction of Diabetes Utilizing Classification Algorithm," in *Proceedings - 2022 6th International Conference on Intelligent Computing and Control Systems, ICICCS 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1607–1611. doi: 10.1109/ICICCS53718.2022.9788441.

[5] A. Grover and A. Joshi, "An overview of chronic disease models: a systematic literature review," *Global journal of health science*, vol. 7, no. 2. pp. 210–227, Mar. 01, 2015. doi: 10.5539/gjhs.v7n2p210.

[6] "DETECTION OF CHRONIC DISEASES USING DEEP VERSUS MACHINE LEARNING TECHNIQUES 2023 MASTER THESIS COMPUTER ENGINEERING."

[7] P. Nagaraj, V. Muneeswaran, A. Dharanidharan, K. Balananthanan, M. Arunkumar, and C. Rajkumar, "A Prediction and Recommendation System for Diabetes Mellitus using XAI-based Lime Explainer," in *International Conference on Sustainable Computing and Data Communication Systems, ICSCDS 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1472–1478. doi: 10.1109/ICSCDS53736.2022.9760847.

[8] H. C. Lagunzad, M. A. C. Impang, M. V. Gonzaga, J. F. Lawan, F. C. Pineda, and R. A. A. Tanjente, "Predicting the Early Sign of Diabetes using ID3 as a Data Model," in *2022 14th International Conference on Computer and Automation Engineering, ICCAE 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 135–139. doi: 10.1109/ICCAE55086.2022.9762442.

[9] S. E. Saeed, B. H. Al Telaq, and A. M. Zeki, “A Comparative Study on Classifying Diabetes Disease using Data Mining Models,” in *2021 International Conference on Data Analytics for Business and Industry, ICDABI 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 438–442. doi: 10.1109/ICDABI53623.2021.9655807.

[10] G. S. Tomar and Institute of Electrical and Electronics Engineers, *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT) proceedings*.

[11] S. Karimah, E. B. Setiawan, and I. Kurniawan, “Implementation of Random Forest in Classification Model of Diabetes Prediction based on Drug Review Content,” in *2021 International Conference on Data Science and Its Applications, ICoDSA 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 228–232. doi: 10.1109/ICoDSA53588.2021.9617218.

[12] P. Amrutha, V. V. Nair, and N. S Nair, “Mellitus Preliminary Analysis using Various Data Mining Algorithms and Metrics,” in *Proceedings of the 6th International Conference on Communication and Electronics Systems, ICCES 2021*, Institute of Electrical and Electronics Engineers Inc., Jul. 2021, pp. 1222–1225. doi: 10.1109/ICCES51350.2021.9489117.

[13] A. Prakash, R. Anand, S. S. Abinayaa, and N. S. Kalyan Chakravarthy, “Normalized Naïve Bayes Model to predict Type-2 Diabetes Mellitus,” in *2021 IEEE International Conference on Emerging Trends in Industry 4.0, ETI 4.0 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ETI4.051663.2021.9619332.

[14] A. A. Khan, H. Qayyum, R. Liaqat, F. Ahmad, A. Nawaz, and B. Younis, “Optimized Prediction Model for Type 2 Diabetes Mellitus Using Gradient Boosting Algorithm,” in *Proceedings of the 2021 Mohammad Ali Jinnah University International Conference on Computing, MAJICC 2021*, Institute of Electrical and Electronics Engineers Inc., Jul. 2021. doi: 10.1109/MAJICC53071.2021.9526257.

[15] M. U. Emon, R. Zannat, T. Khatun, M. Rahman, M. S. Keya, and Ohidujjaman, “Performance Analysis of Diabetic Retinopathy Prediction using Machine Learning Models,” in *Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021*, Institute of Electrical and Electronics Engineers Inc., Jan. 2021, pp. 1048–1052. doi: 10.1109/ICICT50816.2021.9358612.

[16] Z. T. Al-Ars and A. M. Aldabbagh, “Predicting the Early Re-admission of Diabetic Patients Using Different Data Mining Techinques,” in *2021 4th International Conference on Electrical, Computer and Communication Technologies, ICECCT 2021*, Institute

of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICECCT52121.2021.9616746.

[17] Dayananda Sagar Academy of Technology & Management, Institute of Electrical and Electronics Engineers. Bangalore Section., and Institute of Electrical and Electronics Engineers, *First International Conference on Advanced Technologies in Intelligent Control, Environment, Computing and Communication Engineering (ICATIECE-2019) : 19th-20th March, 2019 : Dayananda Sagar Academy of Technology & Management.*

[18] China Research Council of Computer Education in Colleges & Universities, Ontario Tech University, IEEE Education Society, and Institute of Electrical and Electronics Engineers, *The 14th International Conference on Computer Science and Education (ICCSE 2019) : August 19 -21, Toronto, Canada*.

[19] Surya Engineering College and Institute of Electrical and Electronics Engineers, *Proceedings of the 3rd International Conference on Computing Methodologies and Communication (ICCMC 2019) : 27-29, March 2019*.

[20] L. Xu, J. He, and Y. Hu, "Early diabetes risk prediction based on deep learning methods," in *2021 4th International Conference on Pattern Recognition and Artificial Intelligence, PRAI 2021*, Institute of Electrical and Electronics Engineers Inc., Aug. 2021, pp. 282–286. doi: 10.1109/PRAI53619.2021.9551074.

[21] Institute of Electrical and Electronics Engineers and RVS College of Engineering & Technology, *Proceedings of the 2nd International Conference on Inventive Research in Computing Applications (ICIRCA 2020) : 15-17 July, 2020*.

[22] S. Samet, M. R. Laouar, and I. Bendib, "Diabetes mellitus early stage risk prediction using machine learning algorithms," in *5th International Conference on Networking and Advanced Systems, ICNAS 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICNAS53565.2021.9628955.

[23] Amity University, Amity University. Amity Institute of Information Technology, Institute of Electrical and Electronics Engineers. Uttar Pradesh Section, and Institute of Electrical and Electronics Engineers, *ICRITO'2020 : IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) : conference date: 4-5 June 2020 : conference venue: Amity University, Noida, India*.

[24] *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, 2019.

[25] *2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE).* IEEE.

[26] N. Dunbray, R. Rane, S. Nimje, J. Katade, and S. Mavale, "A Novel Prediction Model for Diabetes Detection Using Gridsearch and A Voting Classifier between Lightgbm and KNN," in *2021 2nd Global Conference for Advancement in Technology, GCAT 2021*, Institute of Electrical and Electronics Engineers Inc., Oct. 2021. doi: 10.1109/GCAT52182.2021.9587551.

[27] D. Kaur Bhullar *et al.*, "Developing a Predictive Supervised Machine Learning Models for Diabetes," in *7th International Conference on Computing, Engineering and Design, ICCED 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICCED53389.2021.9664833.

[28] Sri Shakthi Institute of Engineering and Technology, Institute of Electrical and Electronics Engineers. Madras Section, All-India Council for Technical Education, and Institute of Electrical and Electronics Engineers, *2020 International Conference on Computer Communication and Informatics : January 22-24, 2020, Coimbatore, India*.

[29] A. S. Sunge *et al.*, "Prediction Diabetes Mellitus Using Decision Tree Models; Prediction Diabetes Mellitus Using Decision Tree Models," 2019.

[30] Sri Sairam Engineering College. Department of Information Technology and Institute of Electrical and Electronics Engineers, *2019 proceedings of the 3rd International Conference on Computing and Communications Technologies (ICCCT'19) : February 21-22, 2019, Chennai, India.*

[31] *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).*

[32] O. Bayat, S. Aljawarneh, H. F. Carlak, International Association of Researchers, Institute of Electrical and Electronics Engineers, and Akdeniz Üniversitesi, *Proceedings of 2017 International Conference on Engineering & Technology (ICET'2017) : Akdeniz University, Antalya, Turkey, 21-23 August, 2017*.

[33] M. Farag, "Machine Learning."

[34] "3 Types of Machine Learning You Should Know | Coursera." https://www.coursera.org/articles/types-of-machine-learning (accessed Mar. 05, 2023).

[35] "Elements of Machine Learning. (Basic framework for designing… | by pratyush kumar | Medium." https://medium.com/@pratyush057/elements-of-machine-learning-e09ebf16af19 (accessed Mar. 05, 2023).

[36] Q. He *et al.*, "Landslide spatial modelling using novel bivariate statistical based Naïve Bayes, RBF Classifier, and RBF Network machine learning algorithms," *Science of the Total Environment*, vol. 663, pp. 1–15, May 2019, doi: 10.1016/j.scitotenv.2019.01.329.

[37] L. Li *et al.*, "Naive Bayes classifier based on memristor nonlinear conductance," *Microelectronics J*, vol. 129, Nov. 2022, doi: 10.1016/j.mejo.2022.105574.

[38] C. Zhu, C. T. Brown, B. Dadashova, X. Ye, S. Sohrabi, and I. Potts, "Investigation on the driver-victim pairs in pedestrian and bicyclist crashes by latent class clustering and random forest algorithm," *Accid Anal Prev*, vol. 182, Mar. 2023, doi: 10.1016/j.aap.2023.106964.

[39] X. Yuan *et al.*, "Spatiotemporal dynamics and anthropologically dominated drivers of chlorophyll-a, TN and TP concentrations in the Pearl River Estuary based on retrieval algorithm and random forest regression," *Environ Res*, vol. 215, Dec. 2022, doi: 10.1016/j.envres.2022.114380.

[40] "Machine Learning Random Forest Algorithm - Javatpoint." https://www.javatpoint.com/machine-learning-random-forest-algorithm (accessed Mar. 04, 2023).

[41] Q. Li, X. Wang, Q. Pei, X. Chen, and K.-Y. Lam, "Consistency preserving database watermarking algorithm for decision trees," *Digital Communications and Networks*, Jan. 2023, doi: 10.1016/j.dcan.2022.12.015.

[42] E. Laber, L. Murtinho, and F. Oliveira, "Shallow decision trees for explainable k-means clustering," *Pattern Recognit*, vol. 137, May 2023, doi: 10.1016/j.patcog.2022.109239.

[43] "Decision Tree - GeeksforGeeks." https://www.geeksforgeeks.org/decision-tree/ (accessed Mar. 04, 2023).

[44] M. M. dos Santos Freitas *et al.*, "KNN algorithm and multivariate analysis to select and classify starch films," *Food Packag Shelf Life*, vol. 34, Dec. 2022, doi: 10.1016/j.fpsl.2022.100976.

[45] "KNN (K-Nearest Neighbors) #1. How it works? | by Italo José | Towards Data Science." https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d (accessed Mar. 02, 2023).

[46] “A Short Introduction to K-Nearest Neighbors Algorithm | Algorithms, Blockchain and Cloud.” https://helloacm.com/a-short-introduction-to-k-nearest-neighbors-algorithm/ (accessed Mar. 04, 2023).

[47] A. D. Boualem, K. Argoub, A. M. Benkouider, A. Yahiaoui, and K. Toubal, “Viscosity prediction of ionic liquids using NLR and SVM approaches,” *J Mol Liq*, vol. 368, Dec. 2022, doi: 10.1016/j.molliq.2022.120610.

[48] “Support Vector Machines Tutorial - Learn to implement SVM in Python - DataFlair.” https://data-flair.training/blogs/svm-support-vector-machine-tutorial/ (accessed Mar. 04, 2023).

[49] E. E. Başakın, Ö. Ekmekcioğlu, and M. Özger, “Developing a novel approach for missing data imputation of solar radiation: A hybrid differential evolution algorithm based eXtreme gradient boosting model,” *Energy Convers Manag*, vol. 280, Mar. 2023, doi: 10.1016/j.enconman.2023.116780.

[50] “A hands-on explanation of Gradient Boosting Regression | by Vagif Aliyev | Medium.” https://vagifaliyev.medium.com/a-hands-on-explanation-of-gradient-boosting-regression-4cfe7cfdf9e (accessed Mar. 04, 2023).

[51] T. Vaulet *et al.*, “Gradient boosted trees with individual explanations: An alternative to logistic regression for viability prediction in the first trimester of pregnancy,” *Comput Methods Programs Biomed*, vol. 213, Jan. 2022, doi: 10.1016/j.cmpb.2021.106520.

[52] “Linear Regression: The Beginner’s Machine Learning Algorithm.” https://elleknowsmachines.com/linear-regression/ (accessed Mar. 04, 2023).

[53] X.-M. Li *et al.*, “Discrimination of Pb-Zn deposit types using sphalerite geochemistry: New insights from machine learning algorithm,” *Geoscience Frontiers*, vol. 14, no. 4, p. 101580, Jul. 2023, doi: 10.1016/j.gsf.2023.101580.

[54] “Bar chart - Wikipedia.” https://en.wikipedia.org/wiki/Bar_chart (accessed Mar. 05, 2023).

[55] “Scatter Plots | A Complete Guide to Scatter Plots.” https://chartio.com/learn/charts/what-is-a-scatter-plot/ (accessed Mar. 05, 2023).

[56] E. J. Jamshidi, Y. Yusup, J. S. Kayode, and M. A. Kamaruddin, “Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: A case study on surface

water temperature," *Ecol Inform*, vol. 69, Jul. 2022, doi: 10.1016/j.ecoinf.2022.101672.

[57] "Detecting and Treating Outliers | How to Handle Outliers." https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/ (accessed Mar. 05, 2023).

[58] M. P. Mateo and G. Nicolas, "Mapping capability of linear correlation statistics for characterization of complex materials using laser-induced breakdown spectroscopy," *Anal Chim Acta*, vol. 1227, Sep. 2022, doi: 10.1016/j.aca.2022.340260.

[59] L. Zhang and L. Wang, "Optimization of site investigation program for reliability assessment of undrained slope using Spearman rank correlation coefficient," *Comput Geotech*, vol. 155, Mar. 2023, doi: 10.1016/j.compgeo.2022.105208.

[60] A. K. J and S. Abirami, "Aspect-based opinion ranking framework for product reviews using a Spearman's rank correlation coefficient method," *Inf Sci (N Y)*, vol. 460–461, pp. 23–41, Sep. 2018, doi: 10.1016/j.ins.2018.05.003.

[61] A. Prosperi, P. A. Korswagen, M. Korff, R. Schipper, and J. G. Rots, "Empirical fragility and ROC curves for masonry buildings subjected to settlements," *Journal of Building Engineering*, vol. 68, Jun. 2023, doi: 10.1016/j.jobe.2023.106094.

[62] I. Ruisánchez, A. M. Jiménez-Carvelo, and M. P. Callao, "ROC curves for the optimization of one-class model parameters. A case study: Authenticating extra virgin olive oil from a Catalan protected designation of origin," *Talanta*, vol. 222, Jan. 2021, doi: 10.1016/j.talanta.2020.121564.

RESUME

Feras KHALEL graduated from Sultan Pasha Al-Atrash High School in Aleppo in 2010, then began studying for a bachelor's degree at the University of Aleppo, Department of Communications Engineering, and did not complete his education due to the country's circumstances. After that, he moved to Al-Sham University and began studying for a bachelor’s degree in the Department of Computer Engineering, graduating in 2020, then moved to the city of Karabük to start his master’s study in the Department of Computer Engineering.