ABSTRACT

Diabetes is a chronic condition that arises from insufficient or ineffective pancreatic released insulin by the body.Diabetes, particularly if left unchecked, can result in a host of issues that impact nearly all of the body's organ systems. Heart illness, stroke, renal disease, nerve damage (neuropathy), eye impairment or blindness, and foot issues are some of these complications. Early detection lowers the chance of serious complications and enables prompt care.Machine Learning algorithms can be used to analyze large datasets in order to find trends and risk factors and study medical data such as blood tests,genetic information and patient history to detect diabetes at an early stage and the patient can be treated accordingly. Also,by offering tailored, data-driven insights that can result in an earlier diagnosis, better treatment outcomes, and an enhanced quality of life for patients, machine learning (ML) algorithms hold the potential to completely transform the detection and management of diabetes.In order to create an accurate learning model to identify diabetes early on, this study compares the effectiveness of numerous machine learning techniques, including Naive Bayes, Support Vector Machine, Artificial Neural Network, Decision Tree, and Random Forest.Out of all of them, Random Forest did the best, according to the simulation results.

Chapter 1

INTRODUCTION

Diabetes is a chronic condition that occurs due to high blood sugar levels. It happens when the pancreas does not produce sufficient amounts of insulin or when our body cannot effectively utilize the insulin produced. Under-production of insulin affects the regulation of blood glucose and leads to an increase of blood sugar levels or hyperglycemia or diabetes. Diabetes raises the risk for damage to the nerves, blood vessels, heart, kidney and many other vital organs. Identifying diabetes in its initial stages holds the potential to prevent adverse outcomes through timely medical intervention. Predicting if a person is diagnosed with diabetes includes analyzing various factors such as medical history, symptoms, lifestyle and indicative tests. Common symptoms of diabetes are polydipsia, polyuria, sudden weight loss, weakness, visual blurring, delayed healing, obesity and other crucial factors like high blood pressure and high cholesterol levels.

Machine Learning, a great transformative approach to medical sciences, plays a pivotal role in diagnosing diabetes. These algorithms have the ability to study and analyze large and diverse datasets and extract useful information for the early detection and prevention of adverse outcomes. Machine learning models can examine a variety of parameters like a patient's demographics,lifestyle choices, genetic information and patient history. Since diabetes is one of the complex diseases with several subtypes and causes, personalization offered by these models can help improve a patient's response to a particular medicine and help prevent the adverse effects caused by it to a larger extent. Furthermore, machine learning models are particularly good at identifying minor indicators of diabetic problems like retinopathy by evaluating medical pictures like retinal scans. This improves diagnostic precision and makes it possible to take prompt action to stop the disease from getting worse. Thus, machine learning algorithms can efficiently reduce the burden on healthcare by simplifying the prediction of several diseases like diabetes and reducing their outcomes.

The data from the likely-diagnosed people is obtained by blood tests, genetic information, sleep patterns and treatment responses.The next step is cleaning, standardization, encoding and feature selection of the data to form a well-organized dataset suitable for further analysis and that also aids in predicting if the person is diagnosed with diabetes or not. In our dataset, the categorical values were encoded into binary values, the missing values in the dataset were identified and no missing value was found, the outliers were detected using a box plot and no outlier was found. Utilizing machine learning techniques facilitates early detection of diabetes. The dataset is split across training and testing sets. Supervised machine learning algorithms are ideal for prediction of diseases as they are trained using labelled data. Labelling allows the data to learn the patterns and relationships between features and their desired outcome. This learning process builds a mapping function. The algorithms learn to map the combination of features to the corresponding label. Once trained, the model is evaluated on testing data or the data it hasn't seen before during training. This ensures the model isn't simply memorizing the training data and can generalize to unseen examples. The predictions are then compared to the actual labels in the testing data. This allows us to calculate metrics like accuracy or overall correctness of predictions and precision or correctness of positive predictions on unseen data. The algorithm’s performance is assessed using various evaluation metrics like accuracy, precision, recall and F1-score. Evaluation metrics indicate how well the model is performing on a specific data set and help in comparing different models and algorithms. Based on the performance metrics on the testing data, the model could be refined by adjusting hyperparameters or trying a different algorithm. This project aims to compare

the performance of diabetes prediction of several supervised learning techniques – Naive Bayes algorithm, Support Vector Machine algorithm, Artificial Neural Network algorithm, Decision Tree algorithm, Random Forest algorithm. UCI's dataset was utilized to study the significant features necessary for the prediction of diabetes. We used only those features from prior research that showed a significant difference between diabetic and non-diabetic characteristics. The dataset was divided into the training and testing set in the ratio of 70:30 respectively. Five of the aforementioned algorithms were used to train as well as assess the model and the results revealed that random forest performed better than the other techniques. Fig 1.1 shows the workflow of the project

Fig 1.1 Workflow diagram of the project

Chapter 2

LITERATURE REVIEW

2.1 RELATED WORKS

A study published in [1] sought to examine the efficaciousness of various supervised learning approaches—Decision Tree Learning, k-Nearest Neighbor, Naive Bayesian, Artificial Neural Networks, and Support Vector Machines—in the literature with regard to disease gene prediction. Developing an effective supervised learning method for the prediction of novel disease genes was the main goal. It also evaluated Random Forest, a relatively recent ensemble learning technique based on decision trees.The outcomes demonstrated that Random Forest performed better than all other methods and that the results remained constant as the proportion of positive training samples changed. Table 1 provides a summary of the study by outlining the paper's interesting features, thorough implementation, and shortcomings.

Table 1 Overview of the paper

Name of the paper :	**Detailed implementation of the paper:**	**Engaging aspects of the paper:**	**Drawbacks of the paper:**
Comparative Analysis of Machine Learning Techniques Based on Classification for the Prediction of Novel Disease Genes	1) Aim: This study compares the efficaciousness of various supervised learning approaches—including There are several methods for predicting disease genes in the literature, including Decision Tree Learning, k-Nearest Neighbor, Naive Bayesian, Artificial Neural Networks, and Support Vector Machines.	1)This work is the first of its type and analyzes the efficacy of different supervised learning algorithms in disease gene prediction using extensive biomedical annotation data.This research also evaluates a relatively new ensemble learning technique based on decision trees: Random Forest.	1) In biomedical research, it is often the case that observing a link does not always mean that it does not exist, making it practically impossible to build a negative training set, which ideally should consist of actual non-disease genes. Thus, the control set, which included all proteins other than known illness proteins and necessary proteins, was used to randomly pick the negative training set. Both the positive training set and this training set had the same size.

	2) <u>Building the feature set:</u> 2.1) Using a protein interaction network compiled from the Online Predicted Human Interaction Database (OPHID), topological properties of proteins were calculated. 2.2) Protein length and the number of annotated gene ontology items were determined by gathering sequence data and gene ontology terms of proteins from UnitProtKB, respectively. 2.3)Also, protein activities may be more readily distorted by mutation if there are more domains and binding sites. Based on this finding, the number of domains and binding sites for each protein was determined using the BioMart program and the protein database InterPro. 2.4) If a protein has more domains and binding sites, mutations may be able to alter its actions more easily. Based on this discovery, the BioMart tool and the protein database InterPro were used to calculate the number of domains and binding sites for each protein.	2) The techniques tested differ in the training algorithms as well as in the way they construct non-disease gene class.All these methods, however, are predicated on the idea that, in the context of supervised training, illness gene and non-disease gene are distinct entities.	2) Since all the supervised learning techniques discussed in this study treat disease gene prediction as a binary-class classification problem, it is necessary that both positive and negative training datasets should be constructed without any ambiguity. On the other hand, genes that are not recognized to be necessary nor associated with diseases are usually selected for negative training samples. The use of real non-disease genes as the negative training set in machine learning algorithms based on binary-class classifier training for disease gene prediction problems may have this drawback.
	To the training set, ten-fold cross validation was applied	3) In order to extract characteristics and identify genes and	3) To overcome this issue, several other learning methods like

	in order to assess and compare different supervised learning strategies, and the F-measure was utilized to gauge performance.	proteins, interactome, genomic, and proteome data were gathered from a variety of data sources. We used just those features from prior studies that showed a statistically significant distinction between genes associated with disease and those not. Positive and negative samples are typically included in the training set for binary-class classification problems. The known illness genes gathered from OMIM formed the foundation for the positive training samples. All the proteins in the interaction network belonged to the "control set," that is, the non-disease proteins, except for the essential and known illness proteins. Finally, a random selection was made from this "control set" to create the negative training set, which had the same size as the positive training set.	semi-supervised learning and unary class classification techniques are proposed.In semi-supervised learning, the classifier is learned from both labeled (i.e., known disease genes) and unlabeled (i.e., the unknown genes) data.In Unary class classification, the classifier learns only from positive samples.Comparisons and analysis of such methods should also be considered for disease gene prediction problem.
	4) <u>Results:</u> 4.1)Positive training examples were taken	4) This study compares the efficaciousness of	

	from all known disease genes in order to train and evaluate the classifiers. The total of Precision, Recall, and F-measure was determined how well each supervised learning technique performed overall. 4.1) With respect to the precision, ANN achieved the best accuracy (i.e., 75.7%)and k-NN performed worst (i.e., 63.1%) 4.2) With respect to Recall, RF achieved the highest value (i.e., 75.9%) and NB obtained the lowest (i.e., 42.4%). 4.13) DT and RF both have similar high precision scores (69.6% for DT and 68.1% for RF), however recall of RF is much higher(75.9%) than DT. Hence it can be observed that RF has the best F-measure score. 4.2) The positive training set was varied from 10% to 90% of known illness genes because a classifier's performance depends on the size of the training set and the number of known disease genes utilized as positive training examples. When known disease genes changed, the majority of techniques remained steady. When more known disease genes were utilized as	five widely used supervised learning approaches in early experiments. The strategies were used to the problem. The results of the experiments indicate that Random Forest performed best (F-measure: all methods are stable with changes in the size of the positive training set), while NB performed worst.	

	positive training samples, overall performance improved and RF continued to outperform other approaches while NB fared the worst.		

A study in [2] provides a review of machine learning techniques used throughout disease detection and prediction.The research specifically attempted to use machine learning methods like Support Vector Machine (SVM), Decision Tree (DT), K-nearest neighbors (KNN), and Naive Bayes (NB) to address the challenges related to disease classification and precise diagnosis. Table 2 gives a thorough summary of the study, including information on how it was implemented, its interesting features, and any possible downsides.

Table 2 Overview of study

Name of the paper:	**Detailed implementation of the paper:**	**Engaging aspects of the paper:**	**Drawbacks of the paper:**
Using Machine Learning Algorithms for Patient Disease Classification and Prediction	1) Aim: The aim of the paper is to propose and implement a novel machine learning approach to aid in the accurate diagnosis of diseases using real patient data collected from Al-Kadhimiya Teaching Hospital. Specifically, the study uses machine learning techniques to handle the difficulties associated with disease classification and precise diagnosis, such as Support Vector Machine (SVM), Decision Tree (DT), K-nearest neighbors (KNN), and Naive Bayes.	The primary objective of the researchers was the diagnosis of seven diseases in addition to a normal classification: anemia, chronic renal failure, delta hepatitis, high fat content, jaundice, lipid problems, and liver dysfunction.With an accuracy of 84%, the results demonstrated that the SVM and DT approaches performed better than the other algorithms. The researchers emphasized the importance of machine learning techniques in the medical sector for accurate disease diagnosis and early	1) Since the main focus of the work is on employing well-known machine learning techniques (SVM, DT, KNN, and NB) to identify diseases using a dataset from a single hospital, it lacks innovation. It doesn't present any brand-new methods or approaches to the diagnosis of illness.

		detection. Overall, the paper highlighted the potential of machine learning algorithms in improving disease diagnosis and healthcare outcomes.	
	2) The workflow involves the following key steps: 2.1.**Data Collection:** Real data from the Iraqi Kadhimiya Hospital, comprising diagnoses of seven diseases, was collected from 1000 hospitalized patients. 2.2.**Dataset Description:** The dataset included information on the prevalence of each disease, with numbers ranging from 35 to 303 cases. 2.3.**Pre-processing:** Normalization was done as part of the pre-processing of the dataset to put the data inside the [0-1] range. To facilitate model training and evaluation, the dataset was subdivided into training sets, comprising 70%, and testing sets, comprising 30%. 2.4.**Machine Learning Method:** The pre-processed dataset was subjected to a variety of machine learning techniques in order to		2) Although the paper mentions the dataset collected from the Al-Kadhimiya Teaching Hospital, it does not provide sufficient details about the characteristics of the dataset, such as data collection methods, potential biases, missing data handling, or data quality assessment. The effectiveness and dependability of machine learning models can be strongly impacted by these variables.

	diagnose diseases. K-nearest neighbors (KNN), Support Vector Machine (SVM), and Naive Bayes (NB), and Decision Tree (DT). Each algorithm was briefly described, along with its advantages and disadvantages. 2.5 **Evaluation Metric:** The accuracy, recall, and F-score of the suggested model were among the metrics used to assess its performance. These metrics were computed to evaluate the classification models' accuracy for every illness class. 2.6.**Results Analysis:** We showed and examined the outcomes of running each method on the dataset. Along with an overall accuracy for the entire model, each disease class's accuracy, precision, recall, and F-score were reported.		

Another study done by[3] proposed the development of an early diabetes detection model.This is crucial for mitigating the risk of developing complications associated with diabetes, such as renal failure, heart attacks, stroke, blindness, and amputation of the lower extremities. The paper focused on employing an ensemble classification strategy to effectively analyze a massive volume of data and extract usable information for diabetes prediction. Table 3 provides a comprehensive overview of the study, detailing its implementation, engaging aspects, and potential drawbacks.

Table 3 Overview of paper

Name of the paper:	**Detailed implementation of**	**Engaging aspects of**	**Drawbacks of the**

	the paper:	**the paper:**	**paper:**
Identification of Diabetes Through Machine Learning Algorithms	1) Aim: The purpose of this work is to use data mining and machine learning approaches to build a strategy for reliably predicting diabetes at an early stage. This is essential for reducing the chance of developing diabetes-related complications, including heart attacks, strokes, blindness, kidney failure, and lower limb amputations. Using an ensemble classification technique to efficiently evaluate large amounts of data and extract actionable information for diabetes prediction is the main goal of this paper.	A number of machine learning techniques are used, such as Random Forest, Gradient Boosting, XGBoost, Decision Tree, Naive Bayes, K Nearest Neighbors, Logistic Regression, and LightGBM. With an accuracy of 98.1%, the Random Forest algorithm outperforms the others, closely followed by the Decision Tree algorithm at 98%, and the XGB and LGBM algorithms at 97.6%. Various algorithms yielded accuracy values between 77% and 88%. The study finds that, out of all the studied algorithms, Random Forest offers the highest accurate prediction of diabetes.	1) Difficulty in finding datasets beyond the PIMA Indian dataset, which may lack real-time data or have insufficient size. Overfitting in tiny datasets can lead to poor performance on additional testing data.
	2) The work-flow involves the following key steps: 2.1.**Data Collection and Description:** For model training, the authors employed a diabetes dataset with 2000 occurrences and 9 characteristics. 2.2. **Data Preprocessing:** -Missing Values: The authors addressed		2)Some studies overlook essential features or bundle them for convenience, which may lead to suboptimal model performance. Proper feature selection is crucial for optimizing model performance

	missing values using a KNN imputer, which uses the values of the K nearest neighbors to impute missing values. -Managing Imbalance: To handle class imbalance, the authors employed random oversampling, which replicates rows from the minority class to balance the dataset. -Feature Scaling: Feature scaling was performed using Standard Scaler to normalize the range of independent variables. 2.3.**Exploratory Data Analysis (EDA):** EDA was conducted to better understand the dataset. 2.4.**Model Selection and Training:** - Several machine learning techniques, such as K Nearest Neighbors, Logistic Regression, Decision Tree, Naive Bayes, Random Forest, Gradient Boosting, XGBoost, and LightGBM, were chosen by the authors of the research to train the data. - Each algorithm was trained on the preprocessed data. 2.5.**Model**		

	Evaluation: - The accuracy of the trained models was tested using a separate test dataset. - To assess each model's efficacy, performance metrics like recall, accuracy, and precision were evaluated. 2.6.**Comparison and Analysis:** - To find the best algorithm for diabetes prediction, the models were compared using performance criteria. - The strengths and weaknesses of each algorithm were analyzed.		

In summary, these aforementioned studies provide an overview on how to analyze the datasets and compare several machine learning algorithms for disease detection.These studies helped us to achieve our goal to compare the five supervised learning techniques namely: Naive Bayes, To find the optimal learning algorithm, combine Support Vector Machine (SVM), Artificial Neural Networks (ANN), Decision Trees (DT), and Random Forest (RF).

Chapter 3

IMPLEMENTATION

3.1 Methodology

The dataset, which we acquired from the UCI Machine Learning Repository, has 16 features and 521 instances. The dataset was first shown, and then exploratory data analysis (EDA) was carried out to gain a deeper understanding of the information. Subsequently, appropriate steps were implemented to cleanse the data, which included eliminating null, missing, and anomalous values, transforming it into numerical formats, and ultimately detecting and eliminating any existing outliers.30% of the dataset was used for testing, and the remaining 70% was used for training.Subsequently, the effectiveness of various machine learning methods was examined using multiple metrics, including accuracy, precision, recall, F1-score, and confusion matrix.

3.1.1 Dataset attributes:

The dataset used consisted of 521 rows and 17 features.Fig 3.1 shows the description of each feature. The dataset was split into 70% for training and 30% for testing.

Attribute	Description
Age	Age in years
Gender	gender indentity (male or female)
Polyuria	Profuse urination and increased urinary frequency
Polydipsia	Excessive thirst and compulsive drinking of water
Sudden weight loss	Involuntary decrease in body weight
Weakness	Fatigue and decrease in muscle strength
Polyphagia	Extreme hunger and increased appetite
Genital thrush	Yeast infection in genital areas
Visual blurring	Lack of sharpness, clarity of vision
Itching	Irritating sensation that makes a person want to scratch
Irritability	tendency to react to stimuli with the experience of negative affective states
Delayed healing	Impaired healing with failure to progress through normal stages of curing
Partial Paresis	Moderate degree of muscular weakness
Muscle stiffness	Rigid, cramped muscles with convulsive movement
Alopecia	Hair loss from areas of body where hair is usually found
Obesity	Build-up of excessive body fat due to large intake of calories
Class	Positive:Diagnosed Diabetes,Negative: Healthy

Fig 3.1 Description of features present in the dataset

3.1.2 Dataset preprocessing

A)Categorical values

During the EDA, firstly the categorical values present in our dataset in the form of Yes/No were replaced by numerical values in the form of 0/1.This process is known as ordinal encoding.By classifying our data into binary values, we make our data easily interpretable by the different machine learning models thus increasing the overall performance of the model.Fig3.2 shows our dataset after ordinal encoding.

Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
40	1	0	1	0	1	0	0	0	1	0	1	0	1	1	1	1
58	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1
41	1	1	0	0	1	1	0	0	1	0	1	0	1	1	0	1
45	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0	1
60	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
55	1	1	1	0	1	1	0	1	1	0	1	0	1	1	1	1
57	1	1	1	0	1	1	1	0	0	0	1	1	0	0	0	1
66	1	1	1	1	1	0	0	1	1	1	0	1	1	0	0	1
67	1	1	1	0	1	1	1	0	1	1	0	1	1	0	1	1
70	1	0	1	1	1	1	0	1	1	1	0	0	0	1	0	1

Fig 3.2 Dataset after ordinal encoding

B)Missing values

Missing values refers to the situation of lack of observations. Carefully identifying and handling missing values is a crucial step in data preprocessing as it can lead to biased or improper results if not addressed properly. Efficiently removing missing values helps produce more accurate statistical metrics like mean,median and standard deviation which can in turn improve the accuracy of the results obtained.Missing values in the dataset were identified and no missing value was found in our dataset as shown in the figure 3.3.

```
df.isnull().sum()
```

```
Age                 0
Gender              0
Polyuria            0
Polydipsia          0
sudden weight loss  0
weakness            0
Polyphagia          0
Genital thrush      0
visual blurring     0
Itching             0
Irritability        0
delayed healing     0
partial paresis     0
muscle stiffness    0
Alopecia            0
Obesity             0
class               0
dtype: int64
```

Fig 3.3 Number of missing values present in the dataset

C)Removing Outliers:

After the identification of missing values, we tried detecting outliers in the dataset by plotting the box plot for each feature.A box plot or a box-and-whisker plot is a graphical representation of the dataset that is used to indicate the distribution of data along with its central tendency and variability. It is proved to be useful in determining the outliers present and also provides a better understandability of the data.The whiskers in the plot are the lines extending from box to the maximum and minimum values.The data points falling outside these lines are considered to be outliers.They can be seen as as separate data points lying outside the lines.Fig 3.4 shows the detection of outliers in the dataset using box plot. As it is clearly seen in the above figure that there are no potential outliers in most of the cases, we found our dataset to be free from outliers.

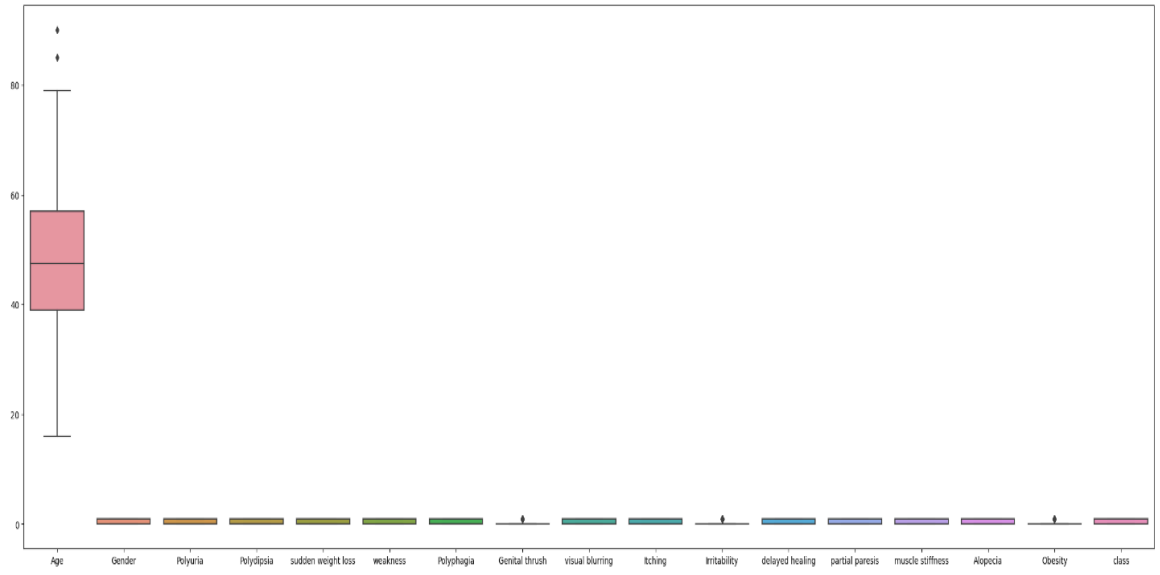

Fig 3.4 Box Plot

3.1.3 Algorithms used

A) Naive Bayes:

The Naive Bayes algorithm is a supervised learning algorithm, meaning that it learns from training data first and uses the Bayes theorem to predict the value of test data. It is applied to the resolution of classification issues. It works well for figuring out how likely it is that an event will occur given past information of comparable events. To put it simply, it categorizes the input according to some prior understanding of its associated features. In particular, it makes the assumption that the characteristics used to categorize the data are unrelated to one another.Since the dataset we used comprised binary values, we employed the Bernoulli Naive Bayes technique.A variation of Naive Bayes called Bernoulli Naive Bayes is very helpful for datasets with values of either 0 or 1.

B) Artificial Neural Network(ANN):

Artificial Neural Networks (ANNs), which are modeled after the neural network in the human brain, are composed of several layers that are added to the computational model in a sequential fashion, with each layer's output acting as the input for the subsequent layer. The layers that lie between the outermost and innermost levels are referred to as "output layers" and "input layers," respectively, while the intermediate layers are called "hidden layers."Depending on the model's

complexity and requirements, there may be one or many hidden layers.Perceptrons are processing elements that mimic the synaptic connections seen in the brain. They may be coupled to all or a subset of the neurodes in the subsequent layers.Weighted signals in the model stimulate the electrical excitation of a nerve cell. The strengthening of the neural pathways is simulated by multiplying these weighted signals with the input signals. Adjusting these weights can help to improve the model's accuracy. With the help of this potent machine learning algorithm, intricate patterns and relationships can be discovered in the data.

3)Support Vector Machine(SVM)

The Support Vector Machine (SVM) is a supervised learning technique that is mainly employed in machine learning for classification tasks.The goal of this approach is to produce the best feasible decision boundary (hyperplane) for classifying the data points so that the test data point is placed in the appropriate category.By calculating the greatest distance between the hyperplane and the closest data point from each class, this hyperplane optimizes the margin between the classes. It determines the hyperplane in accordance with the critical points (support vectors) that are selected first.

4)Decision Tree

A decision tree is a tree-based method where each leaf node represents the result, internal nodes indicate the dataset's features, and branches represent the decision rules.In order to divide the data, it first chooses the best feature according to various standards, such as information gain for classification issues or the Gini index. Subsets of the data are then created according to the selected feature values.Until the maximum depth of samples is achieved or there is no more progress in impurity or error reduction, this process is repeated recursively.

5)Random Forest

As the name suggests, Random Forest is also a tree based learning algorithm but instead of relying on one decision tree, it contains a number of decision trees based on the subsets of the sample dataset and finds out the average in order to improve the accuracy of the model. This algorithm proves to be better than Decision tree algorithm as the larger the number of trees in the forest, the higher is the accuracy rate of the model.

3.2 Testing Plan

Table 4 Metrics for Testing Data

S.no	**Algorithms Used**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
1	Naive Bayes	87.17%	93.61%	86.27%	89.79%
2	Support Vector Machine(SVM)	93.58%	95.09%	95.09%	93.58%
3	Artificial Neural Networks(ANN)	96.15%	96.15%	98.03%	97.08%
4	Decision Tree	97.43%	98.03%	98.03%	98.03%
5	Random Forest	99.35%	100%	99.01%	99.5%

3.3 Result Analysis

The model's accuracy, precision, recall, and F1-score were the metrics used to evaluate it.The number of accurately predicted observations divided by the total number of predicted observations is the measure of accuracy. It merely offers a general indicator of the model's performance.Over all the cases that are anticipated to be positive, precision finds the number of instances that are actually positive.The ratio of accurately anticipated positive values to all predicted positive values is used to determine it.A low false positive rate is indicative of high precision.The percentage of all positive cases that the model successfully detects is measured by recall.The ratio of genuine positives to all true positives + false negatives—true values that the model ignores—is used to compute it.F1-Score is calculated as the harmonic mean of precision and recall. It takes the average of two values but gives more weightage to the lower value. This means that if the model is performing well with one of the metrics and poor with another then that model will have a lower F1-score. Fig 3.5 shows the comparison for five supervised learning techniques for Diabetes detection.With respect to precision, Random forest performed the best(100%) and Naive Bayes performed the worst(93.6%).When considering Recall as well, Random forest performed the best(99.01%) and Naive bayes performed the worst(89.79%).F1-Score as described earlier, is the harmonic mean of Precision and recall and whenever either of them is small, the F-measure also gives smaller value. By looking at the figure, we can clearly see that in the case of recall, decision trees and Random forest have similar values(98.03%) for DT and 99.01% for RF),suggesting that both are almost equally sensitive in retrieving the true diabetic features. However RF outperformed for precision score(100%) indicating that it can correctly predict all the diseased patients. This resulted in Random forest having the best F1-Score(99.5%).A confusion matrix is a table that is used to examine how well a model performs with a given set of test data. Four elements make up the 2X2 matrix: true positive, true negative, false positive, and false negative predictions.The number of instances in which the model accurately anticipated the true values is known as a true positive (TP), and the number of correctly predicted false values is known as a true negative (TN).False negatives (FN) are the negative values that the model mistakenly predicts, and false positives (FP) are the situations in which the model incorrectly predicts true values. The confusion matrix's layout is displayed in Table 5.Fig 3.6 shows the comparison of the five supervised learning techniques with the help of confusion matrices for each of them.It is clearly shown in the figure that ANN,Decision Trees and Random Forest has the highest number of True positives indicating they can detect the diseased patients correctly and also they have the least Number of False Positives suggesting a patient is rarely assumed to be healthy in these cases. Thus confusion matrix analysis suggests that ANN,Decision trees and Random Forest algorithms performed the best among all.

Table 5 Layout of Confusion Matrix

	Predicted Negative(0)	Predicted Positive(1)
Actual Negative(0)	True Negatives(TN)	False Positives(FP)
Actual positive(1)	False Negatives(FN)	True Positives(TP)

Fig 3.5 Performance Classification of five supervised learning techniques

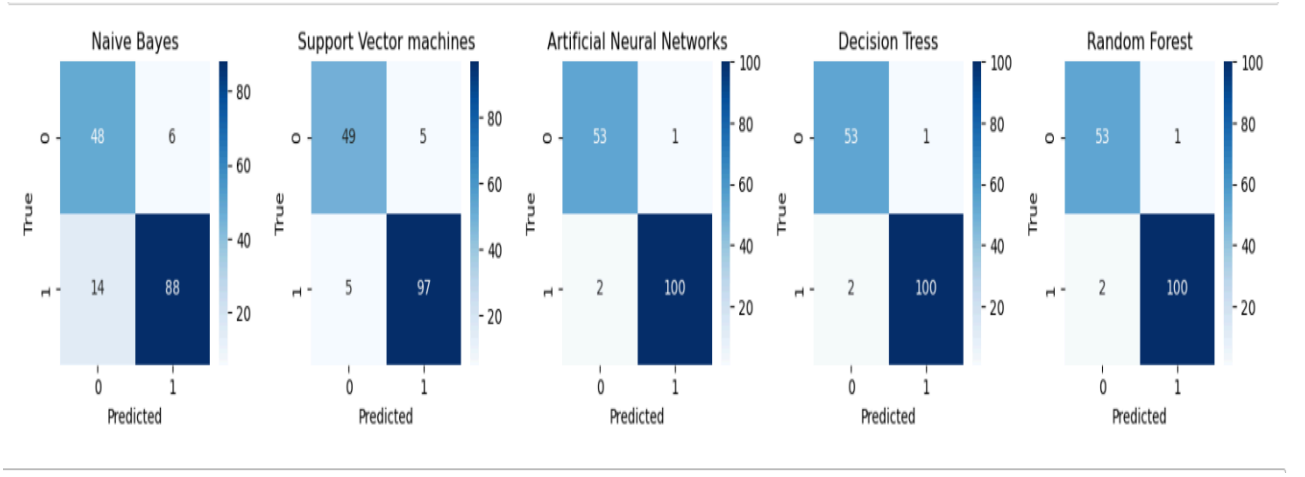

Fig 3.6 Confusion matrix analysis for different supervised learning techniques

Chapter 4

CONCLUSION AND FUTURE SCOPE

This work uses data from the UCI Machine Learning Repository, which includes 521 cases and 17 characteristics, to assess different machine learning methods for early diabetes detection. Naive Bayes, Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees, and Random Forests are among the algorithms that have been tested. With an F1 score of 99.5%, 99.01% recall, and 99.35% precision, Random Forests scored better than the others. This suggests that they are better at predicting diabetes while reducing false positives. Confusion matrices were utilized to shed light on model performance and demonstrate how well ANN, decision trees, and random forests identify patients. These results improve the accuracy of disease diagnosis by furthering machine learning-based diagnostics.

With a precision of 99.35% and a recall of 99.01%, the Random Forest algorithm showed remarkable performance in the diabetes test, attaining perfect scores in both precision and recall at 100%. It also achieved the maximum F1 score of 99.5%, demonstrating its exceptional capacity to precisely identify diabetes cases while reducing false positives. Confusion matrices evaluation offered additional insights into model performance, emphasizing the efficiency of random forests, decision trees, and artificial neural networks (ANNs) in accurately identifying patients while reducing false negatives. These findings demonstrate the potential of supervised learning to improve illness detection accuracy and substantially advance machine learning-based diagnostics.

Machine learning algorithms are able to provide rapid, personalized, and proactive solutions for early diabetes detection thanks to real-time datasets.By examining trends or patterns in the patients' health condition, they can be valuable in supplying the most recent and up-to-date information that can be helpful in identifying the most important elements for diabetes prediction. In the future, this research will use more exact and accurate real-time datasets to test other machine learning techniques, including logistic regression and KNN.