ML In Predicting Diabetes In The Early Stage

Niveditha S[1], Jyothi M [2], Keerthana R[3], Priyanka H M[4], Amrutha P[5]

[1] Head of Department and Assistant Professor, Computer Science and Engineering Department, Jnanavikas Institute of Technology,Karnataka, India
[2]Undergraduate Student, Computer Science and Engineering Department, Jnanavikas Institute of Technology, Karnataka, India
[3]Undergraduate Student Computer Science and Engineering Department, Jnanavikas Institute of Technology, Karnataka, India
[4]Undergraduate Student, Computer Science and Engineering Department, Jnanavikas Institute of Technology, Karnataka, India
[5]Undergraduate Student, Computer Science and Engineering Department, Jnanavikas Institute of Technology, Karnataka, India

---

Abstract-*Diabetes is a chronic ailment that could result in a catastrophe for the world's healthcare system 382 million people worldwide have diabetes, according to the International Diabetes Federation. By 2035, this will increase to 592 million. Diabetes is a disease characterized by high blood glucose levels. The signs of this raised blood sugar level include increased thirst, appetite, and frequency of urinating. One of the main causes of stroke, kidney failure, heart failure, amputations, blindness, and kidney failure is diabetes. Our bodies convert the food we eat into sugars like glucose when we eat. Then, we anticipate insulin to be released from our pancreas. Our cells can be unlocked by insulin, allowing glucose to enter and empowering us. The most prevalent forms of the disease are type 1 and type 2, but there are other varieties as well, including gestational diabetes, which develops during pregnancy. Data science's newest field, machine learning, studies how computers learn via experience. The objective of this work is to develop a system that can more correctly conduct early diabetes prediction for a patient by combining the results of several machine learning methodologies.*

Keywords- Decision tree, K closest neighbor, Logistic Regression, Support vector machine, Accuracy, Machine Learning, Diabetes.

INTRODUCTION

Diabetes is a disorder that is spreading swiftly, even in children If we are to understand diabetes and how it develops, we must first understand what happens in the body when there is no diabetes. We get sugar (glucose) from the foods we eat, particularly those that are heavy in carbohydrates. Our body's primary source of energy comes from foods high in carbohydrates. Everyone requires carbohydrates, including those who have diabetes. Examples of foods that contain carbs include bread, cereal, pasta, rice, fruit, dairy products, and vegetables (especially starchy vegetables). Some glucose is transported to our brain in order for us to think and function effectively. The remaining additionally to our liver, where it is stored as energy that the body uses later. Insulin is required for the body to burn glucose for fuel. The pancreas' beta cells create insulin. Insulin acts as a key to open door to allow glucose to enter the cell from the bloodstream, insulin binds to the cell's doors and opens them. Hyperglycemia occurs when glucose builds up in the bloodstream, and diabetes happens when the pancreas is unable to create enough insulin (insulin deficit) or the body is unable to utilize the insulin produced (insulin resistance). Diabetes Mellitus is characterized by elevated blood sugar (glucose) levels. Diabetes Subtypes:

A person with type 1 diabetes has a weaker immune system and cells that are unable to produce adequate insulin. There are currently no reliable studies demonstrating the causes of type 1 diabetes, and no successful preventative measures are available.

Type 2 diabetes is distinguished by either insufficient insulin synthesis by the cells or by the body's incorrect insulin use. This kind of diabetes affects 90% of people with diabetes, making it the most common. Its occurrence is influenced by both hereditary and environmental factors. Diabetes Causes : Diabetes is primarily caused by genetics. It is caused by at least two faulty genes on chromosome 6, The chromosome that controls how the body responds to
different antigens.

Potentially, viral infection could have an impact on how type 1 and type 2 diabetes develop. Viruses such as hepatitis B, CMV, mumps, rubella, and coxsackievirus, according to research, enhance the likelihood of developing diabetes.

Women who experience high blood sugar levels early in pregnancy develop gestational diabetes. It will reoccur in two-thirds of the cases throughout additional pregnancies. When gestational diabetes was present throughout a pregnancy, there is a high likelihood that type 1 or type 2 diabetes will emerge.

LITERATURE SURVEY

1. Yasodha et al:

Categorization on a number of datasets is used to assess whether or not a person has diabetes. The hospital's data warehouse, which has 200 instances with nine attributes, was used to create the data set for the diabetic patient. Both blood tests and urine tests are mentioned in these instances of the dataset .WEKA can be utilized to classify the data in this study's implementation due to its great performance on tiny datasets. The 10-fold cross validation approach is used to review and compare the data after that. We use the naïve Bayes, J48, REP Tree, and Random Tree algorithms. Among the others, J48 performed the best, with an accuracy of 60.2%.

2. Aiswarya et al:

Attempts to diagnose diabetes by studying and evaluating the patterns that emerge in data through classification analysis using Decision Tree and Naive Bayes algorithms. The study's purpose is to propose a faster and more successful method of identifying the sickness, which will aid in the timely treatment of the patients. The study discovered that utilizing a 70:30 split, the PIMA dataset, and cross validation, the J48 technique delivers an accuracy rate of 74.8% and the naïve Bayes method produces an accuracy rate of 79.5%. dataset with dichotomous values, which means that the class variable has two possible outcomes and may be easily handled if detected earlier in the data preprocessing stage and can aid in improving the prediction model's performance.

3. Gupta et al:

The study evaluates how well the same classifiers perform when used with different platforms, similar parameters (accuracy, sensitivity, and specificity), like MATLAB and RapidMiner. It also seeks to discover and compute the accuracy, sensitivity, and specificity percentages of various categorization algorithms. The algorithms JRIP, J Graft, and Bayes Net were utilized. J Graft has the highest accuracy (81.3%), sensitivity (59.7%), and specificity (81.4%), according to the data Furthermore, it was discovered that WEKA outperforms MATLAB and RapidMiner. 2021 Innovations in power and advanced computing technologies (i-PACT), Rahul S G, Rajnikant Kushwaha, Sayantan Bhattacharjee, Agniv Aditya, K Somasekhar Reddy, Durri Shahwar. computing technologies (i-PACT).

4. Lee et al:

Concentrate on utilizing the decision tree algorithm CART on the diabetes dataset after applying the resample filter to the data. The author emphasizes the issue of class imbalance and the importance of addressing it before using any approach to boost accuracy rates. The bulk of class imbalances occur in datasets with dichotomous values, indicating that the class variable has two possible outcomes. If this imbalance is detected earlier in the data preprocessing step, it may be easily addressed and will help to improve the prediction model's accuracy.

METHODOLGY

The diabetes data set was developed. We will learn about the different classifiers used in machine learning to predict diabetes in this part. We will also present the technique we proposed to increase precision. In this paper, five alternative methodologies were used. The various strategies are discussed more below. The accuracy metrics of machine learning models are the output. Predictions can then be made using the model. The collection contains 2000 diabetes cases. The readings are used to establish whether or not the patient has diabetes.

Traditional and cutting-edge machine learning approaches are employed to forecast diabetes in its early stages in the work presented. The same dataset is used to run eight well-known machine learning algorithms, and the outcomes are analyzed using the same metrics to discover the most practicable technique that provides the greatest classification performance result. Decision tree, random forest, support vector machine, XG Boost, K-nearest neighbor, Naive Bayes, artificial neural network, and convolutional neural network are the machine learning methods employed.

ARCHITECTURE DIAGRAM

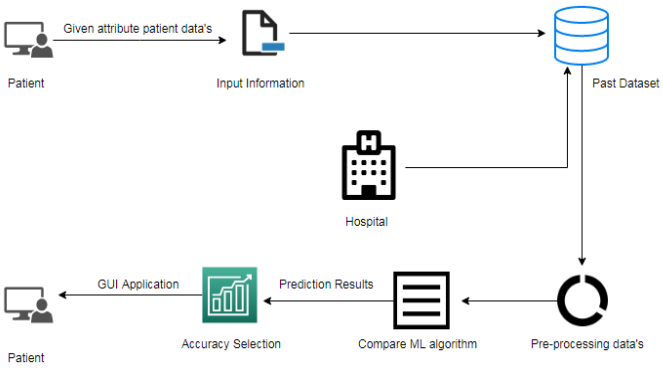


Fig-1 Software Architecture

One of the most pressing real-world medical challenges is the early detection of diabetes. In this work, concerted efforts are made to develop a system that can predict diabetes. Our pancreas is then supposed for insulin production. The ability of insulin to unlock our cells is like a key, enabling the internal flow of glucose and our ability to work. The most common types of diabetes are type 1 and type 2, although there are others, such as gestational diabetes, which develops during pregnancy. Machine learning is a new area in data science that explores how machines learn from experience. The purpose of this work is to develop a system that can more correctly predict early diabetes in a patient by combining the findings of multiple machine learning methodologies.

METHODS AND MATERIALS

1. Dataset

The dataset for this study was supplied by the Diabetes and Digestive and Kidney Diseases National Institute and is At the UCI ML Repository, it is accessible to everyone [29]. The main objective of using this information was to identify and determine a patient's likelihood of having diabetes using precise diagnostic information from the dataset. When selecting occurrences from the larger dataset, numerous constraints were encountered. Both the dataset and the problem are supervised binary classification, specifically. Diabetic Pima Indians (PID) dataset had 768 records describing female patients, 500 negative instances (65.1%), and 268 positive instances (34.9%), as well as 9 = 8 + 1 (Class Attribute) attributes.

2. Data preprocessing

Real-world data may have values that are noisy, inconsistent, or missing. If Low data quality may lead to ineffective search results. The data must be preprocessed in order to obtain high-quality findings. Cleaning, integration, transformation, reduction, and discretization are used to preprocess the data. It is vital to increase the data's suitability for data mining and analysis in terms of time, cost, and quality [30].

3. Data cleaning

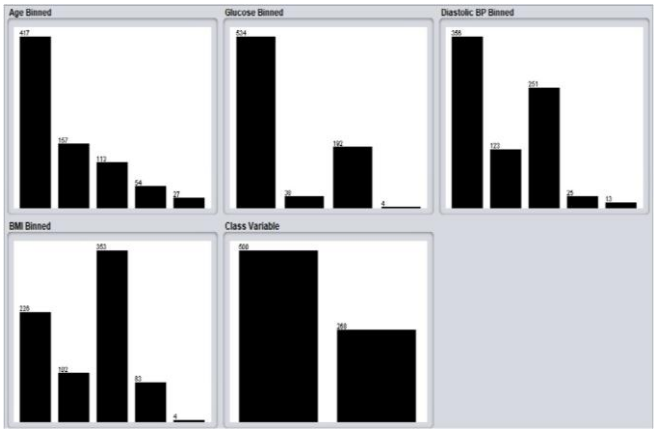
Real-world data may have irregular, inconsistent, or missing numbers. If the data quality is poor, it is probable that no useful results will be discovered. Preprocessing the data is required to get high-quality results. Data preprocessing techniques include cleaning, integration, transformation, reduction, and discretization. It is critical to improve data mining and analysis applicability in terms of time, cost, and quality [30].

4. Data compression

A smaller-volume, condensed version of the dataset that yields the same results is produced through data reduction. (Or a similar) outcome. A dataset's number of attributes has been reduced via dimensional reduction [32]. The principle component analysis method was used to identify essential properties from a large dataset. Age, diastolic blood pressure, BMI, and glucose were all factors in the dataset that were statistically significant.

5. Transformation of data

Smoothing, normalization , and data aggregation are all components of data transformation [33]. The binning method was employed to smooth the data. The property of age has proved effective in categorizing into five groups.



Age(Years)	Age Bins
≤30	Youngest
31–40	Younger
41–50	Middle aged
51–60	Older
≥61	Oldest

Binning of glucose.

Glucose	Glucose Bins
≤60	Very Low
61–80	Low
81–140	Normal
141–180	Early Diabetes
≥181	Diabetes

ALGORITHMS

1. **XG Boost:**

A tree-based machine learning system called XG Boost Extreme Gradient Boosting begins with weak models and finishes with a powerful model. More nodes are added to decision trees in parallel while accounting for the gradient of the loss function. The results of each tree are examined when categorizing an instance, and the result with the most votes is returned as the model's output.

2. K- Nearest Neighbor

A straightforward but efficient machine learning algorithm is K's closest neighbor. A graph representing training data with the expectation that examples from the same classes will be clustered together. An instance's position on a graph is determined using its features when predicting it is given a label, and the k neighbors who are closest to it are recognized. For the categorization of the diabetic dataset, a two- three-dimensional convolutional neural network with convolutional layers was used., with the labels of these neighbors taken into account. Convolution layers employ 8 and 4 filters.

3. The decision tree

The Decision trees, Using a machine learning tool, show how the generated model predicts data. It builds a tree with nodes representing features the branches represent the paths that must be taken following each node, while the leaves represent forecasts. Classes of the given data can be anticipated by travelling from the root to the leaves and picking appropriate branches. The most significant and elective characteristic is located at the root node of the decision tree, indicating the significance of features. In the study being presented, models were built using Gini Information Gain settings for two-level pruning.

4.Random Forest

The decision tree makes misleading predictions when a section of it is constructed wrongly. A machine learning method called random forest seeks to address the overfitting problem. This method combines the predictions of numerous decision trees that were generated at random, and the label that received the most votes is returned as the label for the input data. The voting mechanism for many trees' judgement.

5. SVM: Support Vector Machine

A Each instance in a space is mapped by a support vector machine, which also divides the space into hyperplanes. Each hyperplane represents a class, and each piece of data is mapped to form the classification .Because training costs and time for large datasets may be prohibitive, it is preferable to use SVM for small datasets. The polynomial kernel used in this study has a degree of 3, and the regularization value is set to 0.1.

6.Gaussian Naïve Bayes

The Naive Bayes machine learning algorithm is built on the Bayes theorem. When the dataset is large and contains a large number of features, it does not produce good results since it assumes that all attributes are independent. Gaussian A variation of Naive Bayes that uses the Gauss normal distribution is called Naive Bayes. polynomial kernel used in this study has a degree of 3, and the regularization value is set to 0.1.

7. A neural network is a type of artificial neural network.

A neural network is a type of artificial neural network intended to solve difficult issues by simulating the functions of the human brain. To create predictions, a network with nodes and connections is built using this procedure. At initialization, each link will be assigned random weights, which will be adjusted based on the loss of train data. n nodes are used to produce predictions for an n-class problem. connected to the network's output. And the outputs of each of those n nodes provide the likelihood that a specific set of data belongs to a particular class.

8. Convolutional Neural Network

Artificial neural networks are created to solve complex issues by emulating how the human brain functions. Using this method, a network with Predictions is made using nodes and links. Upon initialization, random weights will be assigned to each link, and weights will be adjusted based on how much train data was lost. The likelihood that a given collection of data belongs When n nodes are set to the network's outputs, each of the n nodes' outputs corresponds to a specific class. Utilize the results to predict a problem with n classes. Performance metrics include accuracy, recall, precision, and f-score.

An accuracy performance indicator measures the proportion of correctly identified data to total data. Despite its popularity, it does not provide full Statistics performed by the model. Precision is defined as the ratio of genuine positives to all data classified as positive. Precision in diabetes patient classification exhibits the model's ability to identify patients while avoiding categorizing healthy individuals as sufferers. The proportion of true positive outcomes to all positive results is referred to as recall. In the case of diabetes classification, it reflects how many patients the model can identify. F score is a valid metric for evaluating model performance since it calculates the harmonic mean of recall and precision.

Accuracy

$$(TP+TN+FP+FN)/(TP+TN+FP+FN) \quad (1)$$

$$\text{Precision} = (TP+FP)/(TP+FP) \quad (2)$$

$$\text{Recall} = (TP+FN)/(TP+FN) \quad (3)$$

$$2*(P*R)/(P+R)\ \text{F-Score} \quad (4)$$

A confusion matrix may also be used to display model performance. An n x n matrix known as a confusion matrix is one in which n represents the quantity of labels in a specific dataset genuine labels are represented by each row, while predicted labels are represented by each column.

9.Aritifical neural network

An important data mining technique is the artificial neural network (ANN), a branch of artificial intelligence research. The An ANN's three layers are input, hidden, and output. Units in the hidden layer convert the input layer into the output layer.. The output of one neuron is used as the input of another layer. An artificial neural network (ANN) recognizes complicated patterns and learns from them. There are billions of neurons in the human brain. A perceptron is a single neuron of this type, and axons connect these cells to other cells. Dendrites interpret input as stimuli after receiving it. Dendrites take in information and convert it into stimuli.

Similar to this, the ANN is made up of numerous nodes connected to one another. A weight represents the connection between two units. An ANN's purpose is the transformation of input into useful output. Introducing "input" refers to the combination of a set of input values linked to a weight vector, which might be positive or negative. The weights can be added using a function, such as y = w1 1x + w x2 2 to send the result to the output. The weighting determines a unit's influence, and the synapse is where a neuron's input signal meets another neuron's output signal. Both supervised and unsupervised learning methods are compatible with ANN. Our study utilized supervised learning because the results are provided.

RESULTS AND DISCUSSIONS

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Fig-2 Standard Form of Confusion Matrix

In order to examine the approaches throughout the dataset, these measures were created using 5fold cross checking. When using k fold cross validation, The dataset has been split into k parts.

The training procedure will be repeated k times, using k-1 folds for training and 1fold for testing each time. This strategy avoids the problem of unbalanced data, allowing model metrics to be monitored more precisely. And 1fold for testing. With this approach, the problem of unbalanced data is avoided, allowing for more precise monitoring of model metrics.

		Predicted	
		Positive	Negative
Actual	True	True Positive 132	False Negative 136
	False	False Positive 59	True Negative 441

(A) Random Forest

		Predicted	
		Positive	Negative
Actual	True	True Positive 176	False Negative 92
	False	False Positive 95	True Negative 405

(B) Artificial Neural Network

		Predicted	
		Positive	Negative
Actual	True	True Positive 181	False Negative 87
	False	False Positive 116	True Negative 384

(C) Clustering

Fig-3 Confusion matrix of proposed models

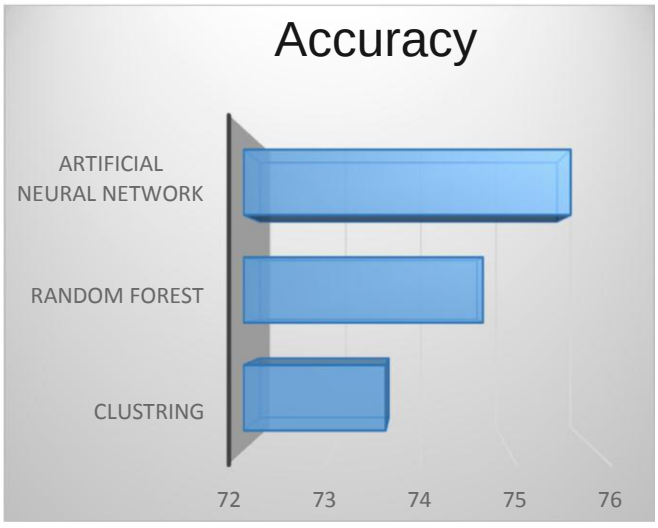
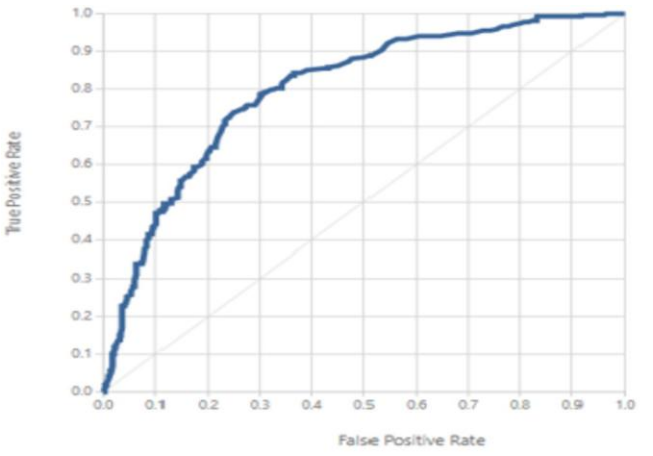
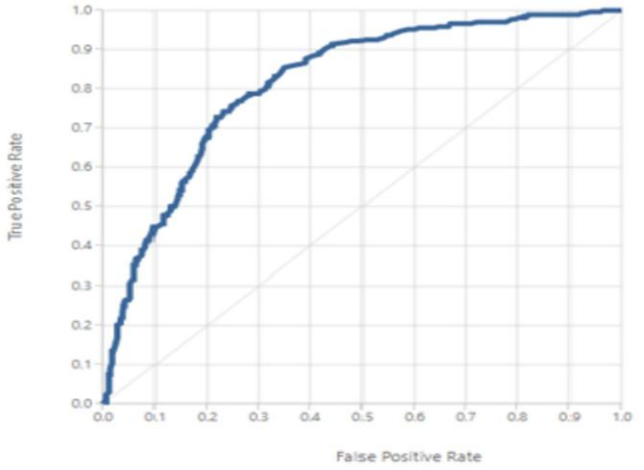


Fig-4 Accuracy



(A) Random forest: Sensitivity= 0.74, Specificity=0.31



(B) ANN: Sensitivity= 0.75, Specificity=0.29

Fig-5 Sensitivity

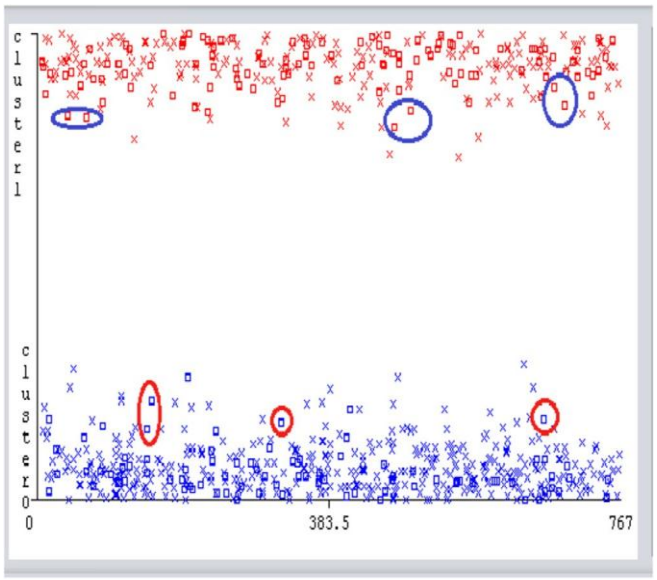


Fig-6 Correct and incorrect clustered instances.

Training Precision	0.81
Testing Precision	0.78

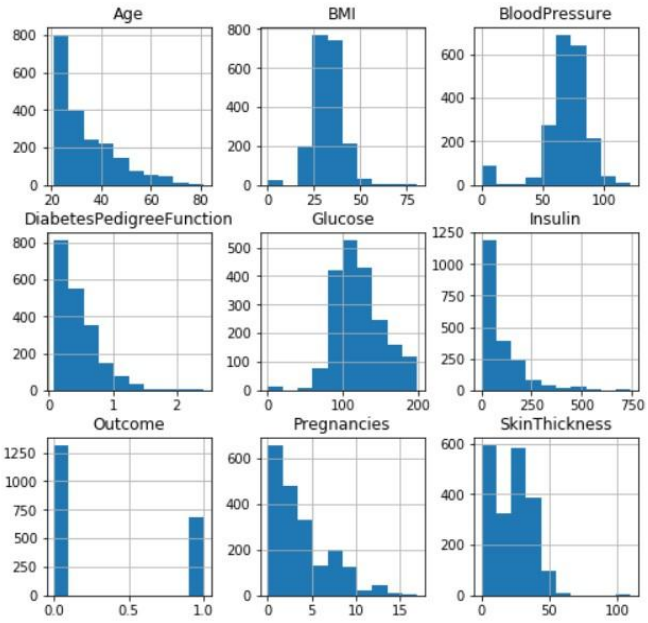


Fig-7 The distribution of features and labels varies.,

A confusion matrix can also be used to demonstrate model performance. An n x n matrix is a confusion matrix. in which n represents how many labels are there in a particular dataset. The labels themselves are represented by each row, while the anticipated labels are shown by each column.

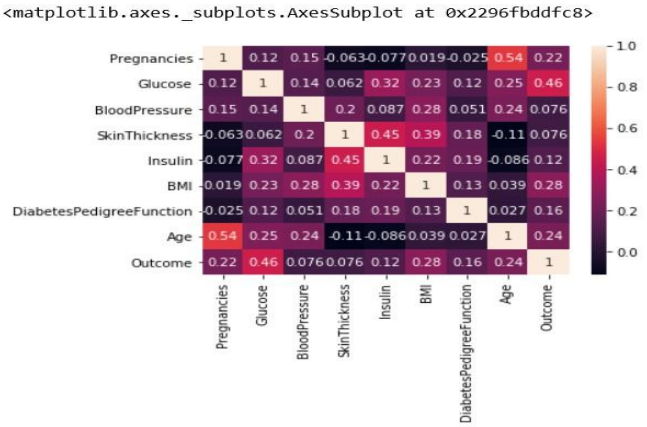


Fig-8 Correlation Matrix

It is clear that none of the attributes have a particularly strong relationship with our result value. Some characteristics have a positive correlation with the outcome value, while others have a negative correlation. Histogram:

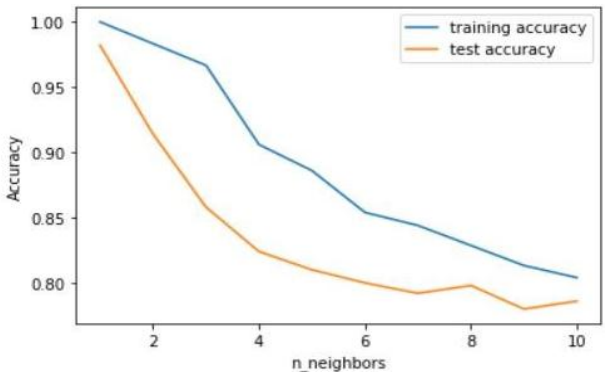


Fig-9 Connection between model complexity and accuracy

It is clear that none of the features significantly influence our result value. Some features and the outcome value have a positive correlation, while others have a negative correlation. Histogram:

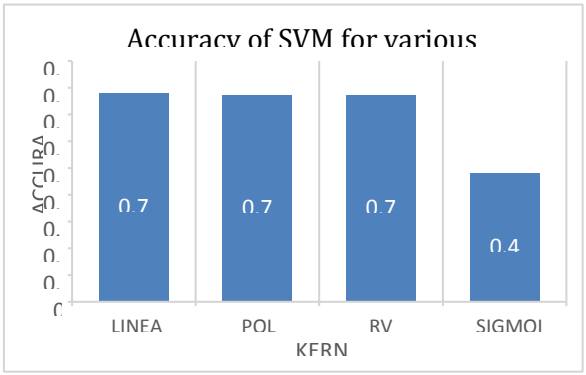


Fig-10 Final Accuracy

This classifier attempts to generate a hyper plane that modifies the distance between the data points and the hyper plane to best distinguish the classes. A number of kernels are used to select the hyperplane. I tried out the linear, poly, rbf, and sigmoid kernels.

CONCLUSION

The early diagnosis of diabetes is one of the most urgent modern medical problems. The goal of this endeavor is to create a system that can anticipate diabetes. This research examines and evaluates five machine learning classification methods utilizing a range of measures. Experiments are being conducted on the John Diabetes Database. With 99% accuracy, the decision tree technique is used to assess the acceptability of the desired system. Diabetes identification is one of the most important medical problems facing society today. This approach requires coordinated efforts to develop a system that can predict diabetes. We investigate and evaluate five machine learning classification technique in this study utilizing various criteria. The research is centered on the John Diabetes Database. Using the Decision Tree approach, experiments assess the suitability of the desired system with 99% accuracy.

Data mining and machine learning methods are beneficial for diagnosing diseases. Various categorization systems based on accuracy, with the capacity to anticipate diabetes early being a key premise. for medical diagnosis of diabetes patients. There is a classification statements of accuracy. The Pima Indian diabetes dataset was subjected to three machine learning methods. In addition, they were trained and validated on a test dataset. The results of our model implementations show that ANN outperforms the other models. The findings from association rule mining revealed a significant correlation between BMI and glucose levels and diabetes. This study's restriction is the use of a structured dataset, however unstructured data will likely also be used in the future.

The categorization outcome demonstrates the identification and labelling of three types of tissues. This tissue classification is used to determine the best course of action for treating diabetic patients' wounds so they recover quickly. However, the outcome needs to be verified using factors like accuracy. Additionally, verification of this detection system using More wound image data sets are being examined, as well as the effectiveness of various segmentation and classification algorithms based on various color and textural features.

Eight machine learning techniques are used in this study to examine the early-stage diabetes risk prediction dataset. Performance indicators including accuracy, recall, precision, and f-score are used to compare the results. The created 1-dimensional convolutional neural network model is the most effective one which, when applied to the data set, has a 99.04% accuracy rate using the 5-fold cross validation schema. On the early-stage diabetes risk prediction dataset, no studies using XG Boost or Convolutional Neural Networks have been published. The findings presented in this research demonstrate that both of these two approaches successfully identify the risk of diabetes in its early stages. Since the evaluated metrics are high, more effort could be put into developing an early-stage diabetes risk prediction application.

REFERENCES

[1] Anjuman, A.A., M.G. Ahamad, and M.K. Siddiqui (2013). Application of data mining: Treatment of patients with diabetes, both young and old.

[2] Bainite, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and

Computing 1, 763–770. doi:10.1007/978-3319-11933-5.

[3] Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima Indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451– 455

[4] Sisodia, D., Shrivastava, S.K., Jain, R.C., 2010. ISVM for face recognition. Proceedings 2010 International Conference on Computational Intelligence and Communication Networks CICN 2010, 554–559doi:10.1109/CICN.2010.109.

[5] Sciences 25, 127–136, King Saud University; Doi: 10.1016/j.jksuci.2012.10.003. Health Monitoring and Tracking of Soldier Using GPS," International Journal of Research inAdvent Technology, vol. 2, no. 4, pp. 291-294, April 2014. P. Kumar, G. Rasika, V. Patil, and S. Bo bade.

[6] "A Real Time Autonomous Soldier Health Monitoring and Reporting System Using COTS Available Entities," Second International Conference on Advances in Computing and Communication Engineering (ICACCE), Dehradun, India, May 2015, pp. 683-687.

[7] "An IoT based patient monitoring system using raspberry Pi", International Conference on Computing Technologies and Intelligent Data Engineering, Kovilpatti, India, January 2016, pp. 1-4.

[8] “Real Time Health Monitoring System of Remote Patient Using Arm7”, R. Shaikh, International Journal of Instrumentation, Control and Automation (IJICA), vol. 1, no.3-4, pp. 102-105, 2012.

[9] International Institute for Strategic Research (3 February 2014). Pages 241-246 of The Military Balance (2014). Routledge, London, ISB9781857437225.no. 3-4, pp. 102- 105, 2012.