2021 18th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE). Mexico City, Mexico. November 10-12, 2021

A Comparative Study with Different Machine Learning Algorithms for Diabetes Disease Prediction

Hafsa Binte Kibria, Abdul Matin, Nusrat Jahan, and Sanzida Islam
Department of Electrical & Computer Engineering,
Rajshahi University of Engineering & Technology, Rajshahi-6204, Bangladesh
Email: hafsabintekibria@gmail.com, ammuaj.cseruet@gmail.com 1510045@student.ruet.ac.bd, mousanzida7@gmail.com

***Abstract*—Diabetic is a disease that occurred when the level of blood glucose is higher than usual, which is also known as hyperglycemia. When the human body is incapable of producing enough insulin(a hormone that produces glucose from food), then this situation leads to diabetes. The rapid increase of this disease makes the researchers work much harder in this area to build a model for diagnosing diabetes efficiently. As in healthcare, the availability of data is high, so it is easy to extract information from those data to diagnose disease and develop a new model for better results. This paper aims to introduce a model that can predict diabetes efficiently with the help of machine learning algorithms. Here logistic regression, SVM, and k nearest neighbor algorithms have been used for the classification of diabetics. After data preprocessing and training, those algorithms gave a good result. Logistic regression provided the best accuracy of 83% for test data. Also, SVM and knn both performed well and showed an accuracy of 82% and 79%, respectively. The proposed model has demonstrated improved results compared with previous work.**

***Index Terms*—Logistic regression, Support vector machine, K nearest neighbor, Machine learning, Diabetes**

I. Introduction

Diabetes mellitus (DM) is frequently referred to as diabetes mellitus (DM). It is a set of disorders characterized by high sugar in the blood [1]. Diabetes mellitus is a condition when the body does not produce enough insulin to satisfy its needs, and so the level of blood sugar abnormally increases.It can lead to many severe complex long-term diseases such as heart attack, kidney failure, stroke, and death. In 1980 approximately 122 million people worldwide were affected by this disease, and this number reached around 422 million in 2014 [2]. In 2040 the figure will hit about 642 million [3]. Also, diabetes was primarily responsible for approximately 1.6 million deaths [4].

The economic growth of a country directly depends on healthcare development. A healthy person is considered an asset to a nation that can perform duties efficiently. That's why improving the quality of health care is a vital task for researchers. Technological use is proved very useful, such as machine learning in the medical sector [5].So researchers have been using it to improve health issues. Machine learning(ML) algorithms have significant importance for the prevention, detection, and treatment of different diseases. It is the most excellent source for improving the health care system. Diagnosing disease manually takes more time, and it gives low accuracy than ML techniques [6]. So ML helps to build an efficient model, which also takes very little time to diagnose disease. That's why the world is getting more and more dependent on data mining techniques.

In the medical industry, the significance of predictive analysis is getting high day by day. So predicting disease with the help of artificial intelligence has gained massive attention to researchers. The goal of machine learning applications is to train a computer that can perform like humans or even better than humans [7]. Generally, to train a model, supervised algorithms are used with data that are labeled, and for evaluation, testing data is used [8]. As the diabetics' data we used in this research contains non-linearity, so it is a little bit challenging to analyze those data [9].

This paper's main objective is to build a system that can predict the presence and absence of diabetes. The first section discusses diabetes and the importance of machine learning in the health care industry. The second segment of this paper talks about other studies regarding this topic. Next, we have discussed the materials that were used for our proposed system. The fourth section covers the idea of the proposed model. In the fifth part, our model's performance is evaluated. And at last, the conclusion is made.

II. Literature Review

We have applied machine learning algorithms to predict disease, and we have chosen diabetes diseases for prediction for our research work. Not all of the algorithms in machine learning always show good accuracy. Depends on the data, the performance of the algorithm varies. Also, the accuracy of a model depends on the tuning of parameters in the algorithms. We need to tune the hyperparameters of the algorithms according to the data type. Here we have addressed various kinds of algorithms for diabetes disease predictions and focussed on the respective algorithm's testing data accuracy. This review only focuses on the classification of diabetes disease.

In this paper [10], diverse sampling techniques were applied to the dataset. Then four data mining algorithms decision tree (DT), naive Bayes (NB), artificial neural network (ANN), and deep learning (DL) were applied for the evaluation. PIMA dataset was used for diabetes prediction, and among the four algorithms, DT gave the highest accuracy. In another study

978-1-6654-0029-9/21/$31.00 ©2021 IEEE

[11], the Weka tool was used to simulate the diabetes dataset. They have used some data mining algorithms and compared the results among them. Four classifier algorithms have been applied, and the support vector machine performed the best and provided an accuracy of 79%. They have also discussed the time taken for training and testing the data and other classification parameters such as recall,f1-score.

Models such as Random forest, different types of SVM (linear, polynomial, radial) and Decision tree were applied on the CKD, heart disease, and diabetics dataset [12]. Features were extracted using chi-square method to find out the essential attributes. The improved SVM radial method gave the best accuracy. The result was measured with accuracy, specificity, misclassification rate, precision, and sensitivity. These researchers proposed a generalized architecture for various disease predictions.

In [13], researchers designed a system using classification algorithms for diabetes disease prediction. They used Pima Indian diabetes dataset. Naïve Bayes, SVM, and decision tree algorithms have been used as classification algorithms. Here naïve Bayes provided the highest accuracy of 76.30%. The WEKA tool was used to perform the experiment. Other classification parameters were also measured along with accuracy. In this work, there is a scope for improving the accuracy.

Another study [14], used Genetic programming (GP) to train and test the database for the classification of diabetes by applying the diabetes data set collected from the UCI repository. Compared to other methods, the results obtained using Genetic Programming have the highest degree of accuracy. There was a major improvement in accuracy, and it also took less time. It appears to be useful for diabetes prediction at an effective cost. In [15] cuckoo search optimized reduction and fuzzy logic classifiers were applied for cardiovascular and diabetes disease prediction. Among the other prediction models, this model performed better. Data were preprocessed and normalized for better accuracy. Fuzzy logic was applied after the reduction of attributes. The model is complex which is a limitation.

In [16], they developed a diabetes prediction method, which purpose is to forecast diabetes that a candidate would suffer at a particular age. The suggested framework was structured to be based on the principle of machine learning by implementing a decision tree algorithm. The findings were satisfactory as the built method performs well to forecast diabetes events at a given age, with better accuracy using the decision Tree. So the vital focus of these studies is to try to increase the accuracy to get better performance using different machine learning algorithms.

III. Materials and Methodology

Here we have done quantitative research. As our goal is to predict disease, so the method we applied was quantitative. For our study, we have used secondary data, and we have solved a practical problem. Other researchers have also used this data for classifying diabetes disease.

A. Data Description

Data has been obtained from the archive of the UCI machine learning repository. Data was prepared before analysis. The data used in this paper is the PIMA Indian dataset. It has 768 instances with eight attributes. These eight attributes have the most influence on predicting people with diabetes. There are eight dependent attributes and one independent attribute having two class labels. These attributes are described in table I. Here 500 samples are those patients who have diabetes and 268 patients with no diabetes. It has also presented a summary of statistics of our dataset.Only the PIMA Indian dataset was chosen because it has been used by numerous researchers, and we wanted to compare our work to theirs by implementing different algorithms.

Figure 1 shows the histogram of all attributes in the dataset. It is a graphical representation of data of bar with various heights. It represents the distribution of numerical data.

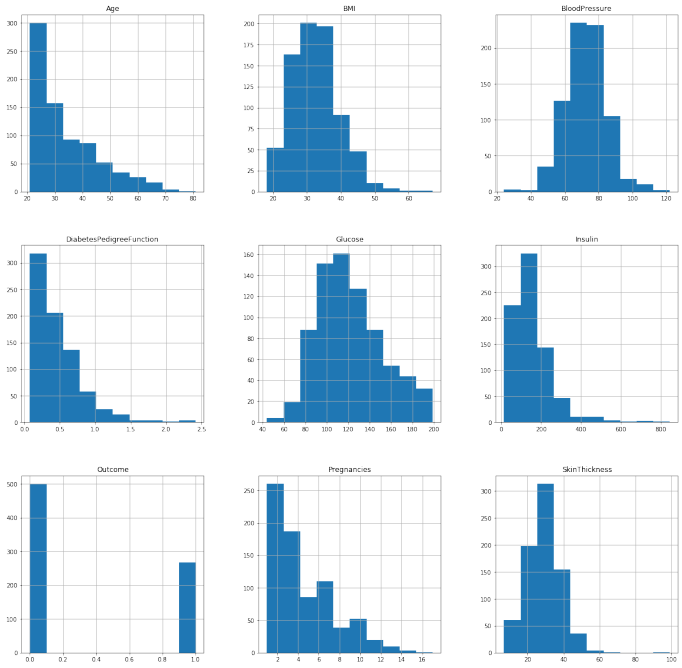

Fig. 1: Histogram of all attributes

B. Data Pre-processing

The preprocessing phase of data is an important stage in machine learning. It is very normal that there is a loss of data and ambiguity in the health sector. Our dataset (PIMA Indian) does not contain any noisy data but has several missing values. The missing values have been identified and replaced by numeric values. This method is called data imputation. The efficient method for data imputation is the use of a model capable of predicting missing values. Here, the K-nearest neighbor model was used for imputation. KNN is a better imputation strategy than the most frequent strategy or using the mean value. For example, if we take the mean value for age column, then all the missing values would've been replaced

TABLE I: Diabetes Disease Dataset

Attribute	Attribute Type	Attribute Description and Range	count	mean	std
Pregnancies	Numeric	Number of times pregnant(0-17)	768	3.845052	3.369578
Glucose	Numeric	Plasma glucose concentration a 2 hours in an oral glucose tolerance test(0-199)	768	120.89453	31.972618
BloodPressure	Numeric	Diastolic blood pressure (mm Hg)(0-122)	768	69.105469	19.355807
SkinThickness	Numeric	Triceps skin fold thickness (mm)(0-99)	768	20.536458	15.952218
Insulin	Numeric	2-Hour serum insulin (mu U/ml)(0-846)	768	79.799479	115.24400
BMI	Numeric	Body mass index weight in kg/(height in m)	768	31.992578	7.88416
DiabetesPF	Numeric	Diabetes pedigree function(0-2.42)	768	0.471876	0.331329
Age	Numeric	Age (years)(0-81)	768	33.240885	11.760232
Outcome	Numeric	Class variable (0 or 1)	768	0.348958	0.476951

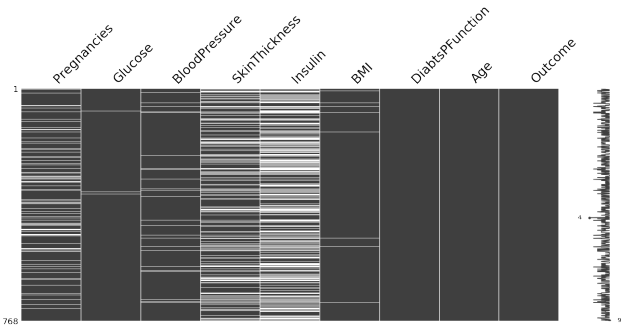

Fig. 2: Graphical representation of missing values

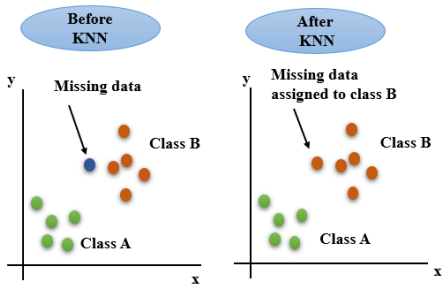

Fig. 3: KNN for data imputation

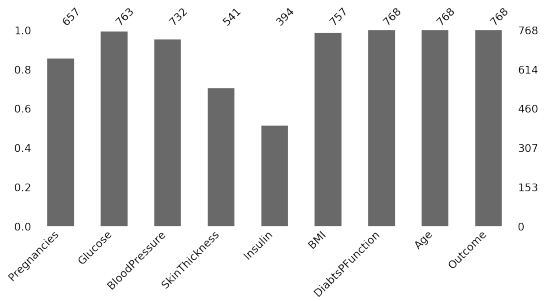

Fig. 4: Visualization of Missing values of all attributes

by the same number, which is not proper as the missing age number should be varied depending on the other symptoms. In KNN, this number was taken considering all the other symptoms in the row, and only that number was taken which fit the best for that row. It gives more accurate results compared to other approaches. That's why this method has been selected. This method locates the nearest samples in the training set and uses the average neighbor points to fill the missing values. The value of k was taken as five for imputation.

The distribution of missing values has been given in figure 2. Since the number of missing values in the dataset is too high, dropping the rows with missing values is not a good idea. We will lose a lot of information if we take this approach. That's why we went with the knn method for imputation. The white represents the missing values. Insulin has the highest number of white lines, which means it has the highest number of missing values.

The value of k was taken as five for imputation. In figure 3 the working of KNN has been displayed for missing data. Here we have shown only using two classes, but practically, there are eight classes in our model as the number of columns is eight. So for missing data, it considers the distance of all classes and selects the minimal distance for replacing the missing value.

Figure 4 displays the visualization of the missing values of all attributes. Out of the nine attributes, three attributes have no missing values at all.From this figure, we can know the actual number of missing values of any attribute. Figure 5 shows the representation after applying knn imputation to the dataset. In this figure, we can see that there are no missing values in the dataset as the missing values are being replaced. Now all the attributes have 768 data points.

C. *Feature extraction*

Feature extraction makes a new set of features from chosen features. It is a method of reducing main features into a minimized feature set. It reduces data volume and makes the

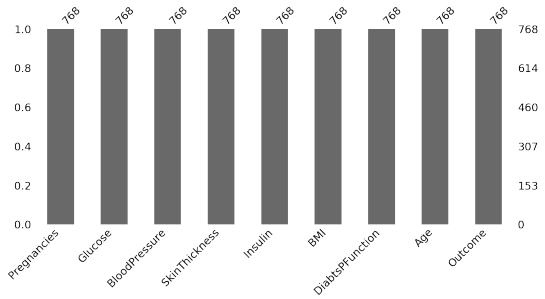

Fig. 5: Visualization After replacing all missing values

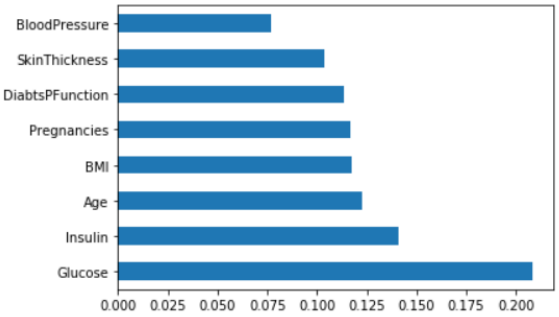

Fig. 6: Important attributes according to extra tree classifier algorithm

model simple. To select important features, we have used extra tree classifier. This algorithm generates a large number of decision trees from the training data. Predictions are made by taking the average of the decision trees' prediction in the case of regression. As we solved a classification problem, so here voting classifier was used for prediction. It selects the important features based on the votes. If the feature gets maximum votes, it means it has the strongest relationship towards the output. Based on the votes, it arranges the features according to their importance in ascending order which has been displayed in figure 6. This technique was applied after replacing the missing values of the dataset. We have used the sklearn library to implement extra tree classifier algorithm.

In the figure 6, the extracted features have been displayed. We have shown the essential features in ascending order from the top of the curve to down. We have used the extra tree classifier to determine the relative importance of all attributes for producing output.

D. Algorithms

1) Logistic regression(LR): Machine-learning is categorized into three kinds of learning: supervised learning, unsupervised learning, and reinforcement learning. Now, under supervised learning, there are two classes of problems. They are classification problems, and the other is a regression. Classification is about predicting a label. And regression is about predicting a quantity. Here we have solved a classification problem using machine learning techniques. The output of a linear regression model is always continuous. But when it comes to logistic regression, the resulting output is categorical. The graph of logistic regression is a curve which is called the S curve or the sigmoid curve. The sigmoid curve maps the relationship between the dependent and independent variables for logistic regression [17]. This is because logistic regression calculates the probability. In linear regression, the dependent variable is always continuous, and it predicts continuous values. But logistic regression uses categorical dependent variables to predict categorical values. LR is based on maximum likelihood estimation. This implies that the coefficients will be selected in such a way that it maximizes the probability of Y (dependent variable) given some value of x (of independent variables). We have tuned the parameter C in logistic regression, which is the strength of the regularization. We took C =100 for our proposed model.

2) KNN: KNN is a supervised algorithm that can be used for both regression and classification predictive problems. It identifies the neighbor and based on the class of neighbors, it gives decision for a new value. The decision depends on the k value. The parameter 'k' in KNN corresponds to the number of closest neighbors for determining the majority of the voting process. The value of k is always set as odd values for each class problem to prevent a tie situation where both classes will have the same votes. The data record that has to be classified first measures the distance between the data and all of the reference points, then it searches at the K closest data in the reference data [18]. A problem in the algorithm is that the complexity in searching the nearest neighbors for each sample.

3) SVM: One of the common types of supervised machine learning models is SVM. For a two-class classification problem, the SVM finds the margin that has the maximum distance from the hyperplane. The more the distance, the easier it will be to classify data. The hyperplane should not be closer to the data points belonging to the other class for better classification. Hyperplane, which is far from the data points from each group, should be chosen. Support vectors are the points that lie nearest to the margin of the classifier [13]. In SVM, we have tuned the hyperparameter c and gamma. C is used to control error, and gamma gives the weight of the decision boundary's curvature. So by adjusting these two hyperparameters, we made our model more efficient to predict the result. The value of c we took was 100, and gamma was taken as .01. the value of C and gamma varies for the different datasets.

To represent the performance of our trained algorithms, we have used some evaluation matrices. The description of them is given in the table II.

IV. Proposed approach

This research aims to develop a system that can predict the presence and absence of diabetes in patients by analyzing the symptoms. There are several steps in this procedure which have been shown in the figure. At first preprocessing step come where data was cleaned and preprocessed.

1) Data collection
2) Replacing missing values (Imputation): KNN method
3) Feature selection
4) Label encoding
5) Normalization: min-max scaler
6) Data split (75% training & 25% testing)
7) Prediction using ML methods

The inputs which will be used for the analysis are age, glucose level, insulin, etc. Eight attributes are the inputs for our algorithms. 75% data was used for training, and 25% was for testing. First, the data were preprocessed, which includes the step of cleaning, replacing missing values. After that, using a min-max scaler, data were scaled from 0 and 1. It handles various types of magnitude in the data and brings all values

TABLE II: Evaluation matrix

Measures	Definitions	Formula
Accuracy(A)	For test data, accuracy is the percentage of correct predictions	A=(TP+TN)/Total no of samples
Precision(P)	It is the amount of positively predicted value	P= TP/(TP+FP)
Recall(R)	It is the percentage of correctly classified data by algorithm	R= TP/(TP+FN)
F1 score(F)	It measures the balance of precision and recall	F=2(P*R)/(P+R)
ROC	Score at all classification thresholds are displayed by receiver operating characteristic curve	

TABLE III: Confusion matrix for KNN

	Actually positive	Actually negative
Predicted positive	116	20
Predicted negative	16	40

TABLE IV: Confusion matrix for SVM

	Actually positive	Actually negative
Predicted positive	110	11
Predicted negative	29	42

TABLE V: Confusion matrix for LR

	Actually positive	Actually negative
Predicted positive	112	11
Predicted negative	23	46

in the same range. After preprocessing, data was trained. Data used for training are labeled, and after training, these trained models predict for the test data that are not labeled. The trained machine learning models will classify the patients who are with and without diabetes. Three algorithms have been for the training. They are support vector machine, k nearest neighbor, and logistic regression. At last, these trained models were used to predict the output for any new data.

For interpretation and viewing of the results, we used Matplotlib. We used the web application of Jupyter Notebook. Scikit-learn library was used, and python was used for coding. Since two-class classifications are our concern, we have chosen algorithms to classify two-class classification problems with good accuracy.

In figure 7, the whole procedure has been displayed. after collecting data, it was preprocessed. The steps of preprocessing have been mentioned in figure 7. Our paper has briefly discussed these steps after preprocessing. Then feature scaling has been done to keep all the values in a range. We have divided our dataset into train and test datasets. The training dataset was used for training with the algorithms, and after training, the trained models predict for the test data. The evaluation matrix has been measured based on correct and incorrect predictions.

V. Result and evaluation

We have used three algorithms for the prediction of diabetics disease. It is a two class classification problem and displays the output as the presence and absence of the disease. Among the three algorithms, logistic regression showed the best accuracy of 83.00% for the test data in table VI. KNN also performed well and provided an accuracy of 82% for the test data. The accuracy for the training data is 78%. From the accuracy of both test and train, we can say that the models are neither overfitting nor underfitting. The dataset is balanced, and that's why we also get a satisfying value for the f1-score and other parameters.

In knn approach, we took the value of k=23 and got an accuracy of 82% for the test data. In the figure 8, the variation of train and test data's accuracy has been displayed with the variation of knn neighbor number.

The confusion matrix for knn, svm and lr have been shown in table III, IV and V respectively.

In table VI, the output of the three algorithms has been shown. F1-score is also highest for logistic regression. Compared with other algorithms, logistic regression worked better. Since our problem is a problem of binary classification, and logistic regression is a good option for two classifications, that's why this algorithm gave us a good performance. Other algorithms such as KNN and SVM also performed well. Here, and the result is knn is closed to the output of logistic regression. It showed an accuracy of 82% for the test data. So we got a pretty good result from both of the algorithms.

In figure 9, the comparison of all algorithms has been shown.

There has been a lot of research on the classification of diabetes disorders so far. To compare them with our models, we have arranged some recent works in table VII. The researchers used numerous algorithms to identify diabetes in these experiments. Some of them provided a good result. And for some approaches, more improved output can be gained using other methods. From table VII it is noticed that the highest accuracy they got in [19] using ANN is 82%. Also, the support vector machine provided an accuracy of 82%.

Roc is the receiver operating characteristic curve that represents the performance of a model. It is a curve related to various threshold values which are true positive rate vs. false positive rate. We calculate the AUC score, which is the area under the curve, from the ROC curve. The greater the area under the curve, the better the model is.

In figure 10, 11 and 12, the ROC curve for SVM, LR and KNN has been displayed. The highest roc score we got from logistic regression that is 79%.

VI. Conclusion

People with diabetes are a widespread concern, and it will be helpful if a machine learning-based system can identify them early. It will make the healthcare system more effective and will reduce time and cost. Furthermore, it will reduce the pressure of the physicians. So here, to identify patients with and without diabetics, a predictive method was created.

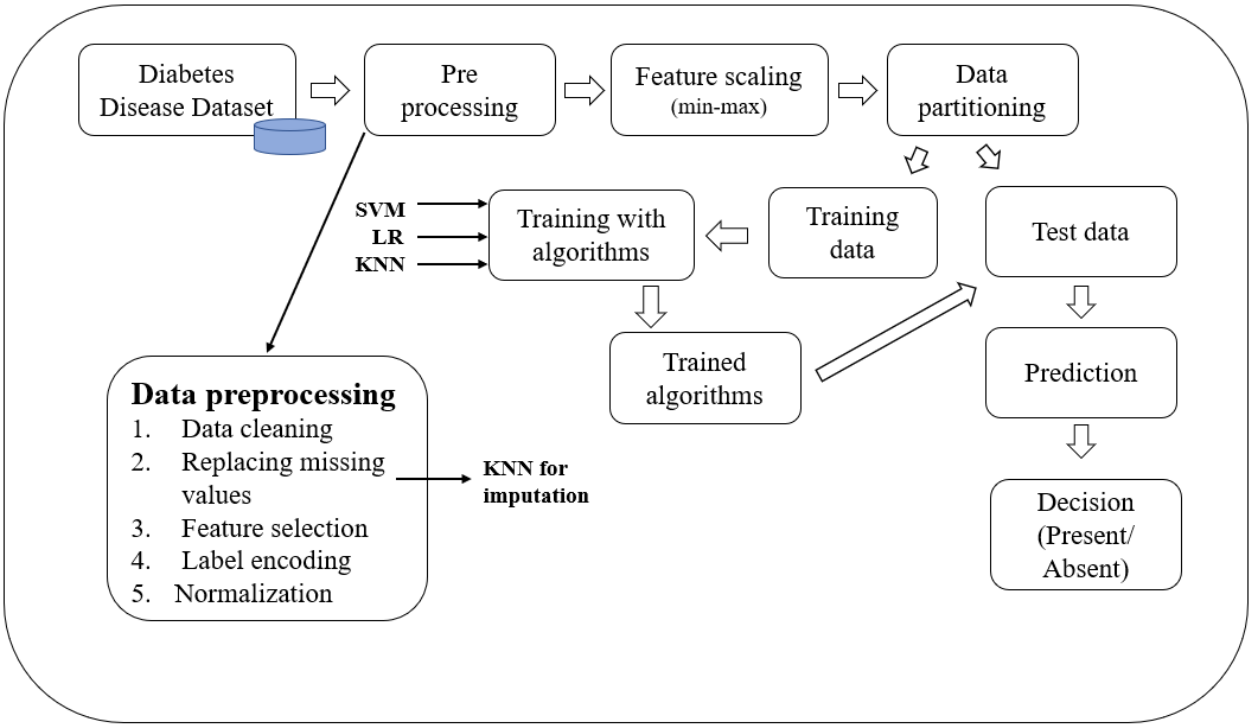

Fig. 7: Architecture of proposed model

TABLE VI: Performance of all algorithms

Classification algorithms	Precision	Recall	F1-score	Accuracy Train	Accuracy Test	Roc-Auc score
Support Vector Machine	79	79	78	76.38	79.16	75.34
Logistic regression	82	82	82	76.04	83.00	78.89
K Nearest Neighbor (KNN)	82	81	81	78.00	82.00	78.36

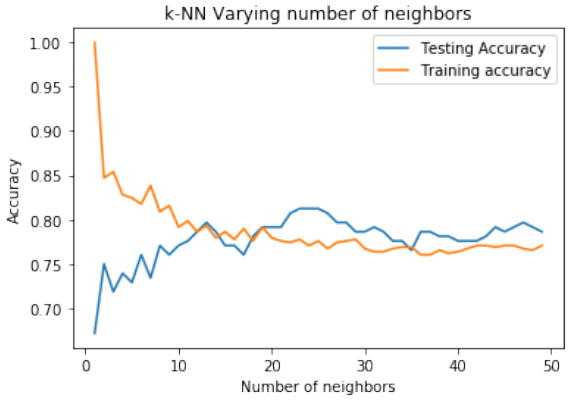

Fig. 8: Variation of the accuracy of train and test regarding the KNN neighbor number

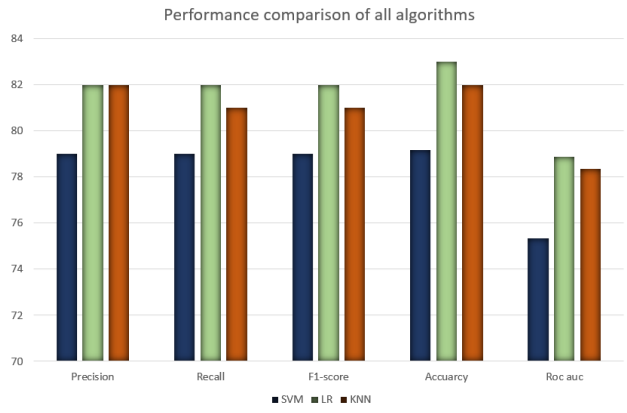

Fig. 9: The performance parameter comparison among all the algorithms

Three machine learning algorithms were used to classify the problem, and the most satisfactory output was given by logistic regression. It gave an accuracy of 83%, where SVM and KNN provided an accuracy of 82% and 79%, which is preferable compared to the recent works. One of the limitations is that we have used the classification algorithms to predict the diabetes

TABLE VII: Previous work with same dataset

Author	Year	Approach	Accuracy
[20]	2021	Artificial neural network	71.35
		XGboost	78.91
		Support vector machine	77.73
		Random forest	75.39
[21]	2019	Naive Bayes	73.48
		Decision tree	73.18
		K-nearest neighbors	63.04
		Decision tree	74.00
[19]	2019	Support vector machine	82.00
		Gaussian naive bayes	80.00
		Artificial neural network	82.00
[22]	2018	Firefly and Cuckoo Search Algorithm	81.00
[23]	2018	Feedforward neural network	82.00
		Naive Bayes	76.30
[13]	2018	Support vector machine	65.10
		Decision tree	73.82

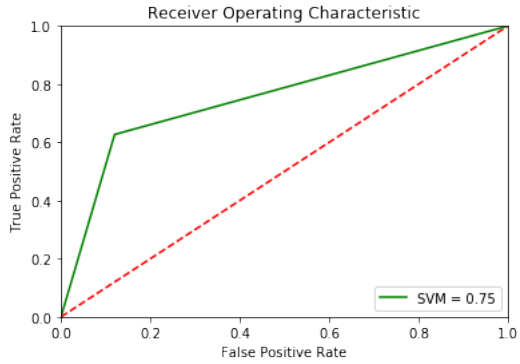

Fig. 10: ROC curve of support vector machine

disease only, however, others diseases can also be predicted using these algorithms. It is also possible to implement other hybrid algorithms to gain better accuracy. In the future, we have planned to use fusion models to get better accuracy.

REFERENCES

[1] A. D. Association *et al.*, "Diagnosis and classification of diabetes mellitus," *Diabetes care*, vol. 33, no. Supplement 1, pp. S62–S69, 2010.

[2] N. R. F. Collaboration *et al.*, "Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19· 2 million participants," *The Lancet*, vol. 387, no. 10026, pp. 1377–1396, 2016.

[3] P. Zimmet, K. G. Alberti, D. J. Magliano, and P. H. Bennett, "Diabetes mellitus statistics on prevalence and mortality: facts and fallacies," *Nature Reviews Endocrinology*, vol. 12, no. 10, p. 616, 2016.

[4] C. Bharath, N. Saravanan, and S. Venkatalakshmi, "Assessment of knowledge related to diabetes mellitus among patients attending a dental college in salem city-a cross sectional study," *Brazilian Dental Science*, vol. 20, no. 3, pp. 93–100, 2017.

[5] P. Yancy and A. Handley, "Can everyone be refreshed?" *The Journal of Continuing Education in Nursing*, vol. 35, no. 2, pp. 80–83, 2004.

[6] R. Birjais, A. K. Mourya, R. Chauhan, and H. Kaur, "Prediction and diagnosis of future diabetes risk: a machine learning approach," *SN Applied Sciences*, vol. 1, no. 9, p. 1112, 2019.

[7] J. B. Heaton, N. G. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios," *Applied Stochastic Models in Business and Industry*, vol. 33, no. 1, pp. 3–12, 2017.

[8] K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, "Enhanced fingerprinting and trajectory prediction for iot localization in smart buildings," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 3, pp. 1294–1307, 2016.

[9] M. Maniruzzaman, N. Kumar, M. M. Abedin, M. S. Islam, H. S. Suri, A. S. El-Baz, and J. S. Suri, "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," *Computer methods and programs in biomedicine*, vol. 152, pp. 23–34, 2017.

[10] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using pima indian dataset," *Journal of Diabetes & Metabolic Disorders*, vol. 19, no. 1, pp. 391–403, 2020.

[11] A. Mir and S. N. Dhage, "Diabetes disease prediction using machine learning on big data of healthcare," in *2018 fourth international conference on computing communication control and automation (ICCUBEA)*. IEEE, 2018, pp. 1–6.

[12] K. Harimoorthy and M. Thangavelu, "Multi-disease prediction model using improved svm-radial bias technique in healthcare monitoring system," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–9, 2020.

[13] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia computer science*, vol. 132, pp. 1578–1585, 2018.

[14] M. Pradhan and G. Bamnote, "Design of classifier for detection of diabetes mellitus using genetic programming," in *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014*. Springer, 2015, pp. 763–770.

[15] T. R. Gadekallu and N. Khare, "Cuckoo search optimized reduction and fuzzy logic classifier for heart disease and diabetes prediction," *International Journal of Fuzzy System Applications (IJFSA)*, vol. 6, no. 2, pp. 25–42, 2017.

[16] K. M. Orabi, Y. M. Kamal, and T. M. Rabah, "Early predictive system for diabetes mellitus disease," in *Industrial Conference on Data Mining*. Springer, 2016, pp. 420–427.

[17] H. B. Kibria, A. Matin, and S. Islam, "Comparative analysis of two artificial intelligence based decision level fusion models for heart disease prediction," in *International Semantic Intelligence Conference*, vol. 2786. ceur-ws.org, 2020, pp. 314–322.

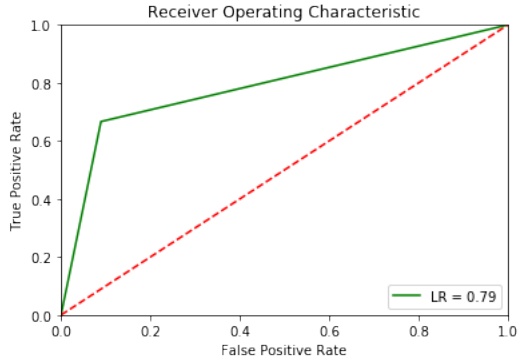

Fig. 11: ROC curve of logistic regression

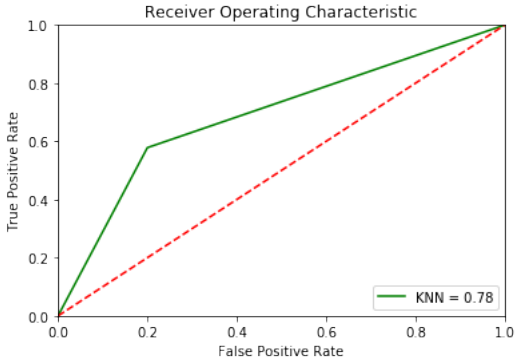

Fig. 12: ROC curve of k-nearest neighbor

[18] H. B. Kibria and A. Matin, "An efficient machine learning-based decision-level fusion model to predict cardiovascular disease," in *International Conference on Intelligent Computing & Optimization*. Springer, 2020, pp. 1097–1110.

[19] P. Sonar and K. JayaMalini, "Diabetes prediction using different machine learning approaches," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2019, pp. 367–371.

[20] P. Tiwari and V. Singh, "Diabetes disease prediction using significant attribute selection and classification approach," in *Journal of Physics: Conference Series*, vol. 1714, no. 1. IOP Publishing, 2021, p. 012013.

[21] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big data*, vol. 6, no. 1, pp. 1–19, 2019.

[22] R. Haritha, D. S. Babu, and P. Sammulal, "A hybrid approach for prediction of type-1 and type-2 diabetes using firefly and cuckoo search algorithms," *International Journal of Applied Engineering Research*, vol. 13, no. 2, pp. 896–907, 2018.

[23] Y. Zhang, Z. Lin, Y. Kang, R. Ning, and Y. Meng, "A feed-forward neural network model for the accurate prediction of diabetes mellitus," *International Journal of Scientific and Technology Research*, vol. 7, no. 8, pp. 151–155, 2018.