KICS The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com

ScienceDirect

ICT Express 7 (2021) 432–439

www.elsevier.com/locate/icte

A comparison of machine learning algorithms for diabetes prediction

Jobeda Jamal Khanam, Simon Y. Foo*

Department of Electrical and Computer Engineering, FAMU-FSU College of Engineering, Tallahassee, FL 32310, USA

Received 20 August 2020; received in revised form 2 January 2021; accepted 11 February 2021
Available online 20 February 2021

Abstract

Diabetes is a disease that has no permanent cure; hence early detection is required. Data mining, machine learning (ML) algorithms, and Neural Network (NN) methods are used in diabetes prediction in our research. We used the Pima Indian Diabetes (PID) dataset for our research, collected from the UCI Machine Learning Repository. The dataset contains information about 768 patients and their corresponding nine unique attributes. We used seven ML algorithms on the dataset to predict diabetes. We found that the model with Logistic Regression (LR) and Support Vector Machine (SVM) works well on diabetes prediction. We built the NN model with a different hidden layer with various epochs and observed the NN with two hidden layers provided 88.6% accuracy.

Keywords: Machine learning; Data Mining; Neural Network; K-fold Cross Validation; Accuracy

1. Introduction

The WHO (World Health Organization) reported that around 1.6 million people die due to diabetes every year [1]. Diabetes is one kind of disease that occurs when the blood glucose/blood sugar level in the human body is very high. According to health experts, diabetes occurs when the human body's gland called the pancreas cannot produce enough insulin (Type 1 diabetes), and the produced insulin cannot be used by the cell of the body (Type 2 diabetes) [2]. When we eat food, after the digestion process, glucose gets released. Insulin is a blood hormone that moves from blood to cells and instructs cells to consume blood glucose and transform it into energy. When the pancreas cannot produce enough insulin, the cells cannot absorb glucose, and the glucose remains in the blood. Hence the blood glucose/blood sugar increases in the blood at a very unacceptable level [3]. Due to high blood sugar, some symptom arises in the human body, such as extreme hunger, intense thirst, and frequent urination. The usual range of glucose levels in the human body is 70 to 99 mg per deciliter. If the glucose level is more than 126 mg/dl, it indicates diabetes. A person is considered to have prediabetes if body glucose concentration is 100 to 125 mg/dl [4]. If the human body's blood sugar level becomes too high, the impending complications can be heart disease, kidney failure, stroke, and nerve damage [5,6]. There is no permanent cure for diabetes [7]. The most common long-term diabetes causes health problems, which are macrovascular and microvascular complications. The macrovascular complication is damage to the large blood vessels of the heart, brain, and legs. Microvascular complication damages the small blood vessels, causing problems in the kidneys, eyes, feet, and nerve [8]. The efficient control of diabetes is possible if it can be detected early. Maintaining an effective fitness system and balanced eating habits can help to prevent diabetes [9]. If a patient has prediabetes, losing bodyweight by getting physical activity can lower the risk of developing Type 2 diabetes. The Center for Disease Control and Prevention (CDC)-led National Diabetes Prevention Program, a lifestyle change program, can help to change a prediabetes patient's lifestyle and prevent developing Type 2 diabetes [10]. The healthcare industry collects an enormous amount of data include hospital records, medical records of patients, and results of medical examinations. For early disease diagnosis, the disease's prediction is analyzed through a doctor's experience and knowledge, but that can be inaccurate and susceptible. Hence the manual decisions can be alarming. The hidden pattern of data can be unnoticed, which can impact decision-making; therefore, patients become deprived of the appropriate treatment. Automated identification with better accuracy is essential for the early detection of diabetes [11–13].

* Corresponding author.
E-mail addresses: jk16c@my.fsu.edu (J.J. Khanam), foo@eng.famu.fsu.edu (S.Y. Foo).

Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS).

https://doi.org/10.1016/j.icte.2021.02.004

Data mining and machine learning have been developing, reliable, and supporting tools in the medical domain in recent years. The data mining method is used to preprocess and select the relevant features from the healthcare data, and the machine learning method helps automate diabetes prediction [14]. Data mining and machine learning algorithms can help identify the hidden pattern of data using the cutting-edge method; hence, a reliable accuracy decision is possible. Data Mining is a process where several techniques are involved, including machine learning, statistics, and database system to discover a pattern from the massive amount of dataset [15]. According to Nvidia: Machine learning uses various algorithms to learn from the parsed data and make predictions [16].

2. Literature review

Several scholars used the machine learning (ML) method to predict diabetes using Pima Indian diabetes dataset (PIDD). The Pima Indian Diabetes dataset (PIDD) having: 9 attributes, 768 records describing female patients. Some closely related works are discussed in this section [17–21].

Alam, T.M.et al. [17] showed 75.7% accuracy by applying the ANN technique on PIDD. Sajida Perveen et al. [22] used a dataset incorporated in this research is obtained from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). The CPCSSN dataset contained in this research includes information related to systolic blood pressure (sBP), diastolic blood pressure (dBP), HDL, triglycerides (TG), BMI, fasting blood sugar (FBS), and gender. They used Bootstrap aggregating, Adaptive Boosting, and the decision tree model. They found for better accuracy, Adaboost can be applied to predict diseases like diabetes, coronary heart disease, and hypertension. Sisodia et al. [18] found that, among the applied machine learning methods SVM, NB, and DT on PIDD, the NB classifier shows better accuracy at 76.30%. Tigga et al. in [19] applied logistic regression on PIDD for diabetic prediction. They found the number of pregnancies, BMI, and glucose level are the most significant variables for diabetes prediction among all features in PIDD. The Pima Indian Diabetes dataset is taken for analysis, and RStudio is used to process and visualize the result. Their model is showing pretty good prediction with an accuracy of 75.32%. In Amour Diwani et al.'s study [20], all the patient's data are trained and tested using 10 cross-validations with Naive Bayes and decision tree. Then the performance was evaluated, investigated, and compared with other classification algorithms using WEKA. The results predicted that the best algorithm is Naive Bayes with an accuracy of 76.3021%. In Zou et al.'s [21] study, they applied Random Forest, Decision Tree, ANN for classification algorithm on PIDD after the feature reduction using Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR) methods. They found that Pima Indians' best accuracy is 77.21% obtained from the random forest with the mRMR feature reduction method.

The most critical problem in the machine learning method is to choose the logical features and the appropriate classifier. In our work, we used Pearson's correlation method to find logical features. Our research work is to predict a patient has diabetes or not. In this work, to predict diabetes in a patient, different machine learning classification algorithms like Naive Bayes (NB), SVM, Linear Regression (LR), Adaboost, Random Forest, K Nearest Neighbor (KNN), Decision Tree (DT), and Neural Network (NN) with different hidden layer are used and evaluated on the dataset. The evaluation of the performance of all the classification methods is done with various measurement methods.

3. Methods

3.1. Data, feature, and software tool

In our research, the Pima Indian diabetes (PID) dataset is collected from the UCI Machine Learning Repository, which is originated from the national institute of diabetes and digestive and kidney diseases (NIDDK). In the PID dataset, all the patients are female, and at least 21 years old. The dataset contains information about 768 patients and their corresponding nine unique attributes. Table 1 shows the description of the attributes of this dataset. The nine attributes that are used for the prediction of diabetes are Pregnancy, BMI, Insulin level, Age, Blood pressure, Skin thickness, Glucose, Diabetes pedigree function, and Outcome. The 'outcome' attribute is taken as a dependent or target variable, and the remaining eight attributes are taken as independent/feature variables. The diabetes attribute 'outcome' consists of binary value where 0 means non-diabetes, and 1 implies diabetes [23]. In our research, we used data mining and machine learning algorithms to predict whether a patient has diabetes or not with enhanced accuracy. Obesity dramatically increases people's risk of developing Type 2 diabetes. Table 1 shows that the average body mass index is 32 for the 768 patients. The dataset is for Type 2 diabetes patients, as people with a BMI of 30 or greater are considered obese [24].

We used Weka, an open-source machine learning, and data mining software tool for the diabetes dataset's performance analysis. Weka contains tools for data preprocessing, clustering, classification, regression, visualization, and feature selection [25]. The Neural Network is implemented in the Jupyter Notebook, and the Python programming language is used for coding [26].

3.2. Data preprocessing

Preprocessing helps transform data so that a better machine learning model can be built, providing higher accuracy. The preprocessing performs various functions: outlier rejection, filling missing values, data normalization, feature selection to improve the quality of data. In the dataset, 268 samples are classified as diabetic, and 500 were non-diabetics.

3.2.1. Missing value identification

Using the excel and weka tool, we got the missing values in the datasets, shown in Table 2. We replaced the missing value with the corresponding mean value.

Table 1
The attributes of PIMA dataset.

Attribute	Description	Type	Average/Mean
Preg	Number of times pregnant.	Numeric	3.85
Glucose	Plasma glucose concentration 2 h in an oral glucose tolerance test.	Numeric	120.89
BP	Diastolic blood pressure (mm Hg).	Numeric	69.11
SkinThickness	Triceps skinfold thickness (mm).	Numeric	20.54
Insulin	2-hour serum insulin (μIU/mL).	Numeric	79.80
BMI	Body mass index (kg/m^2).	Numeric	32
DPF	Diabetes pedigree function.	Numeric	0.47
Age	Age (years).	Numeric	33
Outcome	Diabetes diagnose results (tested_positive: 1, tested_negative: 0)	Nominal	–

Table 2
The number of missing values in PIMA dataset.

Attributes	No. of missing values
Preg	0
Glucose	5
BP	35
SkinThickness	227
Insulin	374
BMI	11
DPF	0
Age	0

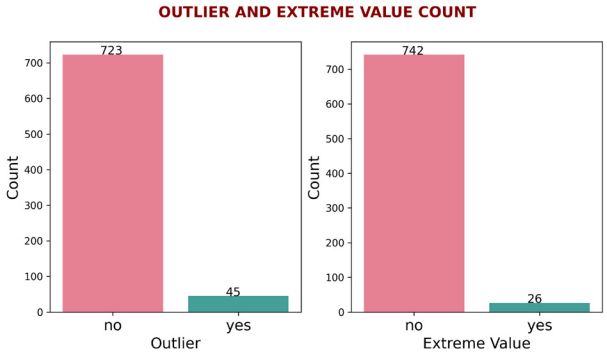

Fig. 1. Outlier and extreme value count.

3.2.2. Outlier identification and removal

Using the Weka tool, we filtered the dataset for detecting outliers and extreme values based on interquartile ranges. The number of outliers and extreme values are shown in Fig. 1, where we can see that there are 45 outliers and 26 extreme values. There were 699 instances after removing these outliers and extreme values from the dataset.

3.2.3. Feature selection

Pearson's correlation method is a popular method to find the most relevant attributes/features. The correlation coefficient is calculated in this method, which correlates with the output and input attributes. The coefficient value remains in the range between −1 and 1. The value above 0.5 and below −0.5 indicates a notable correlation, and the zero value means no correlation. In Weka, the correlation filter is used to find the correlation coefficient, and the results are shown in Table 3. We used 0.2 as a cut-off for relevant attributes. Hence SkinThickness, BP, DPF features are removed. Glucose, BMI, Insulin, Preg, and Age are our most relevant five input attributes.

Table 3
The correlation between input and output attributes.

Attributes	Correlation coefficient
Glucose	0.484
BMI	0.316
Insulin	0.261
Preg	0.226
Age	0.224
SkinThickness	0.193
BP	0.183
DPF	0.178

Table 4
Mean and standard deviation after normalization.

Attributes	Mean	Standard deviation
Preg	0.23	0.20
Glucose	0.48	0.19
Insulin	0.50	0.18
BMI	0.35	0.17
Age	0.20	0.19

3.2.4. Normalization

We performed feature scaling by normalizing the data from 0 to 1 range, which boosted the algorithm's calculation speed [27]. The mean and standard deviation results for all attributes after normalization are shown in Table 4.

In Fig. 2, we can see that, after completing preprocessing, we have 699 samples/instances where 466 patients have no diabetes, and 233 patients have diabetes. After preprocessing, the correlation between input and output attributes is shown in Fig. 3. In Fig. 3, we can see that 'Glucose' and 'Outcome' have a 0.46 correlation coefficient. Hence these are highly correlated.

3.3. Dataset train and test method

After data cleaning and preprocessing, the dataset becomes ready to train and test. We used K-fold cross-validation and 85% train/test splitting method separately to test the different machine learning model's performance. In the train/split

Table 5
Confusion matrices for DT, KNN, RF, NB, AB, LR, SVM classifier.

Test method	LR			KNN			SVM			NB			DT			RF			AB		
		0	1		0	1		0	1		0	1		0	1		0	1		0	1
K-fold cross-validation	0	409	57	0	387	79	0	414	52	0	386	80	0	382	84	0	391	75	0	401	65
	0	105	128	1	95	138	1	110	123	1	91	142	1	96	137	1	100	133	1	117	116
		0	1		0	1		0	1		0	1		0	1		0	1		0	1
Train/test splitting	0	101	18	0	97	22	0	101	18	0	98	21	0	94	25	0	96	23	0	102	17
	1	19	37	1	14	42	1	21	35	1	17	39	1	22	34	1	17	39	1	19	37

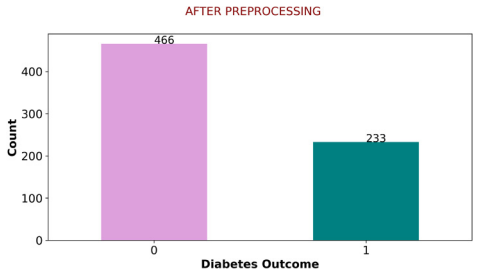

Fig. 2. After preprocessing the number of diabetes and non diabetes patients.

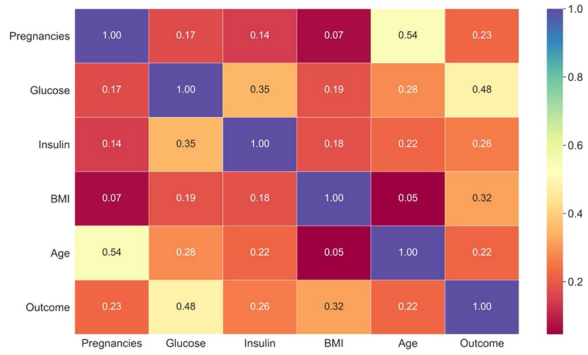

Fig. 3. After preprocessing correlation between input and output attributes.

method, we split the dataset randomly into the training and testing set. In the K cross-validation method, the data is divided into K folds. One-fold is used for validation/testing, and the remaining K-1 folds are used for training. The procedure will continue until every single K fold is a test set. The performance is measured by the average of all recorded scores of the Kth test.

3.4. Design and implementation of classification model

In this research work, comprehensive studies are done on the PIDD applying different ML classification techniques like DT, KNN, RF, NB, AB, LR, SVM, and neural network (NN). We used Kth value = 7 for the KNN algorithm. The proposed model diagram is shown in Fig. 4.

3.5. Neural network model implementation

We built three different neural network models with varying levels of hidden layers. We implemented the neural network with hidden layers 1, 2, and 3 with different epochs (200, 400, 800), and the results are compared. In ANN, the weighted sum of input is processed by the activation function in the hidden layer. We used two types of activation functions in our work, sigmoid and RELU. We used Keras and TensorFlow library to create the neural network models. We used a Sequential class from Keras library. The target variable is the 'Outcome' attribute. In ANN, the optimizer is required to reduce the output error during the backpropagation method. We used SGD (Stochastic Gradient Descent) as an optimizer. The learning rate is a parameter in an optimization algorithm that controls the weight adjustment with respect to loss gradient. We used different learning rates to find an effective one. From the scikit-learn library, we used the train_test_split function to perform the train/test splitting task. We used the cross_val_score function from the scikit_learn library for the K-fold cross-validation task. As the target variable is binary, the 'StratifiedKfold' technique is used in our work.

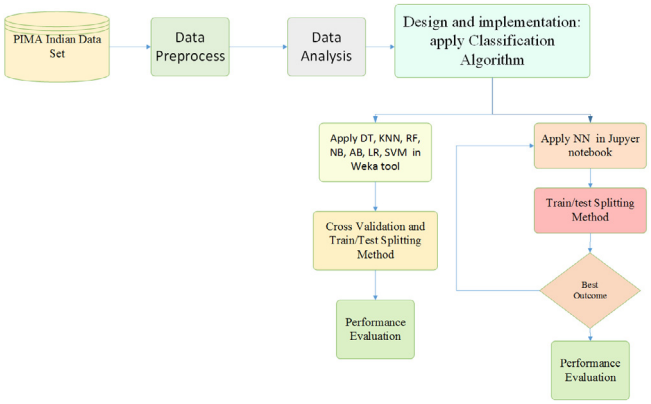

Fig. 4. Proposed model diagram.

3.5.1. Developing a NN model with one hidden layer

At first, we built a neural network with one hidden layer in addition to the input and output layer. We defined the input layer has five neurons, as there are five features. The hidden layer has five neurons and the RELU activation function. The output layer has one neuron and a sigmoid activation function. The model summary of NN with one hidden layer is given below in Fig. 5.

3.5.2. Developing a neural NN with two hidden layers

Here, we have defined a NN model with four dense layers. The first and fourth layers are input and output layers, respectively, having the same input shape, neurons, and activation

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape
=================================================================
dense (Dense)                (None, 5)
_________________________________________________________________
dense_1 (Dense)              (None, 5)
_________________________________________________________________
dense_2 (Dense)              (None, 1)
=================================================================
```

Fig. 5. NN model with one hidden layer.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape
=================================================================
dense (Dense)                (None, 5)
_________________________________________________________________
dense_1 (Dense)              (None, 26)
_________________________________________________________________
dense_2 (Dense)              (None, 5)
_________________________________________________________________
dense_3 (Dense)              (None, 1)
=================================================================
```

Fig. 6. NN model with two hidden layers.

function as NN with one hidden layer. The second layer consists of a hidden layer with 26 neurons, and the third layer consists of a hidden layer with 5 neurons. The activation function of the neurons of each hidden layer is RELU. The model summary of NN with two hidden layers is given below in Fig. 6.

3.5.3. Developing a NN model with three hidden layers

Here, we developed a model having five dense layers. The first and fifth layers are input and output layers, respectively, having the same input shape, neurons, and activation function as NN with one hidden layer. The second, third, and fourth hidden layers have 16, 10, 5 neurons, respectively. The activation function of the neurons of each hidden layer is RELU. The model summary of NN with three hidden layers is given below in Fig. 7.

4. Result and discussion

4.1. Results for ML method DT, KNN, RF, NB, AB, LR, SVM

The accuracy of the machine learning algorithm can be calculated from the confusion matrix. In the abstract term, the confusion matrix is given below.

	Predicted No (0)	Predicted Yes (1)
Actual No (0)	TN	FP
Actual Yes (1)	FN	TP

Here, FP = False Positive, FN = False Negative, TN = True Negative, and TP = True Positive. Eqs. (1)–(4) are used to calculate the performance measurement of the classification method.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FN} + \text{FP})} \tag{1}$$

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape
=================================================================
dense (Dense)                (None, 5)
_________________________________________________________________
dense_1 (Dense)              (None, 16)
_________________________________________________________________
dense_2 (Dense)              (None, 10)
_________________________________________________________________
dense_3 (Dense)              (None, 5)
_________________________________________________________________
dense_4 (Dense)              (None, 1)
_________________________________________________________________
```

Fig. 7. NN model with three hidden layers.

Table 6
The performance measure of all classification methods for K-fold cross-validation and Train/Test splitting method.

Classification	Precision	Recall	F-measure	Accuracy
DT (K-fold)	0.739	0.742	0.741	74.24%
DT (Splitting)	0.735	0.731	0.733	73.14%
RF (K-fold)	0.744	0.750	0.746	74.96%
RF (Splitting)	0.779	0.771	0.774	77.14%
NB (K-fold)	0.753	0.755	0.754	75.53%
NB (Splitting)	0.787	0.783	0.785	78.28%
LR (K-fold)	0.761	0.768	0.761	76.82%
LR (Splitting)	0.788	0.789	0.788	78.85%
KNN (K-fold)	0.747	0.751	0.749	75.10%
KNN (Splitting)	0.804	0.794	0.798	79.42%
AB (K-fold)	0.730	0.740	0.730	73.96%
AB (Splitting)	0.792	0.794	0.793	79.42%
SVM (K-fold)	0.761	0.768	0.759	76.82%
SVM (Splitting)	0.774	0.777	0.775	77.71%

$$\text{Recall} = \frac{(\text{TP})}{(\text{TP} + \text{FN})} \tag{2}$$

$$\text{Precision} = \frac{(\text{TP})}{(\text{TP} + \text{FP})} \tag{3}$$

$$\text{F} - \text{measure} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{4}$$

The confusion matrix of DT, KNN, RF, NB, AB, LR, SVM classifier for cross-validation, and Train/Test splitting is shown in Table 5. The performance measure value of all the classification algorithm used on the PIDD is shown in Table 6. In Table 6, we can see that the accuracy of all classification methods is above 70%. Moreover, LR and SVM both methods are showing better accuracy for both testing methods.

All classifiers' performance based on the different measures with K-fold cross-validation and train/test splitting methods is plotted via a graph in Figs. 8 and 9.

4.2. Results for neural network

In the NN with hidden layer 1, with 200 epochs, we changed the learning rate 0.1, 0.01, 0.005, shown in Table 7. We found that the learning rate at 0.01 provides better accuracy. Hence each case, we used learning rate = 0.01.

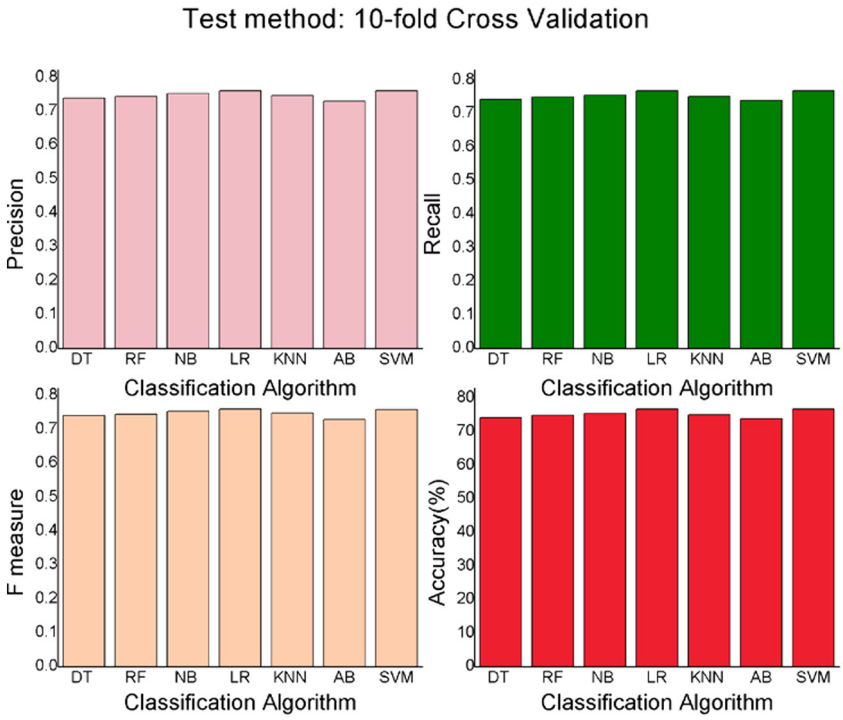

Fig. 8. Graphical presentation of the performance of all classifiers with a 10-fold cross-validation method.

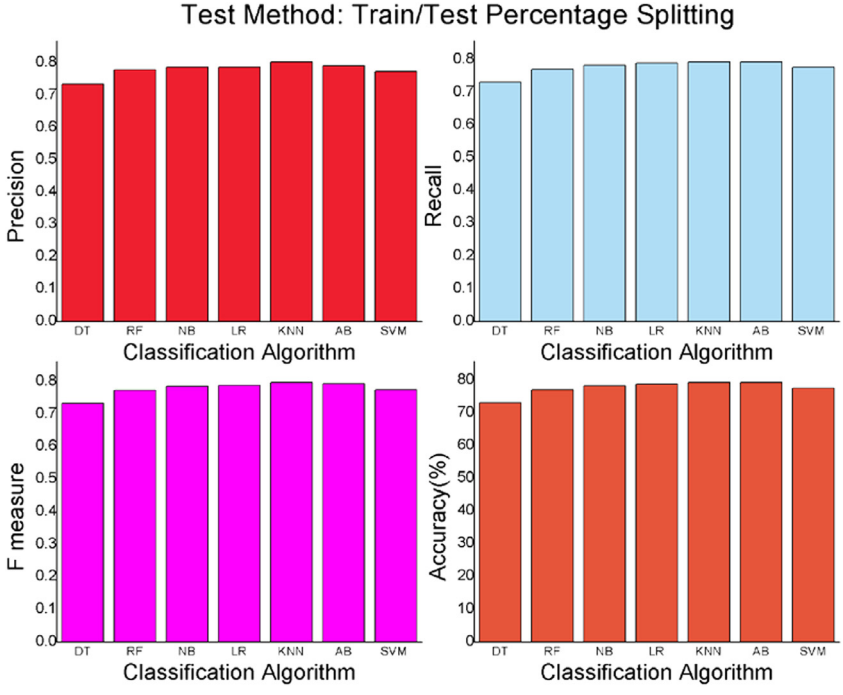

Fig. 9. Graphical presentation of the performance of classifier with train/test splitting method.

Table 8 shows the impact of epoch in a neural network with hidden layers 1, 2, and 3 at the learning rate of 0.01. We found that the NN model with two hidden layers with 400 epochs at a 0.01 learning rate provides the best accuracy of 88.6%. Moreover, the NN model gives more accuracy, Training accuracy, and Testing accuracy among all neural network models. The neural network of our best model NN, which included two hidden layers, is shown in Fig. 10. The ROC curve (receiver operating characteristic curve) for 2 hidden layers with 400 epochs is shown in Fig. 11. With K = 10-fold cross-validation, we also calculated the accuracy for two hidden layers NN model with 200 epochs. The mean accuracy obtained was 76%.

Table 7
Impact of learning rate on accuracy measurement.

Learning rate	Accuracy
0.1	0.829
0.01	**0.838**
0.005	0.800

Table 8
At learning rate 0.01 with hidden layer changes impact on the accuracy.

Hidden layer	Epochs	Accuracy	Training accuracy	Testing accuracy
1	200	0.838	76.43%	83.81%
	400	0.848	77.27%	84.76%
	800	0.829	79.46%	82.86%
2	200	0.876	76.77%	87.62%
	400	**0.886**	78.96%	**88.57%**
	800	0.857	81.65%	87.62%
3	200	0.829	76.77%	82.86%
	400	0.838	83.00%	83.81%
	800	0.790	87.04%	79.05%

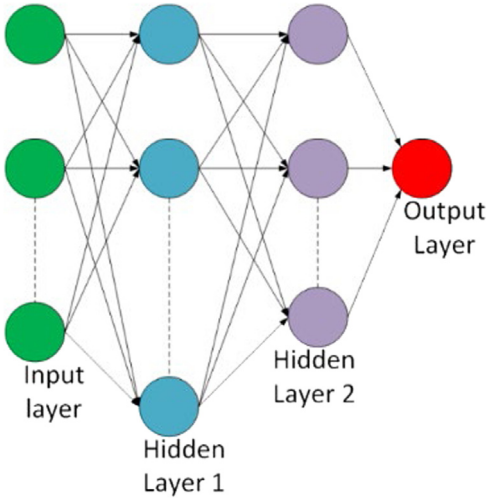

Fig. 10. NN model with two hidden layers.

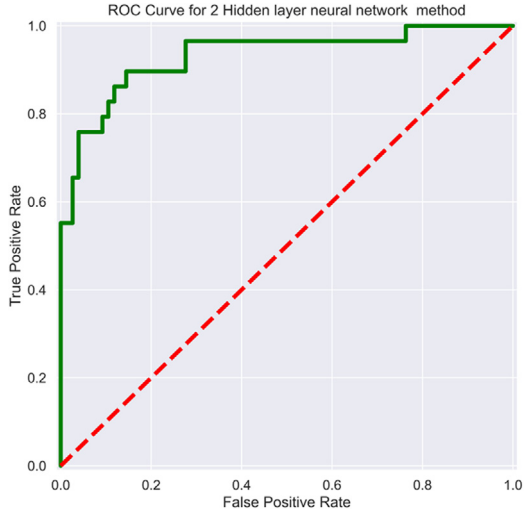

Fig. 11. ROC curve for 2 hidden layer NN with 400 epochs.

Conclusion

Early detection of diabetes is one of the significant challenges in the health care industry. In our research, we designed a system, which can predict diabetes with high accuracy. We preprocessed the data using the WEKA tool. Using the feature reduction method, we dropped three features. We used five input features (Glucose, BMI, Insulin, Pregnancy, and Age) and one output feature (outcome) in the PIMA dataset. We used seven different machine learning algorithms, including DT, KNN, RF, NB, AB, LR, SVM on the PIDD to predict diabetes and evaluated the performance on various measures. All models show good results for some parameters like accuracy, precision, recall, and F-measure. All models provided an accuracy greater than 70%. LR and SVM provided approximately 77%–78% accuracy for both train/test split and K-fold cross-validation method. We also implemented the NN model for diabetic prediction of PIDD. We used the 1, 2, 3 hidden layers in the neural network model varying the epochs 200, 400, 800. Hidden layer 2 with 400 epochs provided 88.6% accuracy, which is the highest accuracy among our implemented model for PIDD. Among all the proposed models, the NN with two hidden layers is considered the most efficient and promising for analyzing diabetes with an accuracy rate of approximately 86% for all varying epochs (200, 400, 800). The accuracy found for logistic regression (78.8571%), Naive Bayes (78.2857%), random forest (77.3429%), and ANN (88.57)% was better than the accuracy of the studies by Tigga et al. [19] (LR ∼75.32%), Sisodia et al. [18] (NB∼76.30%), Amour Diwani et. al [20] (NB∼76.3021%), Zou et al. [21] (RF ∼77.21%), and Alam, T.M.et al. [17] (ANN∼ 75.7%).

CRediT authorship contribution statement

Jobeda Jamal Khanam: Conceptualization, Methodology, Software, Data curation, Writing - original draft, Visualization. **Simon Y. Foo:** Methodology, Supervision, Investigation, Software , Resources, Validation, Writing - review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is funded in part by Florida A&M University and Florida State University, USA.

References

[1] https://www.who.int/health-topics/diabetes.
[2] https://www.medicalnewstoday.com/articles/325018#how-is-the-pancreas-linked-with-diabetes.
[3] https://www.webmd.com/diabetes/diabetes-causes.
[4] https://www.mayoclinic.org/diseases-conditions/prediabetes/diagnosis-treatment/drc-20355284.

[5] https://www.niddk.nih.gov/healthinformation/diabetes/overview/symptoms-causes.
[6] https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html.
[7] https://www.healthgrades.com/right-care/diabetes/is-there-a-cure-for-diabetes.
[8] https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes-long-term-effects.
[9] S.A. Kaveeshwar, J. Cornwall, The current state of diabetes mellitus in India, Australas. Med. J. 7 (1) (2014) 45.
[10] https://www.cdc.gov/diabetes/basics/prediabetes.html.
[11] C.L. Huang, M.C. Chen, C.J. Wang, Credit scoring with a data mining approach based on support vector machines, Expert Syst. Appl. 33 (4) (2007) 847–856, http://dx.doi.org/10.1016/j.eswa.2006.07.007.
[12] J. Chaki, S. Thillai Ganesh, S.K. Cidham, S. Ananda Theertan, Machine learning and artificial intelligence-based diabetes mellitus detection and self-management: A systematic review, J. King Saud Univ. - Comput. Inf. Sci. (2020).
[13] I. Contreras, J. Vehi, Artificial intelligence for diabetes management and decision support: Literature review, J. Med. Internet Res. 20 (5) (2018) e10775.
[14] G. Swapna, R. Vinayakumar, K.P. Soman, Soman KP diabetes detection using deep learning algorithms, ICT Express 4 (4) (2018) 243–246, http://dx.doi.org/10.1016/j.icte.2018.10.005, Elsevier B.V.
[15] M.W. Craven, J.W. Shavlik, Using neural networks for data mining, Future Gener. Comput. Syst. 13 (2–3) (1997) 211–229, http://dx.doi.org/10.1016/s0167-739x(97)00022-8.
[16] https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/.
[17] T.M. Alam, et al., Informatics in medicine unlocked a model for early prediction of diabetes, Inform. Med. Unlocked 16 (2019) 100204.
[18] D. Sisodia, D.S. Sisodia, Prediction of diabetes using classification algorithms, Procedia Comput. Sci. 132 (2018) 1578–1585.
[19] N.P. Tigga, S. Garg, Predicting type 2 Diabetes using Logistic Regression accepted to publish in: Lecture Notes of Electrical Engineering, Springer.
[20] Salim Amour Diwani, Anael Sam, Diabetes forecasting using supervised learning techniques, Adv. Comput. Sci.: Int. J. [S.l.] (ISSN: 2322-5157) (2014) 10–18, Available at: <http://www.acsij.org/acsij/article/view/156>.
[21] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, Predicting Diabetes Mellitus with Machine Learning Techniques, Vol. 9, Frontiers in genetics, 2018, p. 515, http://dx.doi.org/10.3389/fgene.2018.00515.
[22] S. Perveen, M. Shahbaz, A. Guergachi, K. Keshavjee, Performance analysis of data mining classification techniques to predict diabetes, Procedia Comput. Sci. 82 (2016) 115–121.
[23] M. Lichman, Pima Indians diabetes database, ed. Center for machine learning and intelligent systems.: UCI Machine Learning repository.
[24] https://www.cdc.gov/obesity/adult/defining.html.
[25] S.R. Garner, Weka: The Waikato environment for knowledge analysis, in: Proceedings of the New Zealand Computer Science Research Students Conference, Citeseer, 1995, pp. 57–64.
[26] https://en.wikipedia.org/wiki/Project_Jupyter.
[27] H. Benhar, A. Idri, J. Fernández-Alemán, Data preprocessing for decision making in medical informatics: potential and analysis, in: World Conference on Information Systems and Technologies, 2018, pp. 1208–1218.