Diagnosis and Classification of the Diabetes Using Machine Learning Algorithms

Prasannavenkatesan Theerthagiri1 · A. Usha Ruby2 · J. Vidyá1

Abstract
Diabetes mellitus is characterized as a chronic disease that may cause many complications. Machine learning algorithms are used to diagnose and predict diabetes. The learning-based algorithms play a vital role in supporting decision-making in disease diagnosis and prediction. This paper investigates traditional classification algorithms and neural network-based machine learning for the diabetes dataset. Also, various performance methods with different aspects are evaluated for the K-nearest neighbor, Naive Bayes, extra trees, decision trees, radial basis function, and multilayer perceptron algorithms. It supports the estimation of patients who possibly suffer from diabetes in the future. This work shows that the multilayer perceptron algorithm gives the highest prediction accuracy with the lowest MSE of 0.19. The MLP gives the lowest false-positive and false-negative rates with the highest area under the curve of 86%.

Keywords Diabetes prediction · MLP · Machine learning algorithm · Classification of diabetes

Introduction
Diabetes (Diabetes Mellitus—DM) is one of the metabolic disorders with inappropriately raised blood glucose levels. The carbohydrates consumed will be turned into a type of sugar called glucose, and it will be released into the bloodstream. Insulin is a hormone that helps move glucose from the blood to cells. With this chronic condition, the pancreas will produce little or no insulin, and sometimes, the produced insulin will not be absorbed by the cells; this is termed insulin resistance [1].

Presently, diabetes is considered one of the most lethal diseases across the globe, and people are being affected in huge numbers. Around 422 million people are diabetic patients, and about 1.6 million deaths are attributed to diabetes every year. Over the past few decades, the number of cases and the prevalence of diabetes are steadily increasing [2].

DM is classified as type 1, type 2, and gestational diabetes. The condition where the pancreas will produce little or no insulin is type 1 diabetes. If the insulin is not absorbed by the cells or not produced in enough quantity, it is referred to as type 2 diabetes (T2D). During pregnancy, high glucose and sugar levels would increase the risk of complications like hearing loss, dementia, heart diseases, stroke, depression, vision loss, retinopathy, neuropathy, and so on. Early detection plays a prominent role in disease detection. It is one of the crucial causes of cardiovascular diseases, and there is an immense need to support the medical decision-making process. Many researchers in different medical diagnoses have employed various machine-learning techniques [3].

Many researchers are working on medical expert systems, and there has been much contemplation in this field. Medical experts and data analysts collaborate continuously to make this system more accurate and, thus, useful in real life. Recent surveys by the World Health Organization indicated a tremendous increase in diabetic patients and the demise attributed to diabetes every year. Therefore, early diagnosis of diabetes is a significant concern among researchers and medical practitioners [4]. Multitudinous computer-based detection systems were designed and outlined for analyzing and anticipating diabetes. The usual identifying process for
diabetes takes time. Nevertheless, with the rise of machine learning, we can develop a solution to this intense issue [5].

To accurately predict the disorder, a good model that can represent the presence of diabetes through input characteristics is required. With a good model and an accurate detection technique, diagnosis can be made more efficient. Based on the prediction, medical practitioners can envision biomedical diagnosis using engineering tools that can automatically adapt to unexpected future conditions. A long-term prediction algorithm can play a vital role in planning and provisioning. Intelligence systems can learn or adapt and modify functional dependencies in response to new experiences or changes in functional relationships [6]. The contributions of this paper are

- A systematic extensive literature was studied concentrating on AI techniques and applications for diabetes encompassing various traditional machine-learning methods and statistical models.
- This paper explicitly researches the most recent developments in deep learning technologies for the treatment of diabetes in this study.
- Machine learning-based model has been demonstrated competitive performance in a number of other significant works on diabetes.

**Literature Survey**

Adel Al-Zebari et al. have compared the performance of various machine-learning algorithms for diabetes detection. MATLAB classification learner tool has been used in this work, including decision tree, discriminant analysis, SVM (Support Vector Machine), k-NN (k-Nearest Neighbor), Logistic regression, and ensemble learners, and their variants with 26 classifiers are considered. The results are evaluated on a tenfold cross-validation basis and average classification accuracy is considered for performance measures. It has been reported that average accuracy with the range 65.5% and 77.9%, for Coarse Gaussian SVM and LR algorithms, respectively [7].

G. A. Pethunachiyar used SVM with disparate kernel functions for the classification of diabetes. The simulation model of the proposed system includes five phases. After collecting the data, the selection process is carried out by rectifying the errors (inconsistency in data or missing values or wrong information). The data will be divided into training (70%) and testing dataset (30%). The SVM technique has been selected for efficient prediction, and a model has been built. Test data are applied to the model to make the prediction. The linear, polynomial, and radial kernel-based SVM has been implemented in this work. The confusion matrix is used for calculating prediction accuracy. ROC (Receiver-Operating Characteristic curve) is used to evaluate these three kernel functions. Linear kernel with SVM predicts more accurately compared to other kernels [8].

Pahulpreet Singh Kohli et al. have applied various machine-learning techniques on three different disease datasets for disease prediction. Feature selection is carried out by backward modeling using the p-value test. The proposed model includes four phases: initially, the dataset is explored in a Python environment. During data munging, the missing values are replaced with mean and mode values for the continuous and categorical variables, respectively. The features are selected very cautiously to improve the performance of the model. The attributes are eliminated using the backward selection method (based on p-value, it is eliminated). After selecting the features, the model is refitted. Five algorithms were compared, including decision tree, logistic regression, random forest, adaptive boosting, and support vector machine. The dataset has been divided into a training set (90%) and a test dataset (10%). In future data munging, the selection of features and model fitting steps can be automated; a pipeline structure for preprocessing data would improve results. This work reported with the accuracy of 85.71 by SVM with linear kernel [9].

Samrat Kumar Dey et al. have developed a web application using Tensorflow to predict diabetes successfully. This proposed model requires patient data for successful diagnosis, and the techniques like SVM, ANN (Artificial Neural Network), KNN, and Naive Bayes are used to predict the disease. The dataset is divided into two parts: training and testing dataset. Pre-processing of data and data normalization would increase the accuracy of the model. Min Max Normalization model is used to improve accuracy. Thus, the ANN with Min Max Normalization model produced 82.35% accuracy [5].

Sidong Wei et al. have comprehensively explored DNN (Deep Neural Network), Logistic regression, SVC (Support Vector Classifier), Naive Bayes, and Decision tree techniques to identify diabetes. This work has been carried out in four steps: initially, the best preprocessor is identified for the classifier. Then, the parameters are optimized. In the third step, these techniques are compared by accuracy, and later relevance of these features is considered. The features like plasma glucose concentration, age, and the number of times pregnant were found to be more significant. This technique performs well with the accuracy of 77.86%. The three important features based on these results are number of times pregnant, age, and plasma glucose concentration [10].

S. Hari Krishnan et al. used machine-learning techniques to measure blood glucose level. A photoplethysmograph (PPG)-based system is used to determine the glucose parameters, using light sources of three different wavelengths. The light is illuminated on the skin at the wrist, and the reflected light is captured by a photodiode.
receiver. The same is conditioned, digitalized, and sent to the Arduino UNO microcontroller. The PPG signal is derived by the microcontroller in accordance with the blood glucose values. The waveform is preprocessed and segmented to obtain the peak of the signal. To obtain the statistical features like mean, skewness, variance, kurtosis, standard deviation, and entropy, the random forest technique is implemented on the acquired signal. The model is designed and trained to estimate blood glucose from extracted features. The future would focus on estimating the correlation of the feature sets with different machine-learning techniques [11].

M. Shanthi et al. proposed and developed a model for diagnosing T2D using the ELM (Extreme Learning Machine) technique. The ELM mathematical model has one hidden layer feed-forward network, creating random hidden nodes. Parameters are randomly generated for the hidden nodes initially. The next output matrix is calculated, and then, the network’s optimal weight is given as the output. The output is obtained from the characteristics, input weight, and activation functions. The activation functions are a triangular basis, sine, hard-limit, and sigmoid. This model assists medical experts in forecasting T2D. The hidden neuron with the sigmoid function gives a good accuracy of 84.15% [6].

Sajratul Yakin Rubaiat et al. introduced an approach to predict type 2 diabetes using a neural network. This analysis is carried out in two methods: The first method involves data recovery followed by the selection of features. The selected features are inputted into MLP (multilayer perceptron) neural network classifier. The second approach uses the K-means algorithm. The neural network-based method involves three steps such as data recovery (missing data are replaced with mean value to complete the dataset), selection of features (features with more impact on risk factor identification are selected), and Multilayer Perceptron Classifier (hyperparameters are selected). K-means reduces noise very effectively, and its output has been used as a feature for the model. The model can be trained using these two methods and predict whether a person has diabetes at an early stage. The first method is more efficient and requires less computation compared to k-means. Thus, MLP neural network classifier gives an accuracy of 85.15% and the noise reduction with K-Means produces accuracy of 77.08% [12].

Maham Jahangir et al. presented a novel prediction framework that uses AutoMLP (automatic multilayer perceptron) combined with an outlier detection method. This method involves two stages: preprocessing data with outlier detection, followed by training of AutoMLP. In the second stage, it is used to classify the data. Compared to the other architectures of the neural network, AutoMLP gives higher accuracy. The attributes like plasma glucose level, blood pressure, and the number of times pregnant are found to be more relevant. This work had produced an accuracy of 88.7% [13].

Ali Mohebbi et al. used CGM (Continuous Glucose Monitoring) signals for adherence detection in diabetic patients. A considerable amount of signals were simulated using a T2D-adapted version of the MVP (Medtronic Virtual Patient) model. Different classification algorithms were compared using a comprehensive grid search. Logistic regression, Convolutional Neural Network (CNN), and Multi-Layer Perceptron techniques have been used in this work. CNN shows better performance in classification with the accuracy of 77.5% [4].

The majority of works employ accuracy as an integrity metric, which is based on the overall number of accurate forecasts, both diabetic and non-diabetic. The network weights in the supervised learning can be adjusted further to optimize performance for certain learning tasks. In the extreme situation of an imbalance dataset, even if all of the samples in the minor class are incorrectly predicted, a high accuracy may still be attained due to the percentage of the main class as long as the majority of the samples in the major class are properly predicted. A proposed method considers the issues of missing data and class imbalance, and has been provided a solution to the difficulties. Table 1 summarizes the key aspects of literature study with methodology used and performance results.

## Methodology

### Data Pre-processing and Cleaning

The diabetes dataset from the Pima Indians Diabetes (PID) Database [14] is taken for the predictive analysis in this research work. The considered dataset was cleaned using the data preprocessing and data cleaning methodologies, and then, the resulting dataset was considered for several experiments over different classification algorithms. The Pima Indians Diabetes Database contains the patient’s details with diabetes status (Non-Diabetes and Diabetes). The vital patient’s information is used to diagnose and predict Diabetes Mellitus among the population.

The considered diabetes mellitus dataset contains 768 records. The dataset contains features of patients such as (1) number of times pregnant, (2) plasma glucose concentration a 2 h in an oral glucose tolerance test, (3) diastolic blood pressure in mmHg, (4) triceps skinfold thickness in mm, (5) 2-h serum insulin in µU/ml, (6) body mass index (weight in kg/height in m)², (7) diabetes pedigree function, (8) age in years, and (9) class variable (0 or 1).

The data preprocessing and cleaning process (data imputation–mean technique) removes the missing and outlier data values from the dataset. After preprocessing, the resulting dataset is reduced to 722 records with three required relevant features of patient details. There are 722
patient details in the dataset, out of which 474 cases are in the class of ‘Non-Diabetes’, and 248 cases are in the class of ‘Diabetes’ with 46 records missing required essential values. Six numerical features from the dataset are taken as the input attributes, and one feature is considered as the output attribute. The patient's information is presented in Table 2.

The patient features, such as the number of times pregnant, plasma glucose concentration, diastolic blood pressure, body mass index (BMI), diabetes pedigree function (DPF),

| **Table 1** Summary of literature study |
| Authors | Objectives | Methodology | Main outcomes |
| --- | --- | --- | --- |
| Adel Al-Zebari et al. [7] | To study and compare different machine-learning methods for prediction of diabetes | MCLT (MATLAB Classification Learners Tool). They have used DT, SVM, Linear discriminant analysis (LDA), ensemble learners, and LR | Accuracy: DT: 65–75.3%, LDA: 74.1–77.5% LR: 77.9% SVM: 65.5–77.6%, KNN: 72.8–76%, Ensemble Learners: 71.6–76.6% |
| Pahulpreet Singh et al. [9] | To detect the diabetes at an early stage | Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, AdaBoost | Accuracy: LR: 84.42%, DT: 74.03%, RF: 81.82% SVM with linear kernel: 85.71%, SVM with RBF: 66.23%, AdaBoost: 80.52% |
| Samrat Kumar et al. [5] | To develop a web application for predicting the diabetes disease | SVM, KNN, Gaussian Naïve Bayes (GNB), Artificial Neural Network with Min Max Scaler | Accuracy: SVM: 78% KNN: 75.3% GNB: 79.1% ANN: 81.8% |
| Sidong et al. [10] | To early identify and predict diabetes | Deep Neural Network (DNN), SVM, LR, DT, Naïve Bayes | Accuracy: DNN: 77.86%, SVM: 77.6%, LR: 77.47% DT: 76.3%, Naïve Bayes: 75.78% |
| Shanthi et al. [6] | Diagnosis of type 2 diabetes | Extreme Learning Machine (ELM) | Accuracy: Sigmoid function: 84.15% Sine function: 83.6% Hard limit: 64.67% Triangular basis: 77.93% |
| Sajratul et al. [12] | To find a best model that predicts the diabetes (Type 2) | Feature selection Noise reduction with feature selection | Accuracy: K-Means: 77% MLP with feature selection: 85.153% |
| Maham et al. [13] | To predict the diabetes | MLP | Accuracy—88.7% Weighted Mean Recall—88.5 Weighted Mean Precision—85.8 |
| Ali Mohebbi et al. [4] | To detect type 2 diabetes | LR, MLP, CNN | LR–65.2% MLP–72.5% CNN–77.5% |

| **Table 2** Sample record of cleaned dataset |
| --- | --- | --- | --- | --- | --- | --- |
| #Pregnant | Glucose | BP | BMI | DPF | Age | Class |
| --- | --- | --- | --- | --- | --- | --- |
| 8 | 183 | 64 | 23.3 | 0.672 | 32 | Diabetes |
| 1 | 89 | 66 | 28.1 | 0.167 | 21 | Non-diabetes |
| 0 | 137 | 40 | 43.1 | 2.288 | 33 | Diabetes |
| 5 | 116 | 74 | 25.6 | 0.201 | 30 | Non-diabetes |
| 3 | 78 | 50 | 31 | 0.248 | 26 | Diabetes |
| 2 | 197 | 70 | 30.5 | 0.158 | 53 | Diabetes |
| 4 | 110 | 92 | 37.6 | 0.191 | 30 | Non-diabetes |
| 10 | 168 | 74 | 38 | 0.537 | 34 | Diabetes |
| 9 | 129 | 80 | 27.1 | 1.441 | 57 | Non-diabetes |
| 1 | 189 | 60 | 30.1 | 0.398 | 59 | Diabetes |
and age, are considered input variables, and the class is taken as the output variable. Figure 1 illustrates the population with respect to age. Figures 2 and 3 depict diabetes/non-diabetes populations with respect to BMI and BP.

This research analyzes the prediction of diabetes in non-diabetes patients using a different machine-learning algorithm. Different classification models are applied to the diabetes dataset, and its performance in terms of accuracy, error rates, and area curves are evaluated. This work includes evaluating KNN, Naive Bayes, extra trees, decision trees, radial basis function, and multilayer perceptron algorithms.

**Machine Learning Algorithms**

The KNN is one of the most straightforward supervised machine-learning algorithms used to solve regression and classification problems. It assumes the similarity between the new and available data and assigns the new data to the category that is most similar to the available categories. The distance between data points is calculated using Euclidean distance; the distance between two points \((X_1, Y_1)\) and \((X_2, Y_2)\) is \(\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}\), this gives the nearest neighbor [15].

Naive Bayes is one of the popular classification algorithms that are most widely used to get the base accuracy of the dataset. It assumes that all the variables present in the dataset are Naive (not correlated to each other). They are used in real-time prediction, multi-class prediction, spam filtering, sentimental analysis, text classification, recommendation system, etc. Bayes’ rule determines the probability of the hypothesis. The formula used is: \(P(A|B) = \frac{P(B|A)P(A)}{P(B)}\), where \(P(A|B)\) refers to Posterior probability, \(P(A)\) refers to Prior probability, \(P(B)\) is the marginal probability, and \(P(B|A)\) refers to Likelihood probability [16].

An extra tree (Extremely Randomized tree) classifier is an ensemble learning method based on decision trees. It will work by creating many unpruned decision trees from the training dataset. In the case of classification, prediction is carried out using majority voting, and in the case of regression, to make a prediction, the prediction of decision trees [17].

The decision tree is a supervised machine-learning technique that splits the data based on a parameter. The tree contains two entities, namely leaves and decision nodes. The leaves are the final outcomes, and the data are split into the decision nodes. The main issue is the selection of the best attribute as the root node and sub-nodes. Information gain and Gini index techniques can be used for attribute selection. Information gain is calculated using as follows: Information gain = Entropy(S) − [[(Weighted Avg) * Entropy(each feature)]], the entropy metric is used to measure the impurity in the attribute. Entropy is calculated using this formula, Entropy(S) = −\(P(\text{yes})\log_2P(\text{yes}) - P(\text{no})\log_2P(\text{no})\).; \(S\) represents the number of samples, probability of yes and no are represented.

---

Fig. 1 Population vs age

Fig. 2 BMI vs diabetes classes

Fig. 3 BP vs diabetes classes
as \( P(\text{yes}) \) and \( P(\text{no}) \), respectively. It tells us the amount of information a feature provides about the class; the decision tree can be built with this information. The Gini index is calculated as follows: \[ \text{GiniIndex} = 1 - \sum_{j=1}^{N} P_j^2 \] It is the measure of purity or impurity used for decision tree creation in the Classification and Regression Tree (CART) algorithm. An attribute with a low Gini index is preferred [18].

A radial basis function will assign a real value to every input from its domain, and the outcome will be an absolute value and cannot be negative (it is a measure of distance) \( f(x) = f(||x||) \). Mainly, it is used to approximate the functions. The sum \( y(x) = \sum_{i=1}^{N} w_i \phi(||x - x_i||) \) represents radial basis function. These functions act as activation functions [19].

Multi-layered perceptron is a simple, commonly used neural network model, referred to as "vanilla" neural networks. It can be used for various applications like spam detection, image identification, election voting predictions, and stock analysis [20].

Results and Discussion

This section summarizes the prediction results of the KNN, Naive Bayes, extra trees, decision trees, radial basis function, and multilayer perceptron algorithms. The k-fold cross-validation is one of the resampling procedures used to validate the machine-learning models on the limited data sample. In this work, the \( k \) value is chosen as 7. Therefore, it can be called a sevenfold cross-validation resampling method. The sevenfold cross-validation method intends to reduce the bias of the prediction model [21].

Performance Evaluation

Typically, the performance of the machine-learning prediction algorithms is measured using some metrics based on the classification algorithm. In this work, the prediction results are evaluated using the metrics such as accuracy, mean square error (MSE), root-mean-square error (RMSE), Kappa score, confusion matrix, the receiver-operating characteristic-area under curve (ROC_AUC), classification performance indices, sensitivity, specificity, and f1 score values [21–23].

In this work, the prediction accuracy (that is, whether the patient is diabetic or non-diabetic) of different machine algorithms (KNN, Naive Bayes, extra trees, decision trees, radial basis function, and multilayer perceptron algorithms) is determined. Each classification model has a different prediction accuracy based on its hyperparameters and a certain level of improvement over other prediction models. This work considers 70% dataset for training and 30% of the data samples for testing classification algorithms. In this work, each model’s accuracy is compared, and its prediction results are summarized in Table 3.

In Table 3, the classification algorithms, such as KNN, Naive Bayes, extra trees, decision trees, radial basis function, and multilayer perceptron algorithms, have prediction accuracy of 71.7241, 66.8965, 77.2413, 72.4137, 68.2758, and 80.6890, respectively. In contrast, the multilayer perceptron algorithm predicts the diabetes cases (based on the number of pregnant, glucose level, BP, BMI, DPF, age, and diabetes class) more accurately than the other algorithms.

The multilayer perceptron is the feed-forward artificial neural network. The MLP processes the diabetes dataset using non-linear activation functions with three layers of neurons: input, hidden, and output layer. The dataset was tested with several neuron values and classified the given dataset into two forms of classes as diabetes and non-diabetes patients with reduced errors. It works by mapping the given weighted inputs to the output of each neuron among the test data and training data. The testing datasets are classified for each data point based on the error values. That the multilayer perceptron algorithm produces a higher classification rate than the other algorithms.

Cohen's kappa score for the multilayer perceptron algorithm is also higher than other algorithms. Cohen's kappa score estimates the consistency of the classification algorithm based on its predictions. Figure 4 depicts the accuracy scores of different classification algorithms. From Fig. 4, we can clearly see that the multilayer perceptron algorithm has the highest accuracy of 80.68. The multilayer perceptron algorithm has 3.4 to 13.8% improved accuracy compared to KNN, DT, NB, ET, and RBF algorithms. The MLP algorithm works by classifying the data point of the diabetes dataset based on similarity.

Table 4 presents the performance error metrics of the various machine-learning algorithms. The error metrics mean square error and root-mean-square error values for each algorithm are evaluated. The KNN, Naive Bayes, extra trees, decision trees, radial basis function, and multilayer perceptron algorithms have the MSE error rate as 0.2827,

| S. no | Classifier          | Accuracy  | Kappa  | AU_ROC  |
|-------|---------------------|-----------|--------|---------|
| 1.    | K-Neighbors Classifier (KNN) | 71.7241   | 0.3973 | 0.77    |
| 2.    | Decision Tree (DT)   | 66.8965   | 0.3125 | 0.64    |
| 3.    | Naive Bayes          | 77.2413   | 0.5112 | 0.85    |
| 4.    | Extra Trees          | 72.4137   | 0.3770 | 0.70    |
| 5.    | Radial Basis Function| 68.2758   | 0.3213 | 0.70    |
| 6.    | Multi-Layer Perceptron (MLP) | 80.6890   | 0.5740 | 0.86    |
0.3310, 0.2275, 0.2758, 0.3172, and 0.1931, respectively. As per Fig. 5, the multilayer perceptron algorithm produces the lowest error rate, 0.1931 on the prediction of accurate diabetes cases than the other algorithm. The multilayer perceptron algorithm classifies the testing dataset by output mapping between the new (testing) instance \( x_i \) and the existing (training) instance \( y_j \). Therefore, it results in lower error rates.

Similarly, the MLP’s RMSE error rate is also very low (0.43) as compared to the error rates of KNN (0.53), DT (0.57), NB (0.47), ET (0.52), and RBF (0.56) algorithms, as shown in Table 3. As depicted in Fig. 6, the multilayer perceptron algorithm produces the highest consistency (Kappa score) among the evaluated algorithms as 0.57. The KNN, Naive Bayes, extra trees, decision trees, and radial basis function have 0.3973, 0.3125, 0.5112, 0.3770, and 0.3213, respectively.

Figure 7 illustrates the confusion matrix without normalization of the multilayer perceptron algorithm. In all classification algorithms, 30% of the data samples are taken for testing with the 70% training dataset. In this figure, the x-axis represents the percentage of predicted values, and the y-axis represents the percentage of true values. It can be seen that the multilayer perceptron algorithm predicts 79% (true positive) of the diabetes cases correctly, with 12% (false positive) of misclassification.
Similarly, in Fig. 8, the confusion matrix with normalization is depicted.

**Comparative Results’ Analysis**

Figure 9 is the pictorial representation between the false-positive rate and true-positive rate in the form ROC area under the curve. The multilayer perceptron algorithm produces the highest value of 0.86 compared to KNN, Naive Bayes, extra trees, decision trees, and radial basis function algorithms. Figure 10 summarizes the performance metrics such as precision, recall, and confusion matrix of the multilayer perceptron algorithm.

The multilayer perceptron algorithm produces the precision (true-positive rate) value of 0.82 for the non-diabetes cases and 0.78 for the deceased cases. The recall values for the non-diabetes and diabetes cases are 0.89 and 0.67, respectively. Further, the F1 score for non-diabetes and diabetes cases is 0.85 and 0.75, respectively.

The results of the proposed work have been critically compared with different datasets. Table 5 summarizes the comparative analysis with the datasets in the papers [24–27], and [28]. The parameters, such as models used, data source, training process, performance metrics, and accuracy, are compared with the proposed models’ results [29]. The work in [24] uses LSTM with an attention pooling layer for the 7191 patients and gives 75.9% accuracy and 66.2% precision. The work in [25] employs the Restricted Boltzmann Machine (RBM) and RNN for the PID dataset and gives 90.6% of sensitivity and 75% of precision. Similarly, DeepMLP was developed in [26] for the dataset with 4814 participants and produced an AUC of 0.703, with an accuracy of 67.9%. Thus, the comparative study clearly shows that the proposed model is the best, with 86% of AOC and 80.7% of accuracy.

**Conclusion**

It is worth studying much essential to accurately predicting and diagnosing any disease using machine learning. This work explores different machine-learning algorithms and their performances on the diabetes dataset. The results of machine-learning algorithms KNN, Naive Bayes, extra trees, decision trees, and radial basis function are analyzed in this study. All the above algorithms have experimented with prediction accuracy, MSE, RMSE, Kappa score, AUROC, precision, recall, and F1-score. The results show that MLP performs better with the Prima diabetes dataset. In addition, comparing with the results of other classification algorithms, we can see that the MLP has better AUROC at 86% among

**Fig. 8** Confusion matrix (no normalization)

**Fig. 9** ROC_AUC curve

**Fig. 10** Summary of performance metrics scores of MLP algorithm
KNN, Naive Bayes, extra trees, decision trees, and radial basis function. Thus, the results suggest that the MLP algorithm can able to diagnose and classify diabetic patients.

Funding There is no funding source.

Data availability Data available on request from the authors.

Declarations

Conflict of Interest Not applicable.

Research Involving Human Participants and/or Animals Not applicable.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent Not applicable.

References

1. Lebovitz HE. Type 2 diabetes: an overview. Clin Chem. 1999;45(8):1339–45.
2. Mokdad AH, Ford ES, Bowman BA, Nelson DE, Engelgau MM, Vinicor F, Marks JS. Diabetes trends in the US: 1990–1998. Diabetes Care. 2000;23(9):1278–83.
3. Kalyani RR, Corriere MD, Donner TW, Quartuccio MW. Diabetes head to toe: everything you need to know about diagnosis, treatment, and living with diabetes. Baltimore: Johns Hopkins University Press; 2018.
4. Ali M, Tinna BA, Alexander RJ. A deep learning approach to adherence detection for type 2 diabetics. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017.
5. Samrat KD, Ashraf H, Md. Mahbubur R. Implementation of a Web application to predict diabetes disease: an approach using machine-learning algorithm. In: 2018 21st International Conference of Computer and Information Technology (ICCIT), 2018, pp. 21–3.
6. Shanthi M, Ramalatha M, Shivapriya SN, Navaneethakrishnan R. Diagnosis of diabetes using an extreme learning machine algorithm based model. In: 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), 2019.
7. Adel A-Z, Abdulkadir S. Performance comparison of machine-learning techniques on diabetes disease detection. In: 2019 1st International Informatics and Software Engineering Conference (UBMYK); 2019.
8. Pethunachiyar GA. Classification of diabetes patients using kernel based support vector machines. In: 2020 International Conference on Computer Communication and Informatics (ICCCI - 2020), Coimbatore, INDIA, 2020, pp. 22–4.
9. Pahulpreet SK, Shriya A. Application of machine learning in disease prediction. In: 2018-4th International Conference on Computing Communication and Automation (ICCCA); 2018.
10. Sidong W, Xuejiao Z, Chunyan M. A comprehensive exploration to the machine-learning techniques for diabetes identification. In: 2018 IEEE 4th World Forum on Internet of Things (WF-IoT); 2018.
11. Hari-Krishnan S, Vinupritha P, Kathirvelu D. Non-invasive glucose monitoring using machine-learning. In: International Conference on Communication and Signal Processing, 2020, pp. 28–30.
12. Sajratul YR, Md Monibor R, Md Kamrul H. Important feature selection and accuracy comparisons of different machine-learning models for early diabetes detection. In: 2018 International Conference on Innovation in Engineering and Technology (ICIET); 2018.
13. Maham J, Hammad A, Mehreen A, Khawar K, Raheel N. An expert system for diabetes prediction using autotuned multi-layer perceptron. In: Intelligent Systems Conference 2017-8, London, UK, 2017.
14. Kaggle repository: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Accessed 15 Jan 2020.
15. Saxena R. Role of K-nearest neighbour in detection of Diabetes Mellitus. Turk J Comput Math Educ (TURCOMAT). 2021;12(10):373–6.

16. Karim M, Rahman RM. Decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing. J Softw Eng Appl. 2013. https://doi.org/10.4236/jssea.2013.64025.

17. Brownlee J. How to develop an extra trees ensemble with python. Mach Learn Mastery. 2020.***

18. Kulkarni M. Decision trees for classification: a machine-learning algorithm. Xoriant. 2017. ***

19. Prasannavenkatesan T, Vidya J. Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques. In: Expert systems. Cham: Springer; 2022. https://doi.org/10.1111/essy.13064.

20. Plunkett K, Marchman V. U-shaped learning and frequency effects in a multilayered perceptron: implications for child language acquisition. In: Connectionist psychology: A text with readings. Cham: Springer; 2020. p. 487–526.

21. Prasannavenkatesan T. Probable forecasting of epidemic COVID-19 in using COCUDE model. EAI Endors Trans Pervasive Health Technol. 2021;7(26): e3. https://doi.org/10.4108/eai.3-2-2021.168601.

22. Prasannavenkatesan T, Gopala Krishnan C, Nishan AH. Prognostic analysis of hyponatremia for diseased patients using multilayer perceptron classification technique. EAI Endors Trans Pervasive Health Technol. 2021;7(26): e5. https://doi.org/10.4108/eai.17-3-2021.169032.

23. Theerthagiri P. Forecasting hyponatremia in hospitalized patients using multilayer perceptron and multivariate linear regression techniques. In: Concurrency and computation: practice and experience. Berlin: Springer; 2021. p. e6248.

24. Pham T, Tran T, Phung D, Venkatesh S. Predicting healthcare trajectories from medical records: a deep learning approach. J Biomed Inform. 2017;69:218–29.

25. Ramesh S, Balaji H, Iyengar NCS, Caytiles RD. Optimal pre-dictive analytics of Pima diabetics using deep learning. Int J Database Theory Appl. 2017;10(9):47–62.

26. Spänig S, Emberger-Klein A, Sowa J-P, Canbay A, Menrad K, Heider D. The virtual doctor: an interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes”. Artif Intell Med. 2019;100:101706.

27. Nguyen BP, et al. Predicting the onset of type 2 Diabetes using wide and deep learning with electronic health records. Comput Methods Prog Biomed. 2019;182:105055.

28. Ryu KS, Lee SW, Batbaatar E, Lee JW, Choi KS, Cha HS. A deep learning model for estimation of patients with undiagnosed diabetes. Appl Sci. 2020;10(1):421.

29. Prasannavenkatesan T. Stress emotion recognition with discrepancy reduction using transfer learning. In: Multimedia tools and applications. Cham: Springer; 2022. https://doi.org/10.1007/s11042-022-13593-6.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.