Prediction Of Early-Stage Diabetes using machine learning

Abdulaziz Y.I. Abushawish
Department of Computer Engineering
University of Sharjah
Sharjah, United Arab Emirates
u22103344@sharjah.ac.ae

Ali Bou Nassif
Department of Computer Engineering
University of Sharjah
Sharjah, United Arab Emirates
anassif@sharjah.ac.ae

2023 Advances in Science and Engineering Technology International Conferences (ASET) | 978-1-6654-5474-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASET56582.2023.10180804

Abstract— **Diabetes is one of the most chronic diseases for many years. It occurs when the pancreas loses its functionality, where the production of insulin will no longer be sufficient to move the blood glucose to the cells, or when the cells stop reacting with the produced insulin. People that are overweight, living a bad lifestyle, and are lazy and inactive are more likely to be diabetic. Diabetes has multiple symptoms, and the most common symptom is Polyuria which is the production of a large amount of urine. Moreover, other symptoms can appear such as Polydipsia, sudden weight loss, and visual blurring. In Healthcare industries, Data-driven decision-making (DDDM) has been incredibly used for predicting diseases due to its high accuracy and low cost, it will reduce the health risk of the patients by obtaining its symptoms. Machine learning has been implemented significantly in predicting diseases, especially the chronic disease. In this existing work, the utilized machine learning models namely are, a Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), and Radial-Basis Neural Network (RBF) to predict diabetes using Polyuria, Polydipsia, and sudden weight loss from the diabetes risk prediction dataset. K-Nearest Neighbors (KNN) shows better performance compared to the other classifiers with f1-score of 0.98%.**

Keywords — diabetes prediction, healthcare, KNN, SVM, Machine learning.

I. Introduction

Diabetes is one of the most chronic and fatal diseases nowadays. Sudden weight loss, visual blurring, obesity, and Polyuria are some of the diabetes symptoms. There are two main types of diabetes, type 1, and type 2 diabetes. Those types are called insulin-dependent diabetes, where the insulin controls the level of sugar in the blood. According to [1], it shows that there are 422 million people that are diabetic and 1.6 million die from it. The individual's lifestyle, being overweight, and bad eating habits are some of the possibilities that will lead to being diabetic.

As artificial intelligence and machine learning fields are prospering and developing, it has become popularly used in the medical field [2], especially in the process of early chronic disease prediction, such as predicting lung cancer, breast cancer, and diabetes [3]. The early-stage disease prediction influences the mortality from chronic disease, where the statistics show that with the interference of machine learning in the field of healthcare the number of deaths has been reduced due to the prior knowledge along with treating the disease before the outbreak.

In this work, Multilayer Perceptron (MLP), Radial-Basis Neural Network (RBF), Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) have been utilized in applications in the field of healthcare. The application is Early-Stage Diabetes Prediction. The prediction work is organized into six sections as follows; Section II provides us with the related work in the field of disease prediction. Section III elaborates the machine learning classifiers that have been used, followed by the methodologies in section IV. Section V illustrates the results of the models and discussions. Finally, the conclusion is in section VI.

II. Related work

In the past two decades, researchers have devoted their work for implement machine learning in the field of healthcare, trying to predict chronic disease in the early stage. This section is a demonstration of some of the previous related work. In this research [4], using the PIMA Indian dataset with 8 attributes and 768 instances, six types of machine learning algorithms have been used to predict diabetes which are KNN, SVM, RF, NB, LR, and DT. KNN and SVM showed the best results for diabetes prediction with an accuracy of 77%. While on another research paper [5], machine learning tree classifiers have been implemented to predict diabetes, such as Logistic model tree (LMT) and Random tree. The logistic model tree (LMT) showed higher accuracy with 79.31%. In this research [6], a Support Vector Machine (SVM), C4.5 Decision Tree (DT), K-Nearest Neighbors (KNN), and Naive Bayes (NB) have been employed to predict diabetes. As the result of the experimentation, C4.5 Decision Tree had the highest accuracy among the four classifiers of 73.5%. Authors in [7] proposed a web application for the prediction of diabetes, using a Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), and Naive Bayes Algorithm. ANN provided the highest accuracy of 82.35%. [8] proposed a system to predict diabetes in an unexplored region before, in Saudi Arabia. The authors used Multilayer perceptron (MLP), Support Vector Machine (SVM), Naïve Bayes (NB), and K-Nearest Neighbors (KNN), where the accuracy of these models is 77.694%, 63.910%, 69.6742%, and 62.657% respectively. MLP was superior to other classifiers with 77.694% accuracy.

In [7], the authors used Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NB), and Random Forest (RF) for type 2 diabetes prediction. Random Forest had the highest accuracy of 94.10%. Authors in [9] proposed a system for early-stage diabetes prediction with the use of the Pima Indian Diabetes Database (PIDD) by applying Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF) with K-fold cross-validation. The RF in this proposed work showed the highest accuracy with 87.66% while SVM got 80.85%, KNN got 79.24%, and 76.86% for LDA.

Using the same dataset in [10], the authors proposed a deep learning model to predict diabetes. ANN, NB, DT, and DL have been utilized for the prediction. This paper showed that the DL had better accuracy with 98.07% using accuracy, precision, recall, F-measure, specificity, and sensitivity for comparison.

III. Technical Background

This section provides an additional level of information and details of the machine learning classifiers that have been employed.

A. *Multilayer Perceptron (MLP).*

The multilayer perceptron is a feed-forward neural network. it was extensively used in various problems. The structure of the MLP was inspired by the human brain, where MLP contains 3 layers or more, an Input layer, a hidden layer, and an output layer, with at least one neuron on each layer. The activation function is the heart of input to output transformation, where it is chosen based on the needs of the problem. MLP is also a feedforward artificial neural network (ANN), with fully connected neurons as shown in Fig. 1, where each synaptic has its own weight. It is an adjustable weight, which can be updated to reduce the error.

B. *Radial-Basis Neural Network (RBF).*

One of the artificial neural network types, where is commonly used in approximation problems. Despite its similarity with the Multilayer perceptron, it is a unique NN due to its learning speed. MLP structure can consist of one hidden layer or more, while on the other hand, Radial-Basis Neural Network has a fixed number of layers. Input, output, and single hidden layer as shown in Fig. 2. As in MLP, the nodes in the RBF are fully connected, and the weights are adjustable. Usually, the Gaussian function is used as the basis function [11].

C. *Support Vector Machine (SVM).*

A supervised machine learning model, that uses hyperplane to separate the data groups. The hyperplane is simply a straight line in the two dimensions, which is also called a decision boundary. Fig. 3 elaborates on the concept of the SVM. As discussed, the hyperplane will be in the best position when it maximizes the margin between the two groups.

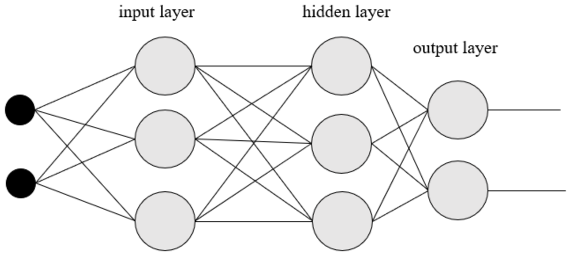

Fig. 1 MLP structure [6].

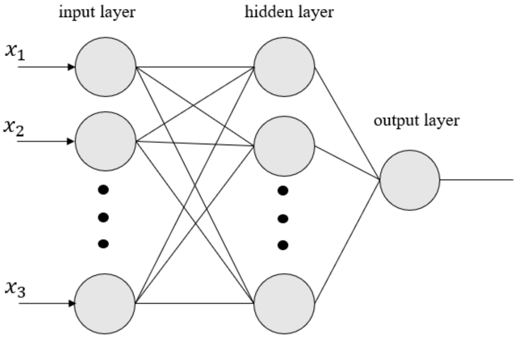

Fig. 2 RBF structure [12].

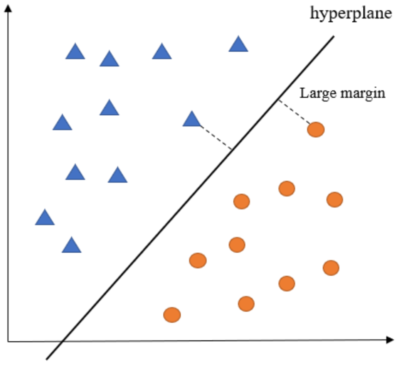

Fig. 3 Support Vector Machine [13].

D. *K-Nearest Neighbors (KNN).*

A simple classifier that falls under supervised machine learning can be easily implemented. The main concept in KNN is the assumption of similar data proximity. Fig. 4 illustrates the structure of KNN. The K in K-Nearest Neighbors is a constant value that the user will define and usually, it is an odd positive integer. Choosing K odd number will help avoid the tied vote.

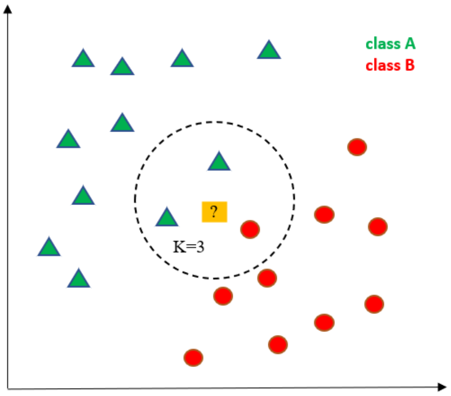

Fig. 4 KNN with K=3 [11].

IV. Methodology

In this section, a description of the utilized dataset will be introduced in addition to the proposed models.

A. *Dataset Description.*

This work aims to study the patients of Sylhet hospital in Sylhet, Bangladesh, and predict whether the patient is diabetic or not. The dataset that has been utilized in this work is the Early-Stage Diabetes Risk dataset. This dataset consists of 15 features and 520 patients. A description of the dataset has been elaborated in Table 1.

TABLE I. DATASET DESCRIPTION

No.	Attributes	Data Types	Values
1	Age	Int64	{20 to 65 years}
2	Gender	object	Male / Female
3	Polyuria	object	Yes / No
4	Polydipsia	object	Yes / No
5	Sudden weight loss	object	Yes / No
6	Weakness	object	Yes / No
7	Polyphagia	object	Yes / No
8	Genital Thrush	object	Yes / No
9	Visual blurring	object	Yes / No
10	Itching	object	Yes / No
11	Irritability	object	Yes / No
12	Delayed healing	object	Yes / No
13	Partial Paresis	object	Yes / No
14	Muscle stiffness	object	Yes / No
15	Alopecia	object	Yes / No
16	Obesity	object	Yes / No
17	Class	object	Positive / Negative

B. *Data Preprocessing*

In order to achieve the purpose of this work, a pre-processing procedure has been established on the utilized dataset. In my case, the target feature output has been changed from positive and negative to one's and zeroes. For the rest of the features, a label encoding have been performed to convert the object to numeric. For instance, the genders were Male and Female, by applying the label encoding, the male will be represented by 1 while the female by 0.

C. *Proposed Models*

After the pre-processing stage, the data is ready for modeling. Using Google Colaboratory, four types of machine learning classification techniques have been employed for predicting diabetes, which is SVM, MLP, KNN, and RBF. The following steps will illustrate the process of the proposed work.

- *First step:* using google colab, a new notebook have been established and the necessary libraries have been imported.
- *Second step:* uploading the desired dataset.
- *Third step:* the missing values have been checked, where in our case there were not any missing values.
- *Fourth step:* starting with the data preprocessing by separating the target feature and applying label encoding.
- *Fifth step:* for better performance, a feature selection has been applied by starting with finding the least correlated features to the target and shown in figure 5 one in addition to dropping them. In this work, the features with the least correlation percentage have been dropped; hence six features have been dropped.
- *Sixth step:* using the holdout set method, which is dividing the data into two sets, training, and testing sets. The testing set was 20% of the whole set, and 80% for training.
- *Seventh step:* implementing the four classifications.
- *Eighth step:* evaluating the model's accuracy, recall, and f1- score and choosing the best model accordingly.

In our proposed work, the default learning rate (α) value in python for Multilayer Perceptron (MLP) have been used, with two hidden layers, the first layer consists of six neurons, and the second consists of five neurons. The activation function of the hidden layers was by default the ReLU activation function. In Support Vector Machine (SVM), the kernel for it was linear [14].

In K-Nearest Neighbors, the default value of K is K=5, but in our proposed work K=3. As discussed in the previous section, the K value must be an odd number for ensuring that there will not be a tie in voting.

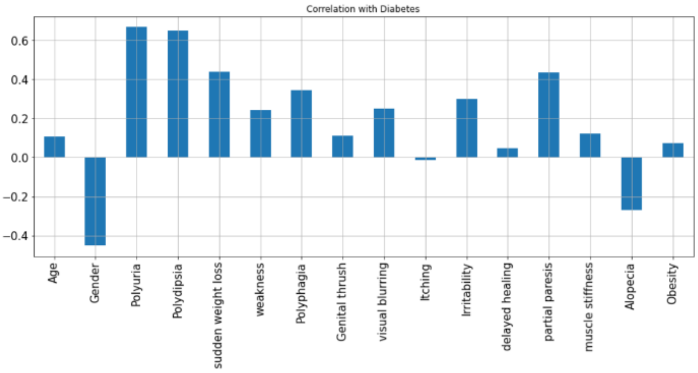

Fig. 5 correlation with Diabetes.

V. RESULTS AND DISCUSSION

As mentioned before, the Holdout set method have been performed in order to calculate the performance of the four machine learning classifiers that have been used. A splitting of 20% for testing and 80% for training.

A. *Evaluation Metric*

The metrics that have been used in this work will be discussed in this section, starting with accuracy:

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

Where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative. Moving to the Recall as defined below.

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

In addition to the Precision.

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

Finally, with the help of both Recall and precision, the F-measure score can be calculated by the following formula.

$$F - measure = \frac{2*Recall*Precision}{Precision+Recall} \quad (4)$$

Along with ROC which refers to Receiver Operating Characteristic. It is a measure of a classifier's performance.

ROC is a graphical representation of the trade-off between the true positive rate and false positive rate at different thresholds.

B. Results comparison

For the evaluation step, finding the accuracy, precision, confusion matrix, and recall helped in determining which of the utilized classifiers had the best performance and accuracy among the four classification methods. Beginning with the MLP, with 0.92% Precision, 0.91% Recall, and 0.91% F1-score. Moving to the SVM classifier, the score was better than MLP with 0.94% of Precision, Recall, and F1-score. Along with SVM, RBF achieved a better score reaching 0.96% in Precision, Recall, and F1-score. Finally, KNN had the better scores in Precision, Recall, and F1-score with 0.98%. The accuracy for MLP, SVM, RBF, and KNN is 0.91%, 0.94%, 0.96%, and 0.98% respectively. Finally, the ROC score is as follows: KNN with 0.98%, RBF with 0.95%, SVM with 0.93%, and MLP with 0.89%.

In ROC, the classifier that shows a curve that is close to the top left corner As shown in figure 6, has the best performance. In figure 6, KNN gives the closest curve to the top left corner among the other classifiers.

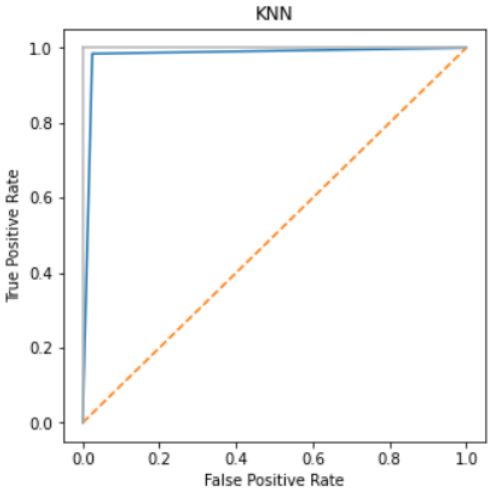

Fig. 6 ROC for KNN.

Overall, the best model have been chosen after experimentation, and according to the accuracy score and f1-score, it can be observed that KNN had the best performance with an accuracy of 0.98%.

VI. Conclusion

In this work, an evaluation to the utilized dataset, which is the Early-Stage Diabetes Risk dataset, have been performed, using machine learning algorithms. Taking into consideration the symptoms of diabetes and predicting the illness at its early stages using machine learning and its classifiers. To predict diabetes, a feature selection, testing, and experimenting with the data with four classifiers, Support Vector Machine (SVM), Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), and Radial Basis Neural Network (RBF) have been done.

In our results, KNN showed the best performance, where it was superior to the other machine learning classifiers. This result can help save lives by predicting diabetes at an early stage.

References

[1] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," *1st Int. Informatics Softw. Eng. Conf. Innov. Technol. Digit. Transform. IISEC 2019 - Proc.*, pp. 11–14, 2019, doi: 10.1109/UBMYK48245.2019.8965556.

[2] N. A. Abujabal and A. B. Nassif, "Meta-heuristic algorithms-based feature selection for breast cancer diagnosis: A systematic review," *Int. Conf. Electr. Comput. Commun. Mechatronics Eng. ICECCME 2022*, no. November, pp. 16–18, 2022, doi: 10.1109/ICECCME55909.2022.9988285.

[3] A. B. Nassif, M. A. Talib, Q. Nasir, Y. Afadar, and O. Elgendy, "Breast cancer detection using artificial intelligence techniques : A systematic literature review," *Artif. Intell. Med.*, vol. 127, no. March, p. 102276, 2022, doi: 10.1016/j.artmed.2022.102276.

[4] A. V. Srinivas, A. Ramya, G. T. Chandralekha, B. Vaagdevi, and K. Anand Goud, "Prediction of Diabetes Using Machine Learning," *Ymer*, vol. 21, no. 5, pp. 485–492, 2022, doi: 10.37896/YMER21.05/54.

[5] D. Vigneswari, N. K. Kumar, V. Ganesh Raj, A. Gugan, and S. R. Vikash, "Machine Learning Tree Classifiers in Predicting Diabetes Mellitus," *2019 5th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2019*, pp. 84–87, 2019, doi: 10.1109/ICACCS.2019.8728388.

[6] M. F. Faruque, Asaduzzaman, and I. H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus," *2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019*, pp. 7–9, 2019, doi: 10.1109/ECACE.2019.8679365.

[7] S. K. Dey, A. Hossain, and M. M. Rahman, "Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm," *2018 21st Int. Conf. Comput. Inf. Technol. ICCIT 2018*, pp. 21–23, 2019, doi: 10.1109/ICCITECHN.2018.8631968.

[8] R. A. Alassaf *et al.*, "Preemptive Diagnosis of Diabetes Mellitus Using Machine Learning," *21st Saudi Comput. Soc. Natl. Comput. Conf. NCC 2018*, no. April, 2018, doi: 10.1109/NCG.2018.8593201.

[9] P. Project-team, "NeuroImage Cross-validation failure : Small sample sizes lead to large error bars," vol. 180, no. June 2017, pp. 68–77, 2018, doi: 10.1016/j.neuroimage.2017.06.061.

[10] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," 2020.

[11] L. Yingwei, N. Sundararajan, and P. Saratchandran, "Performance Evaluation of a Sequential Minimal Radial Basis Function (RBF) Neural Network Learning Algorithm," vol. 9, no. 2, pp. 308–318, 1998.

[12] P. S. Muller and M. Nirmala, "A comparative study of classifiers for early diagnosis of gestational diabetes mellitus," vol. 770, pp. 754–770, 2020, doi: 10.31801/cfsuasm.

[13] N. P. Tigga and S. Garg, "ScienceDirect ScienceDirect Prediction of Type 2 Diabetes using Machine Learning Prediction of Type 2 Diabetes using Machine Learning Classification Methods Classification Methods," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 706–716, 2020, doi: 10.1016/j.procs.2020.03.336.

[14] I. Journal and O. F. Science, "Machine learning algorithms for Diabetes prediction and neural network method for blood glucose measurement," pp. 869–880, 2021.