Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM
e-ISSN: 2788-7421 p-ISSN: 2958-0544

Predicting Diabetes Disease Occurrence Using Logistic Regression: An Early Detection Approach

Ahmad Shaker Abdalrada[1*], Ali Fahem Neamah [2], Hayder Murad [3]

[1] Department of Software, Faculty of Computer Science and Information Technology, Wasit University, Iraq
[2] Department of Computer, Faculty of Computer Science and Information Technology, Wasit University, Iraq
[3] College of Medicine, Wasit University, Iraq

*Corresponding Author: Ahmad Shaker Abdalrada

DOI: https://doi.org/10.30880/ijcsm.2024.05.01.011
Received June 2023 ; Accepted August 2023 ; Available online January 2024

ABSTRACT: Diabetes disease is prevalent worldwide, and predicting its progression is crucial. Several model have been proposed to predict such disease. Those models only determine the disease label, leaving the likelihood of developing the disease unclear. Proposing a model for predicting the progression of disease becomes essential. Therefore, this article proposes a logistic regression model to anticipate the likelihood of Diabetes syndrome incidence. The model exploit capabilities of logistic regression by using sigmoid function. The model's performance was evaluated using the Pima Indians Diabetes dataset and demonstrated high accuracy, sensitivity, and specificity. The prediction accuracy rate was 77.6%, with a sensitivity of 72.4%, specificity of 79.6%, Type I Error of 27.6%, and Type II Error of 20.4%. Furthermore, the model indicates the feasibility of using laboratory tests, such as Pregnancies, Glucose, Blood Pressure, BMI, and DiabetesPedigreeFunction, to predict disease progress. The proposed model can aid patients and physicians in understanding the disease's progression and implementing timely interventions

Keywords: Diabetes disease, Machine learning, logistic Regression, diabetes prediction

1. INTRODUCTION

Diabetes mellitus (DM)is a chronic medical condition characterized by high levels of sugar (glucose) in the blood [1]. The body uses glucose as a primary source of energy, but to be used by cells, it requires insulin, a hormone produced by the pancreas. In people with diabetes, the body either doesn't produce enough insulin, or it can't use it properly, resulting in a buildup of glucose in the bloodstream.

The symptoms of DM are increased thirst and hunger, frequent urination, blurry vision, fatigue, slow healing of wounds, and numbness or tingling in the hands and feet [2, 3]. If left untreated, diabetes can lead to serious complications such as heart disease, stroke, kidney disease, nerve damage, and blindness[4, 5].

DM treatment relies on the type and severity of the condition [6]. Whether using an insulin pump or injections, patients with type 1 diabetes require lifetime insulin therapy. Type 2 diabetics may be able to control their disease with lifestyle modifications such a nutritious diet, consistent exercise, and weight loss, as well as with medications that improve the body's ability to utilize insulin or boost insulin production. Similar lifestyle modifications may be necessary for pregnant women with gestational diabetes, and some may also need insulin therapy..

Adopting a healthy lifestyle is essential for managing or preventing diabetes. This includes eating a balanced diet low in sugar and refined carbs, exercising frequently, maintaining a healthy weight, and controlling stress.[7]. Additionally, diabetics should routinely check their blood sugar levels and take their prescriptions as directed by their healthcare professional. [8].

In conclusion, millions of individuals throughout the world suffer from the terrible medical illness known as diabetes. People with diabetes can lead active, full lives with the right management, but this requires continual care and monitoring to avoid complications and maintain excellent health.

Large databases of clinical and health-related data, such as electronic health records, medical imaging, and genetic data, can be analyzed using machine learning (ML) algorithms to find patterns and make predictions. [9-13]. Predicting and diagnosing diabetes is one of the key uses of machine learning in this field. Additionally, it can be utilized to customize diabetes management and treatment. Healthcare practitioners can customize treatment programs for specific patients by using ML algorithms to find patterns in blood glucose levels and insulin requirements by evaluating a patient's

health data. Better glycemic control, a higher standard of living, and a lower risk of complications can all result from this.

Additionally, ML can be used to create decision support systems that can help doctors choose the best course of treatment for diabetic patients. To identify individuals who may be at danger of consequences like diabetic retinopathy or neuropathy, for instance, predictive models can be utilized. This enables early intervention and treatment.

Numerous research have looked into identifying diabetes by combining blood testing with different risk factors. [14-23]. Although these models have produced encouraging findings, they are unable to depict how the disease develops. The goal of this study is to close this gap because doing so will help to facilitate prompt actions, raise patient awareness, and have a big impact on healthcare providers.

In this study, we use the logistic regression method to introduce predictive model for the occurrence of diabetes disease. Blood tests are used by our model to estimate the likelihood of the condition. Our principal contributions include

- A workable model to predict the likelihood that DM may manifest.
- looking into how the added tests affect disease prediction,
- Utilizing the Pima Indians Diabetes dataset, we evaluated the performance of our model.

This paper is structured as follows: Section 2 provides a review of related works, Section 3 describes our proposed predictive model, and Section 4 presents the analysis of our model's performance, results, and discussion. Finally, in Section 5, we conclude our findings.

2. RELATED WORKS

In recent years, several studies have used ML algorithms to predict diabetes. Each proposed algorithm has its advantages and disadvantages, and the selection of algorithm relies on the type and quality of data available and the specific problem being addressed.

For example, the literature in [14] A reliable and robust framework has been presented for predicting diabetes, which involves various techniques such as outlier rejection, missing value imputation, standardization of data and feature selection. A number of classifiers, including k-nearest neighbor, decision trees (DT), naive Bayes(NB), random forest(RF), AdaBoost, XGBoost, and multilayer perceptron (MLP) , were utilized, with the aim of improving diabetes prediction through a weighted ensemble of different models. The performance metric used was the Area Under ROC Curve (AUC), which was maximized through hyperparameter tuning using the grid search technique. The experiments were conducted using the Pima Indian Diabetes Dataset, and the proposed ensembling classifier demonstrated better results compared to other methods discussed in the article, with sensitivity, specificity, false omission rate, diagnostic odds ratio, and AUC values of 0.789, 0.934, 0.092, 66.234, and 0.950 respectively. The AUC value is 2.00% better than the state-of-the-art results.

Another article in [15] proposed a Decision Support System (DSS) for DM prediction using ML algorithms. Both traditional ML and deep learning techniques are compared in this study. The most commonly used classifiers, (SVM) and (RF), were considered for conventional machine learning method. A fully Convolutional Neural Network (CNN) was employed for Deep Learning (DL) to predict DM patients. The proposed system was evaluated using the publicly available Pima Indians Diabetes database, which consisted tests of 768 people, where 500 of them were control and the rest were patients. The experimental results showed that the overall accuracy obtained using DL, SVM, and RF was 76.81%, 65.38%, and 83.67%, respectively. The results suggested that RF was more effective than deep learning and SVM methods for diabetes prediction.

Different work by [16] focused on developing a system that accurately predicts a patient's risk level for diabetes. They employ classification methods such as DT, Artificial Neural Network (ANN), NB, and SVM. The developed model that using DT demonstrates an accuracy of 85%, NB shows 77%, and SVM shows 77.3%. The results demonstrate significant accuracy in predicting diabetic risk using these methods. The system is designed to extract relevant information from a large amount of diabetes-related data and provide reliable risk predictions for patients.

In this research [17], the prediction of diabetes mellitus was carried out using decision tree, random forest, and neural network algorithms on a dataset of hospital physical examination data from Luzhou, China. The dataset contained 14 attributes, and the models were examined using five-fold cross-validation. To validate the universal applicability of the methods, independent test experiments were conducted using the best-performing methods. The training set consisted of randomly selected data from 68,994 healthy people and diabetic patients, respectively, and to address the issue of data imbalance, data was extracted five times, and the average result was computed. Principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) were used to decrease the data dimensionality. The outcomes indicated that the random forest algorithm achieved the highest accuracy (ACC = 0.8084) when using all attributes.

Additional research in [18] involved using data mining, machine learning (ML) algorithms, and Neural Network (NN) methods to predict diabetes. The Pima Indian Diabetes (PID) dataset, which includes information on 768 patients and nine unique attributes, was used for their analysis. Seven ML algorithms were applied to the dataset to forecast DM. The findings revealed that the LR and SVM classifiers performed well in predicting the disease. Additionally, they developed an NN model with different hidden layers and observed that the NN model with two hidden layers achieved 88.6% accuracy.

The research in [19] aimed to forecast DM by data mining techniques. The Backpropagation algorithm was utilized to expect whether an individual has diabetes or not. In addition, J48, NB, and SVM algorithms were also employed to predict DM. These NNA were structured with an input layer containing eight parameters, one hidden layer with six neurons, and one output layer. To enhance the performance of the model, the 5-fold cross-validation technique and a large value learning rate were implemented. The PIMA Indian dataset was utilized in this study, and the R programming language in RStudio was employed for implementation. The Back propagation algorithm demonstrated a performance of 83.11% accuracy, 86.53% sensitivity, and 76% specificity in predicting diabetes, which indicates an improvement over previous studies. The results were compared with those obtained using J48, Naive Bayes, and Support Vector Machine algorithms.

The objective of the study in [20] was to develop a model with the highest possible accuracy in predicting the probability of diabetes in patients. To achieve this, ML classification algorithms, namely DT, SVM, and NB, were utilized to detect early-stage diabetes. The Pima Indians Diabetes Database (PIDD), sourced from the UCI Machine Learning Repository, was used for the experiments. The achievement of the algorithms was evaluated based on measures such as Precision, Accuracy, F-Measure, and Recall. Accuracy was measured based on correctly and incorrectly classified instances. The results indicated that Naive Bayes outperformed the other algorithms with an accuracy of 76.30%. The results were verified using Receiver Operating Characteristic (ROC) curves in a systematic and rigorous manner.

This research paper in [21] aimed to predict diabetes using significant attributes and to characterize the relationship between these attributes. Various tools, such as clustering, prediction, and association rule mining, were utilized to select significant attributes. The principal component analysis method was used for attribute selection. Their outcomes revealed a robust association between DM and body mass index (BMI) and glucose level, which were identified using the Apriori method. To predict diabetes, ANN, RF, and K-means clustering techniques were employed. The ANN technique yielded the highest accuracy of 75.7% and could potentially assist medical professionals in making treatment decisions.

The study in [22] proposed a highly accurate model for diagnosing diabetes in patients. Additionally, this paper presented an effective diabetes prediction model that improves the classification of diabetes and enhances the accuracy of diabetes prediction using various ML algorithms. Several classifiers were employed for early-stage diabetes prediction, including SVM, LR, RF, DT, K-NN, Gaussian Process Classifier, AdaBoost Classifier, and NB. The models' performances were evaluated based on criteria such as Accuracy, Precision, Recall, F-Measure, and Error.

This study in [23] has utilized the random forest principle to develop an accurate model for diabetes diagnosis. The Pima Indians Diabetes dataset from the UCI repository was used for experiments. Multiple random forests were constructed with varying numbers of trees to determine the optimal forest size, and then compared with other machine learning techniques. Results indicated that the random forest approach outperforms other machine learning methods and is more efficient for diabetes diagnosis.

Despite the impressive results reported by the models proposed in the literature [7-16], they fail to address the issue of predicting the likelihood of diabetes disease occurrence. Moreover, these models are designed to make general predictions such as categorizing patients into healthy or sick categories, and do not provide information on the disease progression of individual patients

3. THE DESIGNED MODEL

This section presents the predictive model designed to estimate the likelihood of DM occurrence. The process of constructing the predictive model is depicted in Figure1, which outlines the various steps involved in this study.

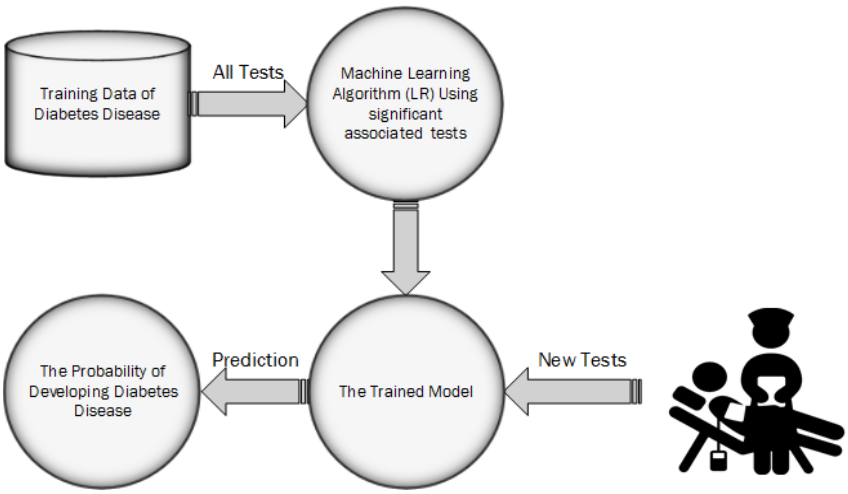

FIGURE 1. -. Illustrate the proposed model

As presented in figure1 above the model has developed to estimate the likelihood of diabetes disease. The model is versatile and can be implemented using many platforms and tools, including the R language. By providing both healthcare professionals and patients with a clear understanding of the disease's progression, the model can aid in timely intervention and increase awareness of the disease's risks. Thus, the proposed model enables effective planning for interventions and facilitates the management of the disease.

To construct our proposed predictive model, we have leveraged the abilities of machine learning algorithms such as logistic regression (LR). For logistic regression the sigmoid function serves as the formula for LR, which is stated as follows [24] :

$$P = \frac{1}{1 + \exp^{-z}} \dots\dots\dots\dots (1).$$

The symbol P denotes the chance of an event occurring (output variable), whereas z is a linear combination function of the input variables. An alternative expression for z is as follows:

$$z = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \cdots \beta_n v_n \dots\dots\dots\dots\dots (2)$$

The symbol β_0 represents the expected mean value of P when v is equal to 0. The parameter n indicates the number of independent variables, while β_n corresponds to the regression coefficient of each independent variable. These coefficients represent the degree of influence of each independent variable on the probability value of P. Lastly, v_n denotes the independent variables.

Utilizing the linear combination function in equation (1) with the employed tests as independent variables for determining the outcome of diabetes disease progression, the formula for the predictive model that estimates the likelihood of diabetes disease progression based on the predictability of the used tests can be expressed as follows:

$$P = \frac{1}{1 + exp^{-(\beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_3 T_3 + \beta_4 T_4 + \beta_5 T_5 + + \cdots + \beta_8 T_8)}} \quad (3).$$

The symbol P represents the probability of diabetes progression occurrence based on the selected independent variables. The parameter β denotes the regression coefficients of each test included in the model, where (T_1 to T_8) correspond to the included tests (i.e., preg, plas, pres, skin, insu, mass, pedi and age), respectively. The dataset section provides a description of each test.

Table 1. Presents the analysis of LR of all covariates

covariates	β	P value	Odds Ratio	CI (%)	
Intercept	-8.405	0.000	0.027	**Lower**	**Upper**
$\boldsymbol{T_1}$ (preg)	.123	.000	1.131	1.062	1.204
$\boldsymbol{T_2}$ (plas)	.035	.000	1.036	1.028	1.043
$\boldsymbol{T_3}$ (pres)	-.013	.011	.987	.977	.997
$\boldsymbol{T_4}$ (skin)	.001	.929	1.001	.987	1.014
$\boldsymbol{T_5}$ (insu)	-.001	.186	.999	.997	1.001
$\boldsymbol{T_6}$ (mass)	.090	.000	1.094	1.062	1.127
$\boldsymbol{T_7}$ (pedi)	.945	.002	2.573	1.432	4.625
$\boldsymbol{T_8}$ (age)	.015	.111	1.015	.997	1.034

LR was employed to determine the influence of the covariates. Table 1 illustrates the outcomes of utilizing covariates such as independent variables and diabetes such as the dependent variable in LR. Table 1 reveals that some of those independent variables are significantly associated with the DM where (P < 0.05) therefore have been used in the predictive model, while insignificant covariates are excluded. The covariates that have a significant association with the DM (P < 0.05) are T_1 (preg) , T_2 (plas), T_3 (pres) , T_6 (mass) and T_7 (pedi). However, the remaining covariates, including T_4 (skin), T_5 (insu) and T_8 (age) , have exhibited an insignificant association with the DM where (P > 0.05). Subsequently, Table 2 displays the coefficients of the covariates that displayed important associations, which were calculated using LR.

Table 2. The analysis of significant covariates

covariates	β	P value	Odds Ratio	CI (%)	
Intercept	-7.955	0.000	0.000	**Lower**	**Upper**
T_1 **(preg)**	0.153	0.000	1.166	1.104	1.231
T_2 **(plas)**	0.035	0.000	1.035	1.028	1.042
T_3 **(pres)**	-0.012	0.017	0.988	0.978	0.998
T_6 **(mass)**	0.085	0.000	1.089	1.059	1.119
T_7 **(pedi)**	0.911	0.002	2.486	1.397	4.423

Table 2 shows the results of LR analysis for the involved variables. The coefficients β for T_1 (preg) = 0.153 and (P value = 0.000) , β for T_2 (plas) = 0.035 and (P value = 0.000) , β for T_6 (mass) = 0.085 and (P value = 0.000) and the β for T_7 (pedi) = 0.911 and (P value = 0.002). The coefficient β for T_3 (pres) = -0.012 and (P value = 0.01).

The positive β coefficients for T_1, T_2 , T_6 and T_7 indicate a positive influence on the occurrence of diabetes disease, while the negative β coefficient for T_3 indicates a negative effect. Based on equation (3) and the results in Table 3, the predictive model was constructed as follows:

$$P = \frac{1}{1 + exp^{-(-7.955 + 0.153 * T_1 + 0.035 * T_2 - 0.012 * T_3 + 0.085 * T_6 + 0.911 * T_7)}} \quad (4).$$

Using the predictive model, new clinic test results can be used to calculate the probability of diabetes disease occurrence. For example, for patient A (T_1= 1, T_2= 189, T_3 = 60, T_6 = 30, T_7 = 0.4), patient B (T_1= 8, T_2= 99, T_3 = 84, T_6 = 35.4, T_7 = 0.39), and patient C (T_1= 1, T_2= 195, T_3 = 84, T_6 = 33, T_7 = 0.5), the chance of DM disease be able to be calculated using equation (4).

The probability of patient A developing diabetes disease is

$$P(A) = \frac{1}{1 + exp^{-(-7.955 + 0.153 * 1 + 0.035 * 189 - 0.012 * 60 + 0.085 * 30 + 0.911 * 0.4)}} = 0.73$$

For instance for patient B, the chance of developing DM disease can be measured as follows

$$P(B) = \frac{1}{1 + exp^{-(-7.955 + 0.153 * 8 + 0.035 * 99 - 0.012 * 84 + 0.085 * 35.4 + 0.911 * 0.39)}} = 0.29$$

Whereas the likelihood of patient C developing DM disease can be considered as follows

$$P(C) = \frac{1}{1 + exp^{-(-7.955 + 0.153 * 1 + 0.035 * 195 - 0.012 * T84_3 + 0.085 * 33 + 0.911 * 0.5)}} = 0.78$$

Moreover, to generalize such model for any dataset, above procedure can apply by discovering the covariates that have a significant association (P < 0.05) with the any diseases and employ equation (3) with the new discovered covariates.

4. THE MODEL EVALUATION

The evaluation and analysis of our proposed predictive model was conducted. Furthermore, this section presents the dataset, setup of experiment, and metrics employed.

4.1 Performance Metrics

The performance metrics used in the current study are Sensitivity, Specificity, Accuracy, Type I Error (α), and Type II Error. These metrics are commonly used in binary classification problems to assess the performance of a predictive model.

Sensitivity measures the proportion of true positive cases that are correctly identified by the model. It represents the number of participants who are correctly predicted with a positive disease. A high sensitivity indicates that the model is good at detecting positive cases, which is important in applications such as medical diagnoses.

$$\boldsymbol{Senstivity} = \frac{\boldsymbol{TP}}{\boldsymbol{FN + TP}} \quad \boldsymbol{(5)}.$$

Specificity measures the percentage of true negative cases that are correctly identified by the model. It represents the number of participants who are correctly diagnosed with a negative disease. A high specificity indicates that the model is good at detecting negative cases.

$$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{FP + TN}} \ (\mathbf{6}).$$

Accuracy computes the proportion of all cases that are correctly predicted by the model, both positive and negative. It represents the total number of participants correctly predicted with a positive and negative disease.

$$\boldsymbol{Accurcy} = \frac{\boldsymbol{TP + TN}}{(\boldsymbol{Total\ nubmer})} \ (\mathbf{7}).$$

Type I Error (α) quantifies the likelihood of detecting patients into a control group. It represents the cases where the model incorrectly predicts a positive disease when the actual diagnosis is negative.

$$\boldsymbol{\alpha} = \frac{\boldsymbol{FN}}{(\boldsymbol{FN + TP})} = \mathbf{1} - \boldsymbol{Sensitivity} \ (\mathbf{8}).$$

Type II Error calculates the probability of spotting control people with the patient group. It represents the cases where the model incorrectly predicts a negative disease when the actual diagnosis is positive.

$$\boldsymbol{Type\ II} = \frac{\boldsymbol{FP}}{(\boldsymbol{FP + TN})} = \mathbf{1} - \boldsymbol{Speci}\text{fi}\boldsymbol{city} \ (\mathbf{9}).$$

Overall, these performance metrics provide important insights into the performance of the predictive model and help to guide the optimization and refinement of the model for real-world applications.

4.2 Experiment

The experiments in this study were conducted applying R language on a PC with an Intel Core 7 processor, 2.5 GB CPU, 12 GB RAM running the Windows 8 operating system. To assess the performance of the predictive model, various measures that we explained in section 4.1 are used. In addition confusion matrix measurements was also calculated.

To validate the model, a ten cross-validation (CV) method was used to obtain a stable evaluation of the generalization error. The entire dataset was randomly divided into 10 subsets, with 9 subsets (90%) used for training and the remaining subset (10%) used for testing. This process was frequent ten times with change in the tested folds.

In addition the sensitivity, specificity, accuracy, Type I Error, and Type II Error were also computed to assess the achievement of the predictive model. Overall, these measures provide a comprehensive assessment of the model's ability to distinguish between malignant and benign tumors.

4.3 Dataset

The Pima Indians Diabetes Database is a dataset that contains information about the Pima Indians, a population of Native Americans in Arizona, USA, and their risk for developing diabetes. The dataset contains information on various medical and demographic factors, including age, BMI, blood pressure, and glucose levels, as well as whether or not an individual developed diabetes within 5 years of the initial examination.

The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) produced the dataset, which is frequently utilized in machine learning and data mining studies. It is frequently used as a benchmark dataset for creating and evaluating diabetes predicative models. As seen in Table 3, the dataset has 768 observations and 8 variables.

Table 3. Present the dataset.

Variables	Description
Pregnancies (preg)	"number of times pregnant"
Glucose(plas)	"plasma glucose concentration after 2 hours in an oral glucose tolerance test"
BloodPressure(pres)	"diastolic blood pressure (mm Hg)"
SkinThickness(skin)	"triceps skin fold thickness (mm)"
Insulin(insu)	"2-hour serum insulin (mu U/ml)"
BMI(mass)	"body mass index (weight in kg/(height in m)^2)"
DiabetesPedigreeFunction(pedi)	"diabetes pedigree function (a function which scores the likelihood of diabetes based on family history)"
Age(age)	"age in years"
The target variable is Outcome, which indicates whether or not an individual developed diabetes within 5 years of the initial examination (0 = no diabetes, 1 = diabetes).	

5. THE RESULTS OF AN EXPERIMENT AND A TEST

In this section, we will examine the experimental findings and give the outcomes in this part. Table 4, which displays the confusion matrix, provides an overview of how well our predictive model performed. Our prediction model has a predicted accuracy of 77.6%, as shown by the confusion matrix. In more detail, out of a total of 768 participants (268 patients and 500 healthy individuals), the model correctly predicted 155 patients and 441 healthy individuals. However, the model incorrectly identified 113 patients and 59 healthy individuals.

We attribute the strong correlation between all of the tested covariates and the DM disease ($P < 0.05$) for the predictive model's successful performance. This shows that using the combination of these tests to identify patients with the condition has a high diagnostic value. These findings suggest that our prediction model may help doctors identify patients who might benefit from additional diagnostic tests or therapy.

Furthermore, we reviewed the results to discover any potential shortcomings or areas for development. The relatively small sample size of our study is one restriction, which may have influenced the accuracy of our model. Future research with bigger sample sizes may assist to validate our findings. In addition, more tests or biomarkers could be added to future models to improve their predictive accuracy. Overall, our findings show that machine learning algorithms have the potential to improve diagnosis accuracy in medical settings.

Table 4. Confusion matrix

Class	Patient	Healthy
Patient	155	113
Healthy	59	441

The predictive model has a sensitivity of 72.4%, which is the percentage of patients that were properly classified as having a positive condition. Furthermore, the model had a specificity of 79.6%, which is the percentage of healthy participants who were accurately labeled as negative. The model's Type I and Type II errors were found to be 27.5% and 20.4%, respectively. These findings illustrate the model's efficacy in forecasting the risk of diabetic disease incidence and imply that it could be used to predict the possibility of other health issues as well.

6. CONCLUSIONS

This research presented a model for predicting the occurrence of diabetic illness. The analysis and evaluation of the model revealed that it is highly efficient and user-friendly, with an accuracy rate of 77.6%, a sensitivity rate of 72.4%, and a specificity rate of 79.6%. The model can be effectively used by healthcare providers to plan timely interventions and increase awareness of the risk of diabetes disease. Furthermore, laboratory tests such as Pregnancies, Glucose, Blood Pressure, BMI, and DiabetesPedigreeFunction were identified as significant predictors of diabetes disease categories. Healthcare providers can use this model if they have access to the results of the conducted tests. The model also has the potential to reduce the workload of trained specialists and enable untrained technicians to screen and process multiple patients objectively without relying on clinicians. We suggest this model for monitoring the progress of diabetes disease and as a tool for web-based and mobile phone interventions. However, the study has limitations in terms of feasibility and resource constraints, such as a small sample size that may limit its generalizability. As a result, we intend to increase the sample size and employ various tests to improve the model's effectiveness in future studies.

Funding

None

ACKNOWLEDGEMENT

None

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] R. Goyal and I. Jialal, "Diabetes mellitus type 2," 2018.

[2] A. Ramachandran, "Know the signs and symptoms of diabetes," *The Indian journal of medical research*, vol. 140, no. 5, p. 579, 2014.

[3] A. S. Abdalrada, J. Abawajy, T. Al-Quraishi, and S. M. S. Islam, "Prediction of cardiac autonomic neuropathy using a machine learning model in patients with diabetes," *Therapeutic Advances in Endocrinology and Metabolism*, vol. 13, p. 20420188221086693, 2022.

[4] V. Grote, S. Becker, and R. Kaaks, "Diabetes mellitus type 2–an independent risk factor for cancer?," *Experimental and clinical endocrinology & diabetes*, vol. 118, no. 01, pp. 4-8, 2010.
[5] T. Al-Quraishi, J. Abawajy, M. Chowdhury, R. Sutharshan, and A. Abdalrada, "Breast cancer risk assessment prediction using an ensemble classifier," *CAINE2017*, 2017.
[6] B. Silver *et al.*, "EADSG guidelines: insulin therapy in diabetes," *Diabetes therapy*, vol. 9, pp. 449-492, 2018.
[7] A. S. Abdalrada, J. H. Abawajy, M. U. Chowdhury, S. Rajasegarar, T. Al-Quraishi, and H. F. Jelinek, "Relationship between angiotensin converting enzyme gene and cardiac autonomic neuropathy among Australian population," in *Recent Advances on Soft Computing and Data Mining: Proceedings of the Third International Conference on Soft Computing and Data Mining (SCDM 2018), Johor, Malaysia, February 06-07, 2018*, 2018: Springer, pp. 135-146.
[8] M. Asif, "The prevention and control the type-2 diabetes by changing lifestyle and dietary pattern," *Journal of education and health promotion*, vol. 3, 2014.
[9] A. S. Abdalrada, J. Abawajy, M. Chowdhury, S. Rajasegarar, T. Al-Quraishi, and H. F. Jelinek, "Meta learning ensemble technique for diagnosis of cardiac autonomic neuropathy based on heart rate variability features," in *30th International conference on computer applications in industry and engineering, CAINE*, 2017, pp. 169-175.
[10] M . Al-Janabi, & Ismail, M. A. (2021). Improved intrusion detection algorithm based on TLBO and GA algorithms. Int. Arab J. Inf. Technol., 18(2), 170-179.
[11] S. Tahzeeb and S. Hasan, "A neural network-based multi-label classifier for protein function prediction," *Engineering, Technology & Applied Science Research*, vol. 12, no. 1, pp. 7974-7981, 2022.
[12] K. Koklonis, M. Sarafidis, M. Vastardi, and D. Koutsouris, "Utilization of machine learning in supporting occupational safety and health decisions in hospital workplace," *Engineering, Technology & Applied Science Research*, vol. 11, no. 3, pp. 7262-7272, 2021.
[13] T. Al-Quraishi, J. H. Abawajy, N. Al-Quraishi, A. Abdalrada, and L. Al-Omairi, "Predicting breast cancer risk using subset of genes," in *2019 6th international conference on control, decision and information technologies (CoDIT)*, 2019: IEEE, pp. 1379-1384.
[14] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516-76531, 2020.
[15] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A decision support system for diabetes prediction using machine learning and deep learning techniques," in *2019 1st International informatics and software engineering conference (UBMYK)*, 2019: IEEE, pp. 1-4.
[16] P. Sonar and K. JayaMalini, "Diabetes prediction using different machine learning approaches," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 2019: IEEE, pp. 367-371.
[17] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in genetics*, vol. 9, p. 515, 2018.
[18] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432-439, 2021.
[19] F. G. Woldemichael and S. Menaria, "Prediction of diabetes using data mining techniques," in *2018 2nd international conference on trends in electronics and informatics (ICOEI)*, 2018: IEEE, pp. 414-418.
[20] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia computer science*, vol. 132, pp. 1578-1585, 2018.
[21] T. M. Alam *et al.*, "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, p. 100204, 2019.
[22] P. Palimkar, R. N. Shaw, and A. Ghosh, "Machine learning technique to prognosis diabetes disease: Random forest classifier approach," in *Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2021*, 2022: Springer, pp. 219-244.
[23] S. Benbelkacem and B. Atmani, "Random forests for diabetes diagnosis," in *2019 International Conference on Computer and Information Sciences (ICCIS)*, 2019: IEEE, pp. 1-4.
[24] H. Nampak, B. Pradhan, and M. Abd Manap, "Application of GIS based data driven evidential belief function model to predict groundwater potential zonation," *Journal of Hydrology*, vol. 513, pp. 283-300, 2014.