

Classification and regression tree model for diabetes prediction

Farah Najidah Noorizan¹, Nur Anida Jumadi^{1,2}, Li Mun Ng¹

¹Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja, Malaysia

²Advanced Medical Imaging and Optics (AdMedic), Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja, Malaysia

Article Info

Article history:

Received Jan 22, 2025

Revised Oct 29, 2025

Accepted Nov 5, 2025

Keywords:

CART

Diabetes mellitus

Evaluation metrics

Hyperparameter tuning

Machine learning

ABSTRACT

Diabetes mellitus is characterized by excessive blood glucose that occurs when the pancreas malfunctions while producing insulin. High blood glucose levels can cause chronic damage to organs, particularly the eyes and kidneys. Diabetes prediction models traditionally use a variety of machine learning (ML) algorithms by combining data from the glucose levels, patient health parameters, and other biomarkers. Prior research on diabetes prediction using various algorithms, such as support vector machine (SVM) and decision tree (DT) models, demonstrates an accuracy rate of approximately 70%, which is relatively modest. Therefore, in this study, a classification and regression tree (CART) multiclassifier model has been proposed to improve the accuracy of diabetes prediction, which is based on three classes: non-diabetic, pre-diabetic, and diabetic. The study involved data preprocessing steps, hyperparameter tuning, and evaluation of performance metrics. The model achieved 97% accuracy while utilizing the value of 5 for the number of leaves per node, the value of 10 for the maximum number of splits, and deviance as the split criterion, which also resulted in a precision of 98%, recall of 97%, and F1-score of 98%, showing that the proposed multiclassifier model can accurately predict diabetes. In conclusion, the proposed CART model with the best hyperparameter setting can enable the highest accuracy in predicting diabetes classes.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Nur Anida Jumadi

Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia (UTHM)

Parit Raja, Batu Pahat, 86400, Malaysia

Email: anida@uthm.edu.my

1. INTRODUCTION

Diabetes mellitus is a metabolic condition characterized by excessive blood glucose and hyperglycemia. This sickness occurs when the pancreas, which produces insulin, malfunctions, or when cells do not adequately respond to the insulin composition [1], [2]. The International Diabetes Federation defines diabetes mellitus as a chronic condition that substantially impacts global health. Approximately 537 million adults (aged 20 to 79) are diagnosed with diabetes, and an estimated 6.7 million diabetes patients worldwide died in 2021. The number of diabetic individuals is expected to rise by 46% by 2045 [3], [4]. There is as yet no cure, and the prevalence is still increasing. High blood glucose levels, characterized by diabetes, and long-term hyperglycemia can cause chronic damage to various organs, particularly the eyes, kidneys, heart, blood vessels, and nervous system [5], [6].

Diabetes is frequently self-managed by patients through many assessments of the glucose level throughout the day, as well as insulin administration via injection or pump, which can be difficult for patients who face numerous challenges in their everyday lives [7]. A conventional blood glucose device uses disposable strips of glucose oxidase biosensors to measure glucose concentration from the acquired fingertip blood [8], [9]. Diabetes prediction models traditionally use a variety of machine learning (ML) algorithms,

including decision tree (DT), logistic regression (LR), and support vector machine (SVM), with various feature sets and datasets, such as glucose levels, health parameters, and other biomarkers [10]. Previous research on diabetes prediction using various ML algorithms, including SVM and DT models, has shown an accuracy rate of roughly 70%, which is rather modest.

The authors in [11]–[13] used various algorithms to predict diabetes using various features, such as personal data and health parameters, including LR, K-nearest neighbors (KNN), Naive Bayes (NB), DT, random forest (RF), artificial neural network, and SVM. These models' accuracy is slightly lower, ranging from 70% to 90%.

Furthermore, the authors in [14]–[17] predicted diabetes using seven different algorithms in predicting diabetes, which were SVM, LR, gradient boosting machine, RF, DT, KNN, and extreme gradient boosting model (XGBoost), and discovered that LR and SVM are useful for diabetes prediction. In addition, [18] developed six ML algorithms to train a dataset using multiple techniques of feature selection for model accuracy improvements. El-Bouhissi *et al.* [19] designed a model for gestational diabetes mellitus prediction using the classifier of deep neural network (DNN), SVM, and RF, which produced an accuracy of approximately 90% to 95%.

The authors in [20]–[22] proposed an ensembling classifier for diabetes prediction based on various ML classifiers, such as KNN, DT, RF, AdaBoost, NB, XGBoost, and multilayer perceptron. In recent years, [23] used Jupyter Notebook to create a new stacking ensemble model for diabetes prediction using RF and LR as base learner models, while [24] used RF, convolutional neural network with long short-term memory network, and sequential dense layers as base learner models, where both studies used the XGBoost model as the meta-learner model. All these models showed an accuracy of 83% to 95%.

This present research proposes a classification and regression tree (CART) model for the accuracy improvement of diabetes prediction. The model employs Gini, deviance, and hyperparameter tuning to identify the optimal splits for efficiency, accuracy, and overfitting avoidance. Contrary to the traditional DT algorithms, such as ID3 [25], which lacks overfitting control and employs multi-branch splits, the use of the CART model aimed to be a more suitable approach for medical prediction. Its capacity to perform classification and regression tasks was utilized to improve its efficiency, interpretability, and accuracy for diabetes prediction.

2. RESEARCH METHOD

A diabetes dataset was collected from the laboratory of a medical city hospital, which contained the data of a total of 1,000 subjects [26]. There were 11 features, which were gender, age, urea level, creatinine ratio, hemoglobin level, cholesterol level, triglyceride level, high-density lipoprotein level, low-density lipoprotein level, very low-density lipoprotein level, and body mass index (BMI). The target output was divided into three classes: class 0 for non-diabetic, class 1 for pre-diabetic, and class 2 for diabetic. The proposed model started with the preprocessing technique and the splitting of the dataset into training and testing using an 80:20 ratio. The CART model was then trained using several hyperparameter tuning settings, which were the number of leaves per node, the maximum number of splits, and the split criterion. Then, the model's performance was validated using the metrics of accuracy, recall, precision, F1-score, and receiver operating characteristic-area under the curve (ROC-AUC). Figure 1 illustrates the overall process of the proposed model's development. A model deployment was created to visualize the predicted diabetes status of patients.

2.1. Data preprocessing

Data preprocessing is critical in ML and data analytics. Preprocessing improves a dataset by removing or imputing missing values, screening outliers, and eliminating noise, ensuring the data are correct and reliable. When working with non-numeric data, categorical processing is essential to data preprocessing. ML models, such as tree-based models and classification metrics, often demand numerical inputs; therefore, categorical variables must be translated properly before being fed into the model.

Figure 1 shows the model's dataset, consisting of 11 input features. Gender was represented as categorical data, with 0 for male and 1 for female, while the remaining data were classified as numerical data. Aside from that, the target data were numerical, with values of 0, 1, and 2 being used as class labels. To conduct the classification tasks, the class labels must be turned into categorical data to distinguish their distinct groups of non-diabetics, pre-diabetic, and diabetic.

2.2. CART

The CART algorithm combines DTs and regression to solve classification and regression issues. It divides a dataset into branches depending on feature values, using Gini impurity for classification and the

mean squared error for regression [27]. Gini and deviance were suitable approaches for classification in this study.

A Gini index assesses the impurity in categorization tasks by calculating the probability of mistakenly classifying randomly chosen data from the split. Deviance, on the other hand, measures how effectively a split improves prediction accuracy by comparing a model's likelihood before and after the split. It is commonly employed in LR or classification tasks, where a lower deviance suggests a better fit and more accurate predictions.

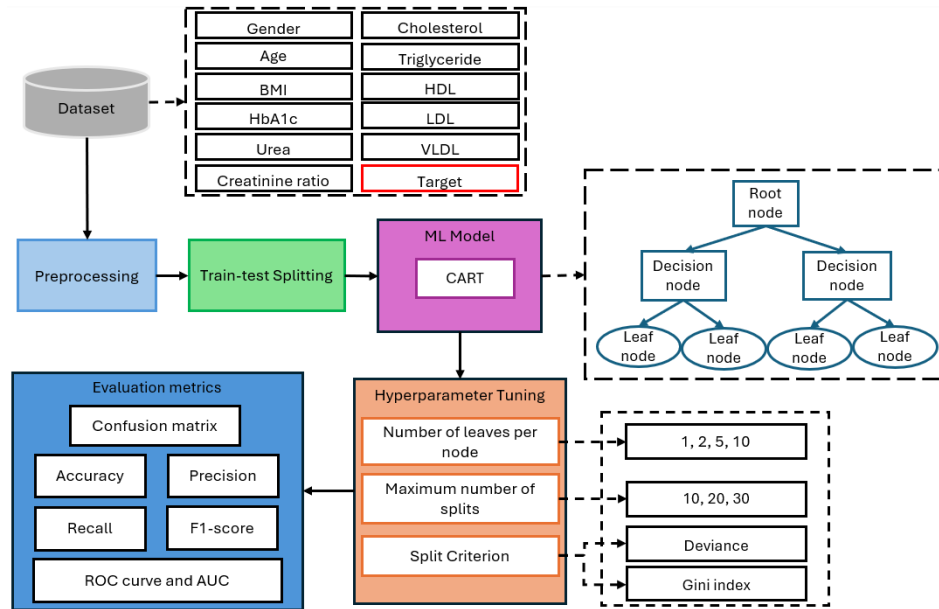


Figure 1. The overall process of the CART multiclassifier for diabetes prediction

2.3. Hyperparameter tuning

In ML, hyperparameter tuning refers to determining the value of a parameter before the learning process begins [28]. CART algorithms often have a fixed set of hyperparameters, such as the maximum number of splits. The number of leaves per node is the smallest number of observations (data points) necessary in a leaf node. Its purpose is to prevent overfitting by prohibiting the tree from splitting further when the number of samples in a node is smaller than this value. The larger numbers result in a simpler tree, while smaller numbers allow for more detailed splits. The parameter of the maximum number of splits restricts the tree's growth, lowering complexity and limiting overfitting by avoiding unnecessary branching. Smaller numbers result in a simpler tree, but larger values provide greater flexibility.

Deviance is frequently utilized in a CART model when working with binary outcomes to assess the goodness of fit of classification models.

$$D = -2 \sum_{i=1}^n \sum_{c=1}^C [y_{i,c} \log(p_{i,c})] \quad (1)$$

The deviation in (1) is calculated using the log-likelihood function in (2), which assesses the fit quality for classification models.

$$\log L(\theta) = \sum_{i=1}^n \sum_{c=1}^C [y_{i,c} \log(p_{i,c})] \quad (2)$$

Where n is the number of observations, and $y_{i,c}$ is the actual class label, which equals 1 if the i -th observation belongs to class c , while $p_{i,c}$ is the predicted probability that observation i belongs to class c . When dividing nodes into a CART tree, deviation is evaluated for each potential split, and the split with the largest reduction in deviance is selected as the optimal decision node.

The Gini index is a statistic that determines how mixed or pure the data is in a DT node. It is computed using (3), where C is the number of classes in the target variables and p_i is the proportion of components in the split that belong to class i .

$$Gini = 1 - \sum_{i=1}^C p_i^2 \quad (3)$$

Deviance is used as the split criterion for several reasons, particularly when a dataset's classes are imbalanced, as in the present study's diabetes dataset, which had a higher proportion of class 2 than classes 0 and 1. In this context, deviance can handle such scenarios more effectively by emphasizing probabilistic separations over pure splits.

2.4. Performance metrics

Classification accuracy is one of the performance evaluation metrics that display how well a model predicts instances based on training data. In this study, the performance metric measurements were divided into the following metrics, which were accuracy, precision, recall, F1-score, and ROC-AUC curve, as expressed in (4) to (8). Table 1 represents the multiclass confusion matrix structure, where the rows represent the true classes, and the columns represent the predicted classes. The term C_0 refers to class 0, C_1 refers to class 1, C_2 refers to class 2, and C_n refers to the n -th class. In addition, TP is defined as true positive, TN as true negative, FP as false positive, FN as false negative, R_i is the rate of the i -th data, and I_f and I_l are negative and positive data, respectively.

Table 1. Confusion matrix structure for multiclass classification

Actual class	Predicted class				
	Classes	C_0	C_1	C_2	C_n
	C_0	TP	FP	TN	TN
	C_1	FN	TP	FN	FN
	C_2	TN	FP	TN	TN
	C_n	TN	FP	TN	TN

The accuracy ratio is the number of true predicted instances, positive and negative, divided by the total number of cases.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Precision is the ratio of expected positive instances to total predicted positive instances.

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Recall is obtained by dividing the number of true positives by the number of true positives plus the number of false negatives.

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

The F1-score is used to assess the overall performance. It weighs the harmonic meaning of precision and recall.

$$F1 - score = \frac{2TP}{2TP+FP+FN} \quad (7)$$

The AUC is a performance metric for a binary classification model, which can be used to differentiate between positive and negative classes. The AUC is the area under the ROC curve, which compares the true positive rate against the false positive rate at different categorization levels.

$$AUC = \frac{\sum R_i(I_l) - I_l(I_l+1)/2}{I_l+I_f} \quad (8)$$

2.5. Model deployment

A deployment model in ML integrates a trained ML model into a real-world system or application to create predictions automatically. This model allows end users to enter data and receive predictions or insights from the model. Figure 2 represents this study's model deployment for diabetes prediction, performed using MATLAB software. This model determined or forecast a patient's diabetes status or class by entering personal information, such as gender, age, BMI, blood sugar, cholesterol, and other health data.

2.6. Pseudocode for diabetes prediction using CART and model deployment

Pseudocode is a simplified, high-level representation of a program or algorithm that combines normal language and programming ideas. It is not created in a particular programming language and does not adhere to strict grammar constraints, making it easier to understand. In this study, model predictions and deployment were implemented using MATLAB software. Table 2 presents the pseudocode that explains the steps for developing the model.

Enter the following details for diabetes prediction:

Gender (0 for male, 1 for female):

Age (years):

Urea (mg/dL):

Creatinine (Cr, mmol/L):

HbA1c (mmol/L):

Cholesterol (mmol/L):

Triglycerides (TG, mmol/L):

High-Density Lipoprotein (HDL, mmol/L):

Low-Density Lipoprotein (LDL, mmol/L):

Very Low-Density Lipoprotein (VLDL, mmol/L):

Body Mass Index (BMI):

Predicted Class:

Figure 2. Model deployment

Table 2. Pseudocode for diabetes prediction using CART and model deployment

Pseudocode for diabetes prediction using CART	
Input: Load and read the dataset of 11 features (gender, age, urea, creatinine ratio, HbA1c, cholesterol, triglyceride (Tr), HDL, LDL, VLDL, and BMI)	
Output: Target class	
1.	Convert the target into categorical data
2.	Split the dataset into training (80%) and testing (20%) subsets
3.	Define hyperparameter tuning
-	Minimum leaf size
-	Maximum number of splits
-	Split criterion
4.	Train the CART model using the training features and target with the defined hyperparameters
5.	Performance of diabetes prediction model based on evaluation metrics:
-	Confusion matrix
-	Accuracy: the overall correct predictions
-	Precision: emphasizes the accuracy of positive predictions
-	Recall: the ability to find all positive cases
-	F1-score: balances between precision and recall
-	ROC-AUC curve: analyzes model performance at various thresholds
6.	Save and simulate the trained model
7.	Prompt the user to input data
-	Gender (0 for male, 1 for female)
-	Age (in years)
-	Urea level (in mgdl/L)
-	Creatinine level (in mmol/L)
-	HbA1c level (in mmol/L)
-	Triglyceride (in mmol/L)
-	High-density lipoprotein (in mmol/L)
-	Low-density lipoprotein (in mmol/L)
-	Very Low-density lipoprotein (in mmol/L)
-	Body Mass Index
8.	Preparing data for prediction
-	Normalization using means and standard deviation values
9.	Displays the prediction result to the user
-	0: Non-diabetic, 1: Pre-diabetic, 2: Diabetic
10.	End

3. RESULTS AND DISCUSSION

This section presents the overall findings and corresponding discussion. Firstly, in subsection 3.1.1, the best hyperparameter tuning setting that improved the model's accuracy is discussed. In subsection 3.1.2, the performance of the developed model through the evaluation metrics is explained. Finally, in subsection 3.1.3, the outcomes after inserting the information into the model deployment are discussed.

3.1. Results

This study investigated the effectiveness of diabetes prediction using an ML model approach, namely the CART model. Unlike the prior research, which had very moderate performance and frequently focused on binary classification, this study attempted to predict three separate diabetes conditions: non-diabetic, pre-diabetic, and diabetic.

3.1.1. Hyperparameter tuning

Table 3 shows the best hyperparameter tuning setting for the model. Values of 5 and 10 were selected for the number of leaves per node and the maximum number of splits, respectively, while the split criterion metric was deviance. This hyperparameter tuning can increase accuracy and performance across multiple evaluation metrics.

Table 3. The best hyperparameter tuning setting for the model

The best hyperparameter tuning parameters	
Number of leaves per node	5
Maximum number of splits	10
Split criterion	Deviance

On the other hand, settings with the combination of values other than 5 and 10, as well as other than deviance, did not perform well, with the model's accuracy ranging from 85% to 93%. For example, the combination of the value of 2 for the number of leaves per node, the value of 20 for the maximum number of splits, and Gini for the split criterion shows an accuracy of 91.67%, with slightly lower recall and a few misclassifications for class 1 and class 2.

3.1.2. Performance of evaluation metrics

In this subsection, the experimental results obtained after training the diabetes dataset using the proposed CART multiclassifier model are described. Table 4 represents the result of the confusion matrix for class 0, where the model accurately identified 17 occurrences, with no misclassifications of other classes. In addition, 10 occurrences in class 1 were appropriately classified. Lastly, the model correctly identified 20 samples for class 2, with only one misclassified as class 0.

Table 4. Result for multiclass confusion matrix

True class \ Predicted class	Class 0	Class 1	Class 2	Total
	Class 0	Class 1	Class 2	Total
Class 0	17	0	0	17
Class 1	0	10	0	10
Class 2	1	0	20	21
Total	18	10	20	48

Table 5 represents the performance measurements of the CART multiclassifier model, evaluated across three classes (0, 1, and 2) based on the model's accuracy, precision, recall, and F1-score. The model's total accuracy was 97%, suggesting that predictions for all classes were correct. For class 0, the model achieved 94% precision, indicating that the prediction was correct. Recall was perfect at 100%, indicating that the model accurately recognized all instances of class 0, and the F1-score was 97%, suggesting a balanced performance for this class. The model performed very well for class 1, with 100% for the precision, recall, and F1-score metrics, suggesting faultless classification for false positives or negatives. For class 2, the model obtained a precision of 100%, indicating that all cases predicted as class 2 were precise. Although the recall and F1-score values were a little lower, around 95% and 98%, respectively, this still indicated a superior performance in this class.

Table 5. Performance measurement of the CART multiclassifier model

Class	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
0	97	94	100	97
1		100	100	100
2		100	95	98

Figure 3 depicts the result of the ROC-AUC curve of the model's performance. The AUC value for class 1 was 1, indicating the model successfully differentiated this class from the other classes with no misclassification. Also, the values for class 0 and class 2 were 0.98 and 0.99, respectively, suggesting a nearly perfect separation with a small possibility of misclassification.

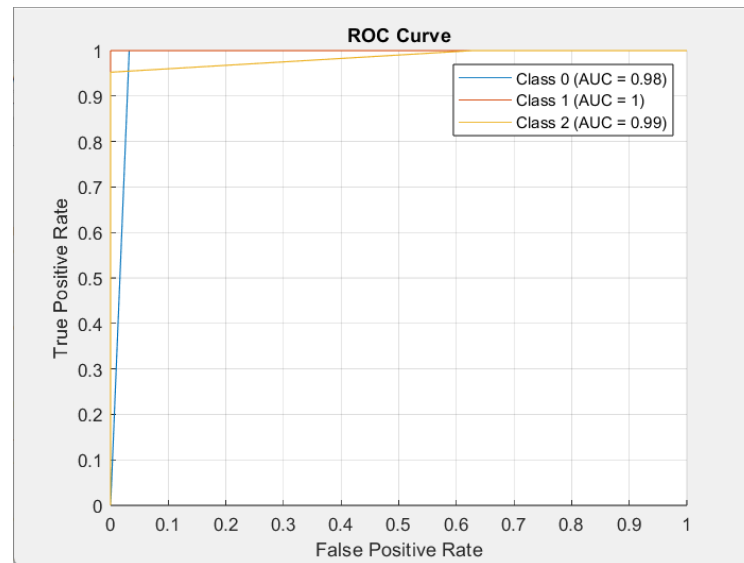


Figure 3. The ROC-AUC curve

3.1.3. Deployment outcomes

Biographical information and health data can be used to determine this model's functionality in identifying a patient's diabetes status. In this subsection, the result of the deployed diabetes prediction model after receiving user input, as depicted in Figure 4, is described. The model was simulated using data from a 40-year-old female patient, with a BMI of 24 and a hemoglobin level of 5.4 mmol/L. The result shows a class of 0, meaning that the patient is not in the category of diabetics. In addition, this predicted scenario of non-diabetic can be proved, as the patient's normal glucose level is below 5.7 mmol/L and the BMI of 24 is within the normal range.

```

Enter the following details for diabetes prediction:
Gender (0 for Male, 1 for Female): 1
Age (years): 40
Urea (mg/dL): 22
Creatinine (Cr, mmol/L): 159
HbA1c (mmol/L): 5.4
Cholesterol (mmol/L): 3.1
Triglycerides (TG, mmol/L): 1.6
High-Density Lipoprotein (HDL, mmol/L): 1.1
Low-Density Lipoprotein (LDL, mmol/L): 1.3
Very Low-Density Lipoprotein (VLDL, mmol/L): 0.7
Body Mass Index (BMI): 24
Predicted Class: 0

```

Figure 4. The deployment outcome

3.2. Discussion

The proposed method in this study was shown to have outstanding performance in predicting diabetes across the three categories, with an overall accuracy of 97%. It greatly outperformed other ML models, such as SVM, KNN, and DT, which had accuracy rates ranging from 70% to 90% [11], [13], [29]. Each class also had strong precision, recall, and F1-score values, especially class 1 (with no misclassification). These findings support the model's capacity to handle multiclass classification effectively, particularly with imbalanced datasets.

Previous studies [11], [12] simulated the dataset using a training and testing split, as well as a validation step to evaluate the model's performance. However, these studies did not emphasize systematic tuning of model parameters, which resulted in poor prediction performance. Therefore, our study suggests that excellent accuracy and balanced performance across all classes are due to the successful combination of data preprocessing, optimal hyperparameter tuning, and the use of deviation as the split criterion. The setting of the number of leaves and the maximum number of splits assists in creating a DT that is neither underfitted nor overfitted. In addition, deviance, which is recognized for dealing with uneven class distributions, helps the model to discriminate between borderline instances.

This study explored a model for diabetes status prediction based on clinical data, which performed well across three classes. However, several constraints should be considered, such as the fact that a dataset that contains many subjects, but is based on a single population group, may limit the model's applicability to other ethnicities or areas. Furthermore, excluding behavioral or lifestyle factors may decrease the model's predictive power. Finally, the current deployment arrangement requires manual input, which might be improved with automation for real-time applications in clinical situations.

4. CONCLUSION

In conclusion, this study found that the CART multiclassifier model is a dependable and accurate technique to predict diabetes status using clinical characteristics. The model achieved great overall accuracy by combining data preprocessing, optimal hyperparameter tuning, and the use of deviation as the split criterion. The model's remarkable performance across various evaluation parameters underlines its potential for early screening and clinical decision support.

In the future, this project could be improved by developing a user-friendly graphical user interface (GUI) to boost the operation of the diabetes prediction model. This GUI allows users to easily, swiftly, and systematically enter all their personal and health information to instantly determine their diabetes status. Furthermore, a visual interface, such as future healthcare-related actions, can be added, making this tool even more beneficial to the healthcare community.

ACKNOWLEDGMENTS

This research was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through GPPS (vot Q644) and Leave a Nest Co., Ltd. through international grants (vot X311 and vot X324).

FUNDING INFORMATION

The funding information can be referred to in the acknowledgement section.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Farah Najidah Noorizan	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Nur Anida Jumadi	✓			✓		✓		✓		✓		✓	✓	✓
Ng Li Mun							✓						✓	

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : **O**riginal Draft

E : **E**diting

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

The authors confirm that there is no conflict of interest related to the manuscript.

DATA AVAILABILITY

The data that support the findings of this study are openly available in the Mendeley repository at doi: 10.1371/journal.pone.0300785.




REFERENCES

- [1] S. A. Mohshim, N. F. N. Nasiruddin, Z. Zakaria, H. M. Desa, D. F. Mohshim, and M. F. Fadzir, "Non-invasive blood glucose monitor using arduino for clinical use," *2023 International Conference on Engineering Technology and Technopreneurship, ICE2T 2023*, pp. 219–224, 2023, doi: 10.1109/ICE2T58637.2023.10540514.
- [2] E. U. Dewi, N. P. Widari, Nursalam, Mahmudah, E. Y. Sari, and Y. F. N. Susiana, "The relationship between diabetes self-care management and blood glucose level among type 2 diabetes mellitus patients," *International Journal of Public Health Science (IJPHS)*, vol. 12, no. 3, pp. 1165–1170, 2023, doi: 10.11591/ijphs.v12i3.22228.
- [3] M. Naresh, V. S. Nagaraju, S. Kollem, J. Kumar, and S. Peddakrishna, "Non-invasive glucose prediction and classification using NIR technology with machine learning," *Heliyon*, vol. 10, no. 7, 2024, doi: 10.1016/j.heliyon.2024.e28720.
- [4] E. Mansour *et al.*, "Review on non-invasive electromagnetic approaches for blood glucose monitoring using machine learning," *Proceedings of the 11th International Japan-Africa Conference on Electronics, Communications, and Computations, JAC-ECC 2023*, pp. 273–276, 2023, doi: 10.1109/JAC-ECC61002.2023.10479620.
- [5] B. Chitradevi, Supriya, N. S. Chandra, T. N. Chitradevi, and H. Alabdeli, "Diabetes mellitus prediction and classification using firefly optimization-based support vector machine," *International Conference on Distributed Computing and Optimization Techniques, ICDCOT 2024*, 2024, doi: 10.1109/ICDCOT61034.2024.10515397.
- [6] A. Maenge, T. Sigwele, C. Bhende, C. Mokgethi, V. Kuthadi, and B. Omogbehin, "Optimizing diabetes prediction using machine learning: a random forest approach," *International Journal of Advances in Applied Sciences (IJAAAS)*, vol. 14, no. 2, p. 454, 2025, doi: 10.11591/ijaas.v14.i2.pp454-468.
- [7] M. Tejedor, A. Z. Woldaregay, and F. Godtliebsen, "Reinforcement learning application in diabetes blood glucose control: a systematic review," *Artificial Intelligence in Medicine*, vol. 104, 2020, doi: 10.1016/j.artmed.2020.101836.
- [8] L. Tang, S. J. Chang, C. J. Chen, and J. T. Liu, "Non-invasive blood glucose monitoring technology: a review," *Sensors (Switzerland)*, vol. 20, no. 23, pp. 1–32, 2020, doi: 10.3390/s20236925.
- [9] A. S. Bolla and R. Priefer, "Blood glucose monitoring- an overview of current and future non-invasive devices," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 14, no. 5, pp. 739–751, 2020, doi: 10.1016/j.dsx.2020.05.016.
- [10] K. Liu *et al.*, "Machine learning models for blood glucose level prediction in patients with diabetes mellitus: systematic review and network meta-analysis," *JMIR Medical Informatics*, vol. 11, no. 1, 2023, doi: 10.2196/47833.
- [11] A. C. Lyngdoh, N. A. Choudhury, and S. Moulik, "Diabetes disease prediction using machine learning algorithms," *Proceedings - 2020 IEEE EMBS Conference on Biomedical Engineering and Sciences, IECBES 2020*, pp. 517–521, 2021, doi: 10.1109/IECBES48179.2021.9398759.
- [12] N. Mohan and V. Jain, "Performance analysis of support vector machine in diabetes prediction," 2020, doi: 10.1109/ICECA49313.2020.9297411.
- [13] L. J. Muhammad, E. A. Algehyne, and S. S. Usman, "Predictive supervised machine learning models for diabetes mellitus," *SN Computer Science*, vol. 1, no. 5, 2020, doi: 10.1007/s42979-020-00250-8.
- [14] 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, doi: 10.1016/j.ict.2021.02.004.
- [15] D. Kaveripakam and J. Ravichandran, "Comparative analysis of machine learning algorithms for diabetic disease identification," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 45, no. 1, pp. 40–50, 2025, doi: 10.37934/araset.45.1.4050.
- [16] N. Ahmed *et al.*, "Machine learning based diabetes prediction and development of smart web application," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 229–241, 2021, doi: 10.1016/j.ijcce.2021.12.001.
- [17] L. Zhang, Y. Wang, M. Niu, C. Wang, and Z. Wang, "Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study," *Scientific Reports*, vol. 10, no. 1, 2020, doi: 10.1038/s41598-020-61123-x.
- [18] S. Gowthami, V. S. Reddy, and M. R. Ahmed, "Type 2 diabetes mellitus: early detection using machine learning classification," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, pp. 1191–1198, 2023, doi: 10.14569/IJACSA.2023.01406127.
- [19] H. El Bouhissi, R. E. Al-Qutash, A. Ziane, K. Amroun, N. Yaya, and M. Lachi, "Towards diabetes mellitus prediction based on machine-learning," *International Conference on Smart Computing and Application, ICSCA 2023*, 2023, doi: 10.1109/ICSCA57840.2023.10087782.
- [20] 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, doi: 10.1109/ACCESS.2020.2989857.
- [21] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A novel approach for feature selection and classification of diabetes mellitus: Machine learning methods," *Computational Intelligence and Neuroscience*, vol. 2022, 2022, doi: 10.1155/2022/3820360.
- [22] B. Natarajan, S. R. Diwakar, R. Annamalai, R. Bhuvaneswari, and S. S. Husain, "An exploration of the performance using ensemble methods utilizing random forest classifier for diabetes detection," 2023, doi: 10.1109/NMITCON58196.2023.10276348.
- [23] A. A. Alzubaidi, S. M. Halawani, and M. Jarrah, "Towards a stacking ensemble model for predicting diabetes mellitus using combination of machine learning techniques," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 12, pp. 348–358, 2023, doi: 10.14569/IJACSA.2023.0141236.
- [24] A. A. Alzubaidi, S. M. Halawani, and M. Jarrah, "Integrated ensemble model for diabetes mellitus detection," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 4, pp. 223–233, 2024, doi: 10.14569/IJACSA.2024.0150423.




- [25] P. Thotad, G. R. Bharamagoudar, and B. S. Anami, "Predictive analysis of diabetes mellitus using decision tree approach," 2022, doi: 10.1109/ASIANCON55314.2022.9909122.
- [26] M. A. Sahid, M. Ul Hoque-Babar, and M. P. Uddin, "Predictive modeling of multi-class diabetes mellitus using machine learning and filtering iraqi diabetes data dynamics," *PLoS ONE*, vol. 19, no. 5 May, 2024, doi: 10.1371/journal.pone.0300785.
- [27] I. D. Mienye and N. Jere, "A survey of decision trees: concepts, algorithms, and applications," *IEEE Access*, vol. 12, pp. 86716–86727, 2024, doi: 10.1109/ACCESS.2024.3416838.
- [28] Monica and P. Agrawal, "A Survey on Hyperparameter optimization of machine learning models," *2024 2nd Int. Conf. Disruptive Technol. ICDT 2024*, pp. 11–15, 2024, doi: 10.1109/ICDT61202.2024.10489732.
- [29] K. Sagar, S. Imtiyaz, A. Arvinth, A. S. Nagaraju, C. R. Prasad, and P. K. Kumar, "Diabetes prediction using support vector machine," 2024, doi: 10.1109/INCET61516.2024.10593364.

BIOGRAPHIES OF AUTHORS






Farah Najidah Noorizan    was born on 25 September 1999 in Kedah, Malaysia. She pursued her diploma at Politeknik Sultan Salahuddin Abdul Aziz Shah and graduated with a Diploma in Medical Electronic Engineering in 2020. In 2024, she received her degree in electronic engineering from Universiti Tun Hussein Onn Malaysia (UTHM). She is now continuing her studies at UTHM for her Master of Electrical Engineering. Her main research interests are biomedical engineering and artificial intelligence. She can be contacted at email: farahnajidah259@gmail.com.



Nur Anida Jumadi    is an associate professor in the Department of Electronic Engineering, Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia (UTHM). Born in 1983, she obtained her bachelor's degree in electrical engineering from Kolej Universiti Tun Hussein Onn (KUiTTHO) in 2006, followed by a master's degree in medical Electronics and Physics from Queen Mary, University of London, in 2008. She earned her Ph.D. in Electrical, Electronic, and Systems Engineering from Universiti Kebangsaan Malaysia (UKM) in 2015. She specializes in medical diagnosis and prognosis using artificial intelligence, focusing on developing non-invasive medical devices based on optical sensors. She is actively engaged in advancing healthcare technology through innovative research. She can be contacted at email: anida@uthm.edu.my.



Li Mun Ng    received her first degree from Universiti Tun Hussein Onn Malaysia (UTHM), bachelor of Electronic Engineering with Honours, Malaysia, in 2018. She also has a master's degree in electrical engineering from UTHM, Malaysia, in 2021. She is pursuing her Ph.D. in Electrical Engineering in the Department of Electronic Engineering, Faculty of Electrical and Electronic Engineering, UTHM. Her main research interests focus on biomedical engineering, signal processing, artificial intelligence, and the application of fuzzy logic. She can be contacted at email: limun.ng@gmail.com.