JSCDM

JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

Vol. 7 No. 1 (2026) 1-10
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Diabetes Prediction Using The Smote-Cart Framework Model for Imbalanced Data Case

Farah Najidah Noorizan[1], Nur Anida Jumadi[1,2]*, Muhamad Amir Irfan Roslan[1], Li Mun Ng[1], Manveer Pal Singh[3], Yukihiro Ishida[4]

[1] *Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia (UTHM), Batu Pahat, Johor, 86400, MALAYSIA*
[2] *Advanced Medical Imaging and Optics (AdMedic), Universiti Tun Hussein Onn Malaysia (UTHM), Batu Pahat, Johor, 86400, MALAYSIA*
[3] *Putra Specialist Hospital Batu Pahat, ICG Healthcare Group, Batu Pahat, Johor, 83000, MALAYSIA*
[4] *SECOND HEART Inc., 22-9 Megurihara, Shoryuji, Nagaokakyo, Kyoto 617-0836, JAPAN*

*Corresponding Author: anida@uthm.edu.my
DOI: https://doi.org/10.30880/jscdm.2026.07.01.001

Article Info

Received: 4 September 2025
Accepted: 8 December 2025
Available online: 10 February 2026

Keywords

Diabetes mellitus, synthetic minority oversampling technique, classification and regression tree, hyperparameter tuning, evaluation metrics

Abstract

Diabetes mellitus (DM) is described by chronic high blood glucose levels, which can result in long-term damage, dysfunction, and organ failure. As a result of technological advancements, many researchers are employing machine learning to predict diabetes. They collect patients' demographics and health information, organizing them into a dataset. However, in most real-world data, the non-diabetic cases exceed the diabetic cases, contributing to bias in the majority class and resulting in low predictive diabetic cases. Therefore, a Synthetic Minority Oversampling Technique (SMOTE) has been proposed to improve diabetic prediction on the dataset samples before training the Classification and Regression Tree (CART) model. The proposed framework involved the preprocessing step (SMOTE and categorical conversion), CART training, hyperparameter tuning, and evaluation metrics. With a combination of 8 leaf numbers per node, a maximum of 10 splits, and deviance as the split criterion, the model achieves an overall accuracy of 98.72%, a precision of 98.94%, a sensitivity of 98.44%, and an F1-score of 98.67%. In conclusion, the proposed SMOTE-CART framework can effectively address the imbalanced data in a diabetes dataset and improve the accuracy of diabetes prediction.

1. Introduction

Diabetes mellitus (DM) is described by chronic high blood sugar levels caused by a metabolic imbalance in which the pancreas fails to generate an adequate amount of insulin or when the body ineffectively uses the insulin it does produce, leading to long-term damage, malfunction, and organ failure [1-3]. In 2021, diabetes affected approximately 6.7 million people globally, as stated by the International Diabetes Federation. By 2045, this figure is anticipated to increase by 46 per cent [4, 5]. The number of diagnosed cases continued to rise, emphasizing solutions to combat diabetes and limit its effects on world health [6].

Diabetes management is self-managed by the patients themselves through many glucose level assessments, such as insulin injections, which generates hardship for people who require daily measurements [7]. The current conventional glucometers use a needle to get a blood sample from a fingertip and a disposable strip to get the readings [8], [9]. Many researchers are utilizing machine learning to predict diabetes, driven by technological advancements. They collect patients' demographic and health information and organize it into a dataset [10]. However, in most real-world data, the non-diabetic cases exceed the diabetic cases, contributing to bias in the majority class and resulting in poor predictive diabetic cases. Previously, several methods and techniques were employed in imbalanced datasets, including the synthetic minority oversampling technique (SMOTE), adaptive synthetic sampling (ADASYN), random oversampling (ROS), and various hybrid techniques.

In [10, 11], the authors study the application of the Synthetic Minority Oversampling Technique (SMOTE) to enhance minority classification in diabetes databases. The researchers found that this method improved the classification performance, especially in precision and recall, ensuring unbiased results towards the majority class, while the author in [13] studies the combination of multiple imputations by chained equations (MICE) and SMOTE to handle missing values in the dataset. Therefore, the model becomes more stable and reliable in making predictions compared to raw, imbalanced data.

The author in [14] explored the hybrid method to address the imbalance class, for example, ADASYN-SMOTE and SMOTE-SVM, finding that ADASYN-SMOTE performed better with an 87.3% accuracy rate compared to the SMOTE-SVM. On the other hand, [15] solved the unbalanced diabetes data using a hybrid sampling approach combining SMOTE and ENN, along with Random Forest (RF) and Support Vector Machines (SVM). In addition, the author [16] proposed a cluster-based resampling method, such as K-means, Agglomerative Hierarchical Clustering (AHC), and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), combined with resampling SMOTE, outperforming the traditional resampling techniques. Furthermore, the authors in [16-21] applied SMOTE to various classifiers, including XGBoost, RF and Gradient Boosting Trees, to tackle the imbalance issue in the dataset. The research demonstrates consistency of precision, sensitivity, and F1-score with RF, and XGBoost outperforms others in handling diabetes classification.

We have collected data from 210 individuals who underwent health screening checkups at Putra Specialist Hospital in Batu Pahat. However, the data is imbalanced, with 195 samples from non-diabetic subjects and the remaining 15 samples from diabetic subjects. We implemented the classification and regression tree (CART) model to predict diabetes; however, the outcomes were biased due to the insufficient positive and negative instances. This is caused by the small dataset size that can affect generalization; in other words, it may be unable to capture the variability of real-world conditions.

SMOTE is a data augmentation approach that can be applied to enhance a model's generalization within a limited dataset by adding new synthetic samples to expand the dataset. Although SMOTE can pose a risk of overfitting and potentially increase the noise or outliers, it is suitable for this research since our dataset does not involve time-series data, where maintaining temporal relationships is crucial.

Therefore, this research proposes the SMOTE-CART framework to address the trade-off between unbalanced data and boost the accuracy in predicting diabetes. The significant contribution of the SMOTE-CART framework is that it offers a novel hybrid technique to tackle imbalanced issues by producing new artificial samples for the minority class prior to training the CART model. Additionally, we proposed a novel optimization implementation that can enhance CART performance in predicting diabetes. Finally, we also showed the real-world deployment of SMOTE-CART through MATLAB. Compared to the previous hybrid method, our findings provide a well-balanced, efficient, and effective approach to handling imbalanced datasets, thereby improving model performance.

2. Methodology

Figure 1 depicts the general process of the diabetes prediction model. The dataset contains six features: gender, age, BMI, smoking status, the number of hours between the last food intake and the health screening, glucose level, and diabetes prediction as the target. Gender, smoking status, and the number of hours between the last food intake and the health screening are represented as categorical data, with values of 0 and 1. The remaining features are expressed in numerical terms. Ethical approval has been obtained from the Research Ethics Committee (REC) Meeting of UTHM prior to data collection. Then, the subject's informed consent was obtained from 210 subjects (adults aged 18 to 85 years old, healthy or with Type 2 diabetes) recruited at Putra Specialist Hospital, Batu Pahat. The diabetes status, which is the target feature, is divided into two classes: Class 0 for non-diabetes and Class 1 for diabetes. The process of diabetes prediction started with the preprocessing step and train-test splitting. The selected machine learning model in this research is the CART model, which implements several hyperparameter tuning settings, for example, the number of leaves per node, the maximum number of splits, and the split criterion. The confusion matrix, accuracy, precision, recall, F1-score, and ROC-AUC curve will be utilized to assess the model's performance. Lastly, a model deployment was performed using MATLAB software.

J. of Soft Computing and Data Mining Vol. 7 No. 1 (2026) p. 1-10

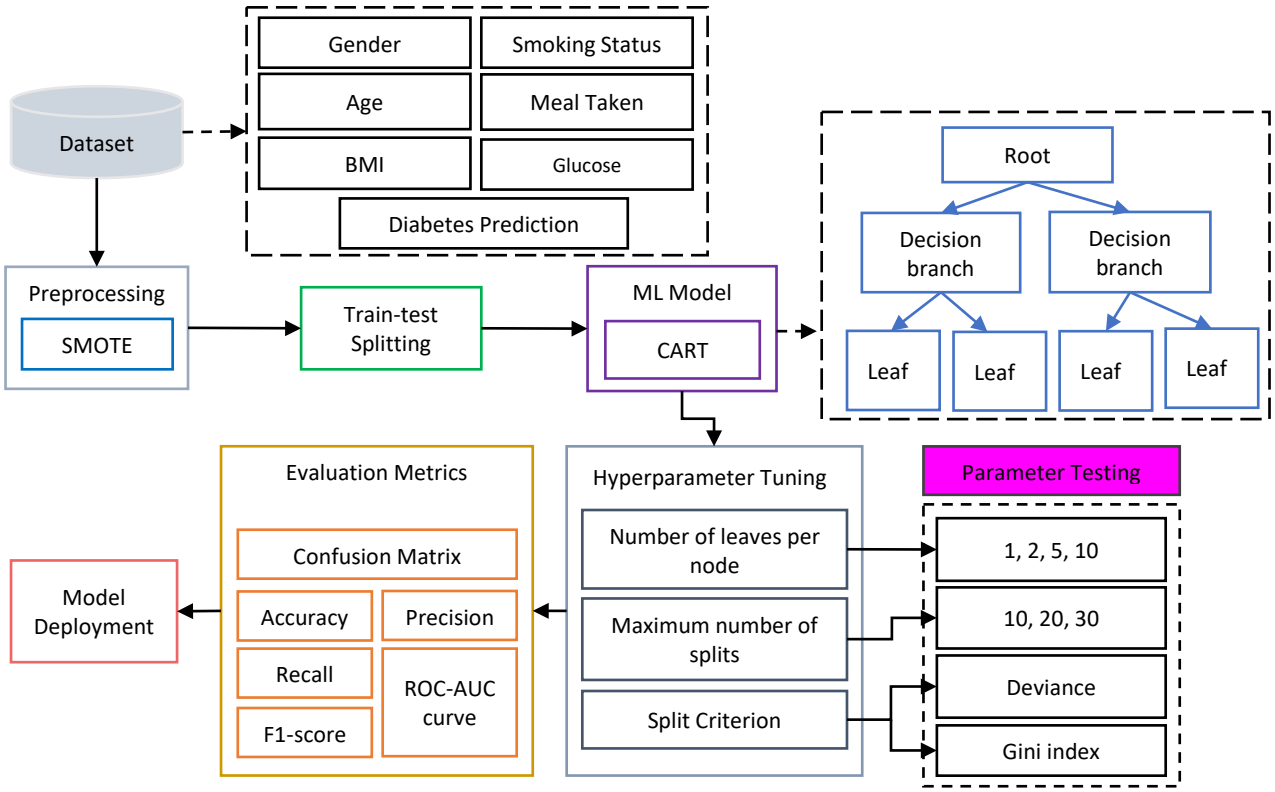

Fig. 1 *Block diagram for diabetes prediction*

2.1 SMOTE and Train-Test Split

SMOTE is used in this study because the classification of diabetes and non-diabetes samples is skewed. It is an oversampling technique which generates artificial samples for the minority class, rather than repeating existing data. A ratio of 80:20 is used to evaluate the model by separating the dataset into two parts. The purpose of training is to carry out the instructed task, while testing assesses how the final trained model performs on the hidden data to prevent overfitting. Twenty per cent of the data will be randomly selected for testing, while the remaining 80% will be used for training.

Table 1 presents the pseudocode for the steps used to address the unbalanced issue using the SMOTE technique. The process begins with loading the dataset containing six input features and a target feature. Preprocessing then follows, where the samples are balanced using the SMOTE approach. The augmented dataset is subsequently divided into two subsets before being trained with hyperparameter tuning via the grid search method. Lastly, six parameters will be used to assess the model's performance, such as the confusion matrix, accuracy, precision, recall, F1-score, and ROC-AUC curve.

Table 1 *Pseudocode for imbalanced data using SMOTE*

Pseudocode for Imbalanced Data using SMOTE
1. Load dataset of 6 input features (age, gender, BMI, glucose, meal timing, smoking status) and target features (Diabetes)
2. Balance the data using SMOTE
3. Split the data into 80:20 train-test
4. Training the CART algorithm using the grid search method of hyperparameter tuning
5. Performance of evaluation metrics
6. End

2.2 Classification and Regression Tree (CART) and Hyperparameter Tuning

CART is an algorithm that combines decision trees and regression analysis to solve classification and regression problems [23]. Gini index and deviance are the criteria for splitting. The Gini index refers to the statistic that determines the purity of data in a decision node, and deviance is calculated using log-likelihood to assess the quality of classification models. Hyperparameter tuning is a parameter in machine learning that must be set before training the model [24]. In CART, the hyperparameters involved in tuning are the number of leaves per node, the maximum number of splits, and the split criterion. The number of leaves per node refers to the minimum number of data points in a leaf branch to prevent overfitting by forbidding the tree from splitting further than the selected value, and the maximum number of splits constrains the tree's growth, minimizing complexity and preventing overfitting by avoiding needless branching.

2.3 Performance Metrics

Accuracy, precision, sensitivity, F1-score, and the ROC-AUC curve are used to evaluate the model's performance. These metrics have been presented in equations (1) through (5), respectively. Table 2 illustrates the binary classification's confusion matrix structure, which displays the position of the number of true positives, true negatives, false positives, and false negatives. Besides that, variable a represents the tested_negative, which refers to the samples in the negative class, and variable b represents all positive samples, referred to as tested_positive. From the structure, a true negative is obtained when the model correctly classifies the negative samples, but if it fails, the samples are categorized as false positives. In contrast, when a model successfully identifies a positive sample, a true positive is obtained, and if it incorrectly predicts a negative sample, the sample is counted as a false negative.

Table 2 *Confusion matrix layout* [25]

Confusion Matrix Structure			
Total number of samples		Predicted Class	
		No a=tested_negative	Yes b=tested_positive
Actual Class	No a=tested_negative	True Negative	False Positive
	Yes b=tested_positive	False Negative	True Positive

Accuracy is the ratio of the summation of true positives and true negatives to the total number of cases. The proportion of the expected positive instances to the total predicted positive instances is referred to as precision. Recall assesses the proportion of positively predicted values among all actual positive samples in the dataset. In addition, the F1-score is used to measure the overall model's achievement by weighting the precision and recall, while the AUC value serves as an indicator for a binary classification model that can differentiate positive and negative classes [26].

All the metrics evaluated by TP as True Positive, TN as true Negative, FP as the False Positive, FN as the False Negative, R_i is the rate of the i_{th} data, and I_f and I_I are the negative and positive data.

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{FI} - \text{score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{4}$$

$$AUC = \frac{\sum R_i(I_I) - I_I(I_I + 1)/2}{I_I + I_f} \tag{5}$$

J. of Soft Computing and Data Mining Vol. 7 No. 1 (2026) p. 1-10

2.4 Model Deployment

Model deployment in machine learning involves integrating the trained model into the real world for user applications through a user interface. Figure 2 represents the user interface of model deployment using MATLAB software. By entering user inputs such as gender, age, BMI, glucose level, the number of hours between the last food intake and the health screening, and smoking status, the model will predict the class of diabetes.

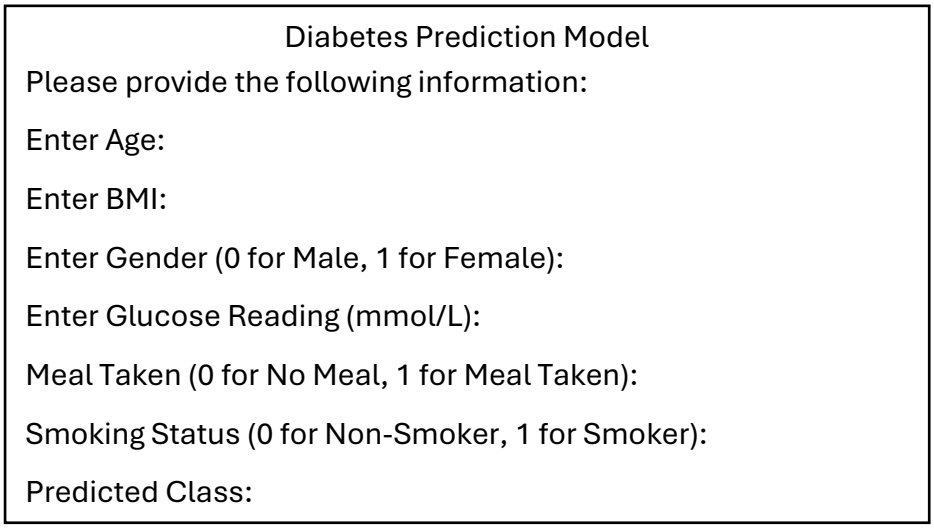

Fig. 2 *User interface of the model deployment*

3. Results and Discussion

Section 3 unveils the findings and their corresponding discussion. Section 3.1 covered four divisions. Firstly, subsection 3.1.1 explains the best selected hyperparameter tuning settings that improve the model's accuracy and overall performance. Then, subsection 3.1.2 explained the potential of the developed model through the evaluation metrics. Subsection 3.1.3 benchmarking SMOTE-CART with Random Forest and CART models, highlighting the results obtained by the selected algorithm compared to others. Finally, subsection 3.1.4 presents the outcomes of model deployment. Additionally, Section 3.2 presents a discussion, and Section 3.3 highlights the limitations of this study.

3.1 Results

This study investigates the effectiveness of the developed SMOTE-CART model in predicting diabetes in imbalanced data conditions. In contrast to the previous study, which demonstrated only modest performance, this work aims to establish a more robust and successful hybrid model for diabetes prediction.

3.1.1 Hyperparameter Tuning

Table 3 presents the optimal hyperparameter tuning settings for the prediction model. A value of 8 is selected for the number of leaves per node and 10 for the maximum number of splits. Then, the chosen split criterion is deviance due to its functionality in imbalanced datasets, as the current dataset has a higher proportion of non-diabetic samples. The combinations of tuning settings increased accuracy and other performance parameters.

Table 3 *The optimal hyperparameter tuning settings*

The Best Hyperparameter Tuning Setting	
Number of leaves per node	8
Maximum number of splits	10
Split criterion	Deviance

3.1.2 Performance of Evaluation Metrics

The experimental results, obtained after training the dataset using the SMOTE and CART framework model, are presented in this section. Figure 3 illustrates the confusion matrix, where the rows represent the actual classes and the columns represent the predicted classes. For Class 0, the model successfully identified 31 samples as Class 0, with one miscategorization as Class 1. Besides, 46 samples of Class 1 were correctly classified with no misclassification. These results demonstrate that the hybrid SMOTE-CART model is a robust, reliable, and valid approach for distinguishing between diabetes and non-diabetes samples. Moreover, the color density in the diagonal pattern suggests that the model is effective and has a strong capacity to distinguish between classes.

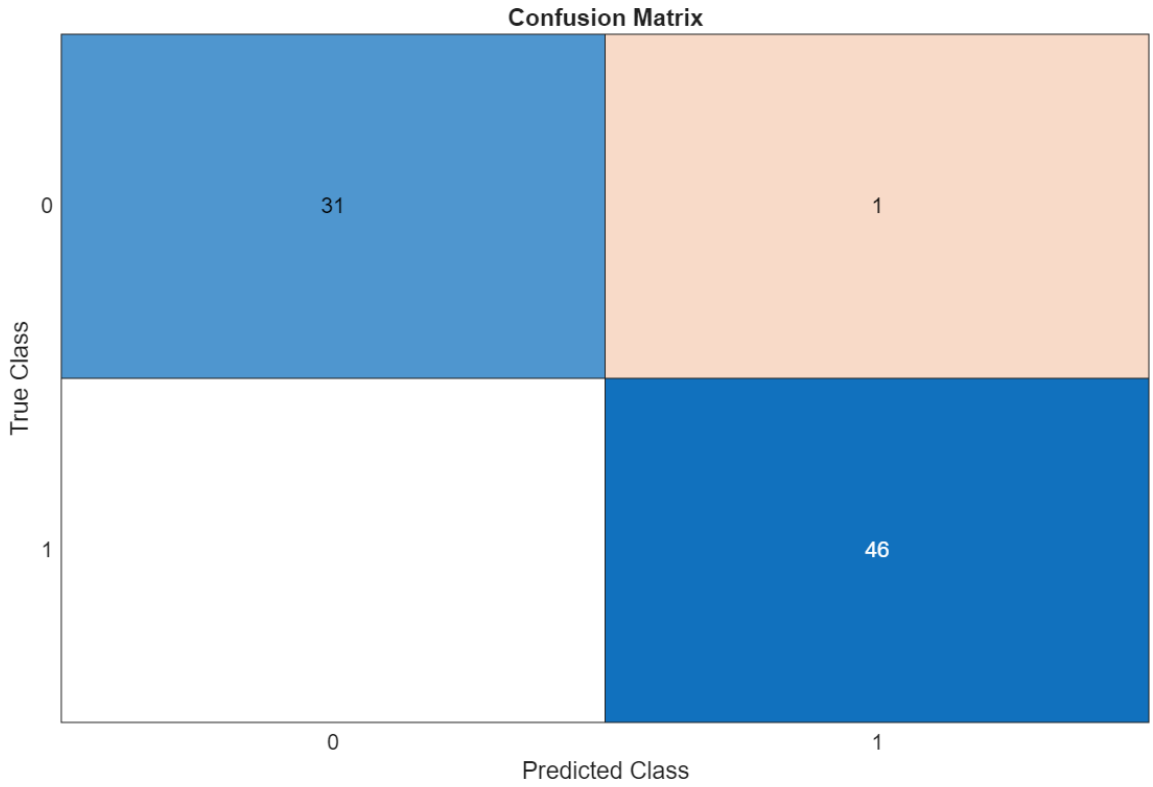

Fig. 3 *Confusion matrix*

The result of the SMOTE-CART performance is presented in Table 4. First, Class 0 achieved an accuracy of 96.88%. indicating that all samples in this class were correctly classified. Then, the 100% precision indicates that all non-diabetic samples were correctly predicted. Then, recall was 96.88%, showing that the ground labels were successfully identified, and the F1-score was 98.41%, reflecting strong consistency in this class.

For Class 1, 100% accuracy was achieved, indicating that all diabetic samples were recognized accurately. A precision of 97.87% indicates that most samples are correctly predicted as Class 1, while the actual instances are successfully identified, achieving 100% recall. A value of 98.92% for the F1-score rate indicates an excellent balance between precision and sensitivity in this class, which suggests that this model is effective in predicting diabetes.

A 98.72% overall accuracy signifies an almost perfect classification of both classes. Furthermore, with precision, recall, and F1-scores of approximately 98.94%, 98.44%, and 98.67%, respectively, indicating a well-consistent ability to differentiate between diabetic and non-diabetic data. In general, the outcome from Class 0 indicates that all the cases of Class 0 (non-diabetic cases) are precisely identified, whereas in Class 1, all actual cases were successfully recognized.

Table 4 *Performance of evaluation metrics*

Class	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
0	96.88	100	96.88	98.41
1	100	97.87	100	98.92
Overall (Class 0 + Class 1)	98.72	98.94	98.44	98.67

J. of Soft Computing and Data Mining Vol. 7 No. 1 (2026) p. 1-10

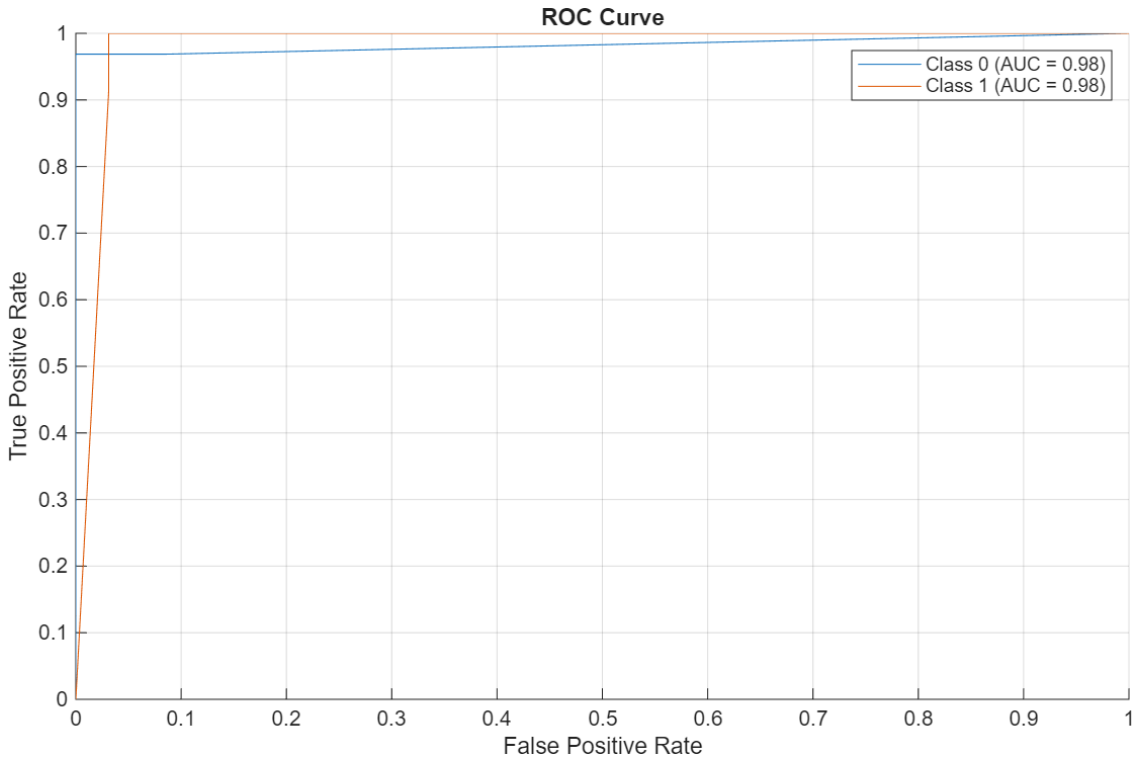

Fig. 4 *ROC curve*

Figure 4 illustrates the ROC curve for the model, where the horizontal line represents the false positive rate and the vertical line represents the true positive rate, plotted under various probabilistic thresholds. This curve explains the ability to discriminate between the two classes across all possible thresholds in binary classification. Furthermore, the Under the Curve (AUC) values for both classes are 0.98. In practice, an AUC value close to 1 indicates perfect differentiation ability, meaning the model can successfully distinguish between individuals with diabetes and those without. Additionally, the equality values of both classes demonstrate that the model performs well in both categories, indicating a balanced predictive capability.

3.1.3 Benchmarking SMOTE-CART with Random Forest and CART Models

The comparison of SMOTE-CART, Random Forest, and CART performance is presented in Table 5. These three algorithms were trained using the same settings: the grid search method and an 80:20 split ratio. Random Forest achieved an accuracy of 88%, with a precision of 63%, a recall of 60%, an F1-score of 61%, and an AUC value of 0.88. This result indicates that RF is one of the reliable models. Additionally, CART achieved 100% accuracy, including all other parameters. However, the AUC could not be executed because of insufficient instances. Lastly, SMOTE-CART achieved higher accuracy than Random Forest, with 98.72%, a precision of 98.94%, a recall of 98.44%, an F1-score of 98.67%, and an AUC curve of 0.98, indicating excellent model performance.

From these findings, the optimal performance of the CART model may indicate overfitting due to insufficient sample size. Meanwhile, the Random Forest exhibits the lowest sensitivity, yet with a good AUC, indicating a reasonable overall performance. Undoubtedly, SMOTE-CART produces a well-proportioned performance with higher accuracy, precision, and recall, making it the most generalizable model among the three algorithms.

Table 5 *Comparison between SMOTE-CART and other models*

Algorithms	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-curve
Random Forest	88	63	60	61	0.88
CART	100	100	100	100	No AUC can be executed because of insufficient positive and negative instances
SMOTE-CART	**98.72**	**98.94**	**98.44**	**98.67**	**0.98**

3.1.4 Model Deployment Outcomes

Figure 5 illustrates the deployment outcomes after simulated input features from a 20-year-old non-diabetic patient. This patient is a female with a body mass index of 20 and a glucose reading of 5 mmol/L, indicating she is healthy. The resulting predicted class is 0, highlighting that this person is a non-diabetic patient. This situation can be proven when the normal glucose level after a meal is less than 7.8 mmol/L, and a BMI of 20 is considered normal.

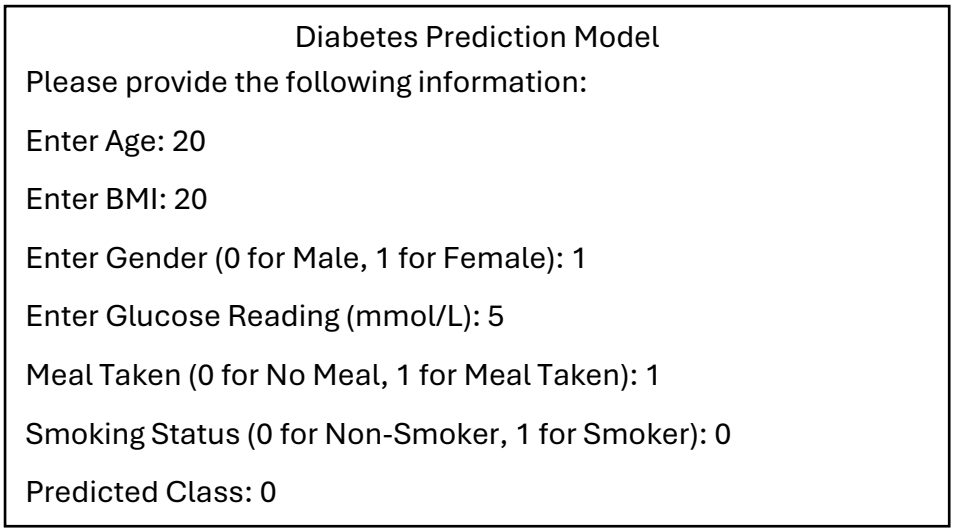

Fig. 5 *Deployment outcomes*

As shown in Figure 5, the deployment outcomes demonstrate the successful implementation of the deployment using MATLAB. The model deployment can be extended using MATLAB App Designer. It can be computerized as a stand-alone application that can be executed on any laptop or PC. Alternatively, using Python or cloud-based deployment can also be considered for model deployment. However, these applications require additional configuration and lack built-in capabilities, and cloud-based deployment relies on the Internet, which might lead to delays and security risks.

3.2 Discussion

This research successfully developed a diabetes prediction model using the Classification and Regression Tree (CART) model, combined with the Synthetic Minority Oversampling Technique (SMOTE), to address the issue of unbalanced data in both diabetes and non-diabetes samples. The model achieved an overall accuracy of 98.72%, outperforming other algorithms, such as Random Forest and traditional CART. Each class 0 and 1 has a higher precision, recall, and F1-score of approximately 98%, respectively. Furthermore, the confusion matrix also indicates that only Class 0 has one misclassification, being classified as Class 1. In addition to AUC values of 0.98 for both classes, this indicates effective discrimination between the classes.

The previous study [17] also applied the SMOTE algorithm for balancing the dataset, but not highlighting the hyperparameter tuning settings, which resulted in a moderate model performance. Therefore, in this study, all combinations of actions, such as data preprocessing, optimizing the number of leaves per node, and the maximum number of splits, as well as determining the split criterion, contributed to producing a model that can achieve high accuracy and effectively detect a person's diabetes status.

However, the dataset used in this study is relatively limited, comprising only 210 samples. Although SMOTE was implemented to lessen the class imbalance, the small number of diabetic cases (15 samples) may introduce unfairness to the model's learning process, limiting its generalizability to more extensive or diverse populations. This study explored a prediction model based on various data types, including demographic information (age and gender), anthropometric data (BMI), and lifestyle-related factors (smoking and meal). By combining these varied inputs, the model demonstrated strong predictive performance in identifying individuals with diabetes.

3.3 Limitations of the Study

One of the study's limitations is that it only uses features related to demographics, anthropometrics, and lifestyle. The exclusion of other clinical criteria, such as cholesterol levels, high blood pressure, and family history of

disease, can impair prediction accuracy. Furthermore, the sample was collected with the cooperation of a single healthcare facility and focused on a few specific areas, which limited the sample's population and diversity.

4. Conclusion

In conclusion, this study demonstrates that the SMOTE-CART hybrid model, which incorporates user health information, is an effective model for predicting diabetes status. This model, which includes numerous settings such as data preprocessing, optimal hyperparameter tuning, and the use of deviance for splitting, performed very well and achieved an excellent accuracy rate.

This project can be improved in the future by transforming the deployment interface model into a graphical user interface (GUI), which would improve usability and make it more visually appealing. This GUI application allows users to enter personal health information more rapidly, making data prediction more efficient and straightforward. Besides considering a more varied dataset source from multiple healthcare facilities to enhance the model's robustness, future research should also examine the use of more diverse data, such as family history and other risk factors, to improve the model's framework accuracy and generalizability.

Acknowledgement

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

Declaration of AI Use in Manuscript Preparation

The authors use Grammarly to assist in grammar checking, paraphrasing, and language editing. All content generated was reviewed and verified by the authors, who take full responsibility for the final submission.

Conflict of Interest

The manuscript has not been published elsewhere and is not under consideration by other journals. All authors have approved the review, agree with its submission and declare no conflict of interest on the manuscript.

Author Contribution

Farah Najidah: ***Data Collection, Methodology, Software, Writing-Original Draft Preparation****; Nur Anida:* ***Supervision, Writing-Reviewing and Editing, Result Validation****; Muhamad Amir Irfan:* ***Assisting in Data Collection****; Ng Li Mun:* ***Proofreading****; Manveer Pal Singh:* ***Assisting in Data Collection and Medical Consultation****; Yukihiro Ishida:* ***Technical Guidance****.*

References

[1] Mohshim, S. A., Nasiruddin, N. F. N., Zakaria, Z., Desa, H. M., Mohshim, D. F., & Fadzir, M. F. (2023). Non-invasive blood glucose monitor using Arduino for clinical use. *2023 Int. Conf. Eng. Technol. Technopreneurship, ICE2T 2023*, 219–224. https://doi.org/10.1109/ICE2T58637.2023.10540514

[2] Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Comput. Methods Programs Biomed., 220*, 106773. https://doi.org/10.1016/j.cmpb.2022.106773

[3] Chitradevi, B., Supriya, Chandra, N. S., Chitradevi, T. N., & Alabdeli, H. (2024). Diabetes mellitus prediction and classification using firefly optimization based support vector machine. *Int. Conf. Distrib. Comput. Optim. Tech. ICDCOT 2024*, 1–5. https://doi.org/10.1109/ICDCOT61034.2024.10515397

[4] Naresh, M., Nagaraju, V. S., Kollem, S., Kumar, J., & Peddakrishna, S. (2024). Non-invasive glucose prediction and classification using NIR technology with machine learning. *Heliyon, 10*(7), e28720. https://doi.org/10.1016/j.heliyon.2024.e28720

[5] Mansour, E., et al. (2023). Review on non-invasive electromagnetic approaches for blood glucose monitoring using machine learning. *Proc. 11th Int. Japan-Africa Conf. Electron. Commun. Comput. JAC-ECC 2023*, 273–276. https://doi.org/10.1109/JAC-ECC61002.2023.10479620

[6] Gowthami, S., Reddy, V. S., & Ahmed, M. R. (2023). Type 2 diabetes mellitus: Early detection using machine learning classification. *Int. J. Adv. Comput. Sci. Appl., 14*(6), 1191–1198. https://doi.org/10.14569/IJACSA.2023.01406127

[7] Tejedor, M., Woldaregay, A. Z., & Godtliebsen, F. (2020). Reinforcement learning application in diabetes blood glucose control: A systematic review. *Artif. Intell. Med., 104*, 101836. https://doi.org/10.1016/j.artmed.2020.101836

[8] Tang, L., Chang, S. J., Chen, C. J., & Liu, J. T. (2020). Non-invasive blood glucose monitoring technology: A review. *Sensors (Switzerland), 20*(23), 6925. https://doi.org/10.3390/s20236925

[9] Bolla, A. S., & Priefer, R. (2020). Blood glucose monitoring—An overview of current and future non-invasive devices. *Diabetes Metab. Syndr. Clin. Res. Rev., 14*(5), 739–751. https://doi.org/10.1016/j.dsx.2020.05.016

[10] Liu, K., et al. (2023). Machine learning models for blood glucose level prediction in patients with diabetes mellitus: Systematic review and network meta-analysis. *JMIR Med. Informatics, 11*(1). https://doi.org/10.2196/47833

[11] Sampath, P., et al. (2024). Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique. *Sci. Rep., 14*(1), 1–15. https://doi.org/10.1038/s41598-024-78519-8

[12] Talebi Moghaddam, M., et al. (2024). Predicting diabetes in adults: Identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm. *BMC Med. Res. Methodol., 24*(1), 220. https://doi.org/10.1186/s12874-024-02341-z

[13] Abdullah, M. N., & Wah, Y. B. (2022). Improving diabetes mellitus prediction with MICE and SMOTE for imbalanced data. *2022 3rd Int. Conf. Artif. Intell. Data Sci. Championing Innov. Artif. Intell. Data Sci. Sustain. Futur. AiDAS 2022 - Proc.*, 209–214. https://doi.org/10.1109/AiDAS56890.2022.9918773

[14] Ramadhan, N. G. (2021). Comparative analysis of ADASYN-SVM and SMOTE-SVM methods on the detection of type 2 diabetes mellitus. *Sci. J. Informatics, 8*(2), 276–282. https://doi.org/10.15294/sji.v8i2.32484

[15] Hairani, H., & Priyanto, D. (2023). A new approach of hybrid sampling SMOTE and ENN to the accuracy of machine learning methods on unbalanced diabetes disease data. *Int. J. Adv. Comput. Sci. Appl., 14*(8), 585–590. https://doi.org/10.14569/IJACSA.2023.0140864

[16] Park, Y. J., & Cheng, K. Y. (2024). A cluster impurity-based hybrid resampling for imbalanced classification problems. *Appl. Intell., 54*(20), 9671–9684. https://doi.org/10.1007/s10489-024-05644-2

[17] Sarayu, M. K., Bhanu, S. A., Deekshitha, K., Meghana, M., & Joseph, I. T. (2024). Diabetes prediction using SMOTE and machine learning. *Proc. - 2024 2nd Int. Conf. Inven. Comput. Informatics, ICICI 2024*, 15–20. https://doi.org/10.1109/ICICI62254.2024.00011

[18] Kaur, A., Gill, K. S., Chauhan, R., & Pokhariya, H. S. (2024). Hyperparameter tuning and SMOTE utilization on a synergistic approach to predicting diabetes. *2024 4th Asian Conf. Innov. Technol. ASIANCON 2024*, 1–4. https://doi.org/10.1109/ASIANCON62057.2024.10837781

[19] Mahmud, S., Islam, B. U., Anik, N. H., & Ghosh, T. (2024). Diabetes prediction: A comparative analysis of machine learning algorithms with SMOTE. *2024 IEEE Conf. Comput. Appl. Syst. COMPAS 2024*, 1–6. https://doi.org/10.1109/COMPAS60761.2024.10796405

[20] Gill, K. S., Anand, V., Upadhyay, D., & Dangi, S. (2024). Diabetes classification using XGBoost classification techniques through machine learning based SMOTE analysis. *2024 3rd Int. Conf. Innov. Technol. INOCON 2024*, 1–4. https://doi.org/10.1109/INOCON60754.2024.10512046

[21] Yakshit, Kaur, G., Kaur, V., Sharma, Y., & Bansal, V. (2022). Analyzing various machine learning algorithms with SMOTE and ADASYN for image classification having imbalanced data. *Proc. 2022 IEEE Int. Conf. Curr. Dev. Eng. Technol. CCET 2022*. https://doi.org/10.1109/CCET56606.2022.10080783

[22] Brandt, J., & Lanzén, E. (2020). *A comparative review of SMOTE and ADASYN in imbalanced data classification* (p. 42). https://www.diva-portal.org/smash/record.jsf?pid=diva2:1519153

[23] Mienye, I. D., & Jere, N. (2024). A survey of decision trees: Concepts, algorithms, and applications. *IEEE Access, 12*, 86716–86727. https://doi.org/10.1109/ACCESS.2024.3416838

[24] Monica, & Agrawal, P. (2024). A survey on hyperparameter optimization of machine learning models. *2024 2nd Int. Conf. Disruptive Technol. ICDT 2024*, 11–15. https://doi.org/10.1109/ICDT61202.2024.10489732

[25] Manna, S. (2022). Small sample estimation of classification metrics. *2022 Int. Conf. Interdiscip. Res. Technol. Manag. IRTM 2022 - Proc.*, 1–3. https://doi.org/10.1109/IRTM54583.2022.9791645

[26] Mahajan, S., Sarangi, P. K., Sahoo, A. K., & Rohra, M. (2023). Diabetes mellitus prediction using supervised machine learning techniques. *2023 Int. Conf. Adv. Comput. Comput. Technol. InCACCT 2023*, 587–592. https://doi.org/10.1109/InCACCT57535.2023.10141734