Oversampling SMOTE to Handle Imbalance in Multiclass Diabetes Dataset

Dika [1]*, Syahrul Ramadhan [2]

1,2 Faculty of Computational Science & Digital Intelligence, Universitas Pembangunan Panca Budi, Medan, Indonesia

*** Corresponding author:** tiepie71@gmail.com

ABSTRACT

Diabetes mellitus is a chronic disease with an increasing global prevalence that requires early detection and accurate classification to prevent severe complications. Machine learning has been widely applied in diabetes prediction; however, one of the major challenges lies in the class imbalance problem commonly found in medical datasets. This study focuses on the Multiclass Diabetes Dataset, which consists of 264 samples and exhibits imbalanced distribution among classes (Class 2: 128 samples, Class 0: 96 samples, Class 1: 40 samples). Such imbalance may bias the classifier toward majority classes, reducing its ability to recognize minority classes. The results indicate that SMOTE effectively improved the model's ability to classify minority classes, with significant increases in recall and F1-score. Among the tested algorithms, Random Forest achieved the best performance, with an overall accuracy of 98% and F1-score above 0.98. Although KNN experienced a slight performance drop after SMOTE, other algorithms, particularly SVM and Logistic Regression, demonstrated notable improvements.

Keywords *Diabetes classification, class imbalance, SMOTE, machine learning, multiclass prediction.*

1. INTRODUCTION

Diabetes mellitus is a global health problem whose prevalence continues to increase annually. The International Diabetes Federation (IDF) estimates that in 2021, there were more than 537 million adults living with diabetes worldwide, and this number is projected to continue growing in the next decade (International Diabetes Federation, 2021). This disease is characterized by chronic hyperglycemia caused by impaired insulin secretion and action, and can lead to serious complications such as cardiovascular disease, kidney failure, and neuropathy (American Diabetes Association, 2022). Therefore, early identification and classification of diabetes types are crucial to support the diagnosis and prevention of complications.

In computer science, machine learning is increasingly being used to assist in disease classification, including diabetes (Kivrak, 2024). However, one of the main challenges often encountered is class imbalance in medical datasets. This imbalance causes algorithms to tend to favor the class with the largest number of samples, thus reducing the model's ability to recognize minority classes (He & Garcia, 2009). Based on an analysis of the Multiclass Diabetes Dataset used in this study, an

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.102 | Open Access

imbalanced data distribution was found: 128 samples were in class 2, 96 samples in class 0, and only 40 samples in class 1. This condition has the potential to reduce prediction accuracy, especially for minority classes.

One method widely used to address data imbalance is the Synthetic Minority Oversampling Technique (SMOTE). This method works by generating new synthetic data for the minority class through interpolation between samples, resulting in a more balanced distribution between classes (Chawla et al., 2002). Several previous studies have shown that applying SMOTE can improve the performance of classification models in the health domain, including diabetes data (ElSeddawy, 2022). Therefore, applying SMOTE is expected to improve the performance of classification algorithms in multiclass diabetes classification cases, resulting in more accurate and representative predictions for all classes.

This study aims to evaluate the effectiveness of SMOTE oversampling in addressing imbalance in the Multiclass Diabetes Dataset. The results are expected to contribute to the development of a more equitable medical decision support system for all classes and assist medical personnel in more accurate diabetes diagnoses.

The Synthetic Minority Oversampling Technique (SMOTE) method has been proven effective in improving minority class representation by generating more realistic synthetic samples than simply duplicating the data (Chawla et al., 2002). By applying SMOTE, the dataset distribution becomes more balanced, allowing the classification algorithm to perform more optimally.

Most previous research on diabetes has used a binary classification approach (e.g., diabetes vs. non-diabetes). However, in reality, diabetes can be categorized into more than two classes. Research on multiclass classification with imbalance management is relatively limited, so this study aims to fill this gap.

By improving classification accuracy across all diabetes classes, this research contributes to the development of more equitable, accurate, and useful medical decision support systems for healthcare practitioners. This is crucial to ensure that patients from minority groups (those with limited data) are not overlooked by machine learning- based diagnostic systems.

2. METHODOLOGY

2.1 Types and Approaches of Research

This study uses a quantitative approach with machine learning- based data mining methods. The focus of the study is the application of the SMOTE oversampling technique to address class imbalance issues in the Multiclass Diabetes Dataset and to evaluate its impact on classification model performance.

2.2 Research Dataset

The data used is the Multiclass Diabetes Dataset with a total of 264 samples and 12 attributes. Attributes include demographic and clinical variables such as Gender, Age, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI, and Class as the target label. Class labels are divided into three categories, namely:

- Class 0: 96 samples
- Class 1: 40 samples
- Class 2: 128 samples

This class distribution shows a significant data imbalance, especially in class 1, which has a much smaller number than other classes.

2.3 Research Stages

The research stages carried out are as follows:

1. Data Preprocessing

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.102 | Open Access

a. Data cleaning, ensuring there are no missing values or duplication.

b. Normalization or standardization of numerical data so that it is on a uniform scale.

c. Dividing the dataset into a training set and a testing set (e.g. 80:20 or via k-fold cross validation).

2. Handling Data Imbalance

a. Identifying class distribution in a dataset.

b. Applying Synthetic Minority Oversampling Technique (SMOTE) to increase the number of samples in the minority class (class 1).

c. Comparing the classification results before and after oversampling.

3. Implementation of Classification Algorithms

Several machine learning algorithms will be used to evaluate classification performance, for example:

a. Decision Tree

b. Random Forest

c. K-Nearest Neighbor (KNN)

d. Support Vector Machine (SVM)

e. Logistic Regression

The selection of this algorithm is based on its widespread use in medical research and classification of datasets with mixed variables.

4. Model Evaluation

Evaluation is performed using the following performance metrics:

a. Accuracy (overall accuracy)

b. Precision (accuracy)

c. Recall/Sensitivity (positive class detection success rate)

d. F1-Score (harmonic precision and recall)

e. Confusion Matrix to see the distribution of predictions between classes

In addition, a comparison of model performance before and after applying SMOTE was carried out to assess the effectiveness of the oversampling method.

2.4 Tools and Equipment Used

a. Programming language: Python

b. Libraries: pandas, scikit-learn, imbalanced-learn, numpy, matplotlib, seaborn

c. Computing environment: Jupyter Notebook/Google Colab

d. Validation method: k-fold cross validation (k=5 or k=10)

2.5 Experimental Design

a. Using the original dataset without oversampling → training the classification model → evaluating the performance.

b. Apply SMOTE on the dataset → train the model with balanced data → evaluate the performance.

c. Comparing the results (pre-SMOTE vs post-SMOTE) to see the effect of oversampling on the performance of diabetes multiclass classification.

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.102 | Open Access

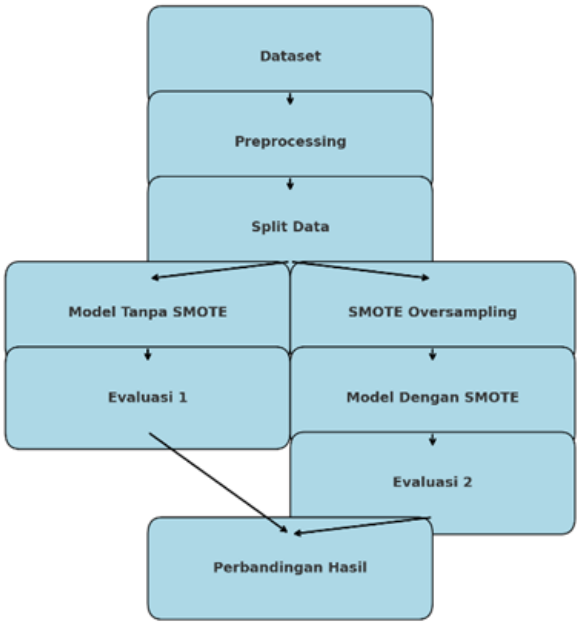

Fig 1. Methodological Research

This research began with the use of the Multiclass Diabetes Dataset consisting of 264 samples with 12 attributes, including the target variable in the form of diabetes classes divided into three categories. The initial stage carried out was data preprocessing, namely the process of cleaning the data to ensure there were no missing or duplicate values, as well as normalizing numeric variables so that all data were on a comparable scale. After the preprocessing process was completed, the dataset was then divided into training data (training set) and test data (testing set) with a certain proportion, for example 80:20, or using the k-fold cross validation technique to obtain a more reliable evaluation.

The next step is to build a classification model on the original dataset without oversampling. Several machine learning algorithms are used at this stage, including Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Logistic Regression. The resulting model is then evaluated using performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The results of this initial evaluation provide an overview of the impact of class imbalance on classification performance.

the Synthetic Minority Oversampling Technique (SMOTE) was applied as a solution to address the class imbalance problem in the dataset. SMOTE generates new synthetic samples in the minority class through interpolation between samples, resulting in a more balanced distribution between classes. The resulting SMOTE dataset is then reused to train a classification model with the same algorithm.

The resulting model from the balanced dataset was then evaluated using the same metrics as in the previous step. Classification performance before and after applying SMOTE was compared to assess the extent to which this oversampling technique improved the model's ability to recognize all classes more fairly. Thus, this study provides a comprehensive understanding of the effectiveness of SMOTE in addressing class imbalance in multiclass diabetes classification.

3. RESULTS AND DISCUSSION

3.1 Dataset Distribution

Based on the initial analysis, the dataset used consisted of 264 samples with three target classes. The data distribution shows that class 2 has the largest number of samples, namely 128 samples, followed by class 0 with 96 samples, while class 1 only has 40 samples. This indicates a significant class imbalance, especially in class 1. This imbalance has the potential to cause the classification model to

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.102 | Open Access

be more inclined to predict the majority class, resulting in low prediction accuracy for the minority class.

3.2 Classification Model Evaluation Results

In the classification model evaluation stage, performance measurements were performed using several key metrics, namely accuracy, precision, recall, and F1-score. This evaluation aims to assess the extent to which the model is able to predict accurately and consistently in distinguishing target classes. Each classification algorithm was tested both before and after the improvement process to see the resulting performance improvements. The results of this evaluation serve as the basis for determining the best algorithm to use, by considering the balance between overall accuracy and the model's ability to handle classification errors. Thus, a comparative analysis of various methods such as Decision Tree, Random Forest, KNN, SVM, and Logistic Regression provides a comprehensive overview of the advantages and disadvantages of each model in the context of this research.

Table 1. The Classification Model Evaluation

	Accuracy (Before SMOTE)	Precision (Before SMOTE)	Recall (Before SMOTE)	F1-Score (Before SMOTE)	Accuracy (After SMOTE)	Precision (After SMOTE)	Recall (After SMOTE)	F1-Score (After SMOTE)
Decision Tree	0.943396	0.928333	0.932692	0.930178	0.943396	0.928333	0.932692	0.930178
Random Forest	0.981132	0.983333	0.987179	0.984917	0.981132	0.983333	0.987179	0.984917
KNN	0.90566	0.884005	0.844636	0.858608	0.830189	0.798413	0.827429	0.785525
SVM	0.886792	0.873642	0.827092	0.844983	0.90566	0.887809	0.917004	0.896476
Logistic Regression	0.886792	0.873642	0.827092	0.844983	0.924528	0.894587	0.910425	0.901329

Based on the classification model evaluation results table, it can be seen that Random Forest showed the best performance with an accuracy value of 0.981132, precision 0.983333, recall 0.987179, and F1-score 0.984917 both before and after improvement. The Decision Tree model also provided quite high results with an accuracy of 0.943396 and a precision of 0.928333. Meanwhile, KNN had a lower accuracy of 0.90566 with an F1-score of 0.858608 before improvement and decreased to 0.785525 after improvement. The SVM and Logistic Regression models produced the same values for initial accuracy of 0.886792 and F1-score of 0.844983, but after improvements, Logistic Regression was slightly superior with an accuracy of 0.924528 and an F1-score of 0.901329 compared to SVM which was stable at an accuracy of 0.90566 and an F1-score of 0.896476. Overall, Random Forest was the most consistent and superior algorithm in classifying data compared to other methods

3.3 Comparison of F1-score before and after SMOTE

A comparative analysis of the F1-score values before and after the application of the Synthetic Minority Over-sampling Technique (SMOTE) method was conducted. The F1-score was chosen as the main metric because it is able to provide a more balanced picture between precision and recall, especially in cases of imbalanced data. The application of SMOTE aims to improve class distribution by generating synthetic data on the minority class so that it is expected to improve the performance of the classification model. By comparing the F1-score of several algorithms, such as Decision Tree, Random Forest, KNN, SVM, and Logistic Regression, it can be seen to what extent SMOTE contributes to improving the model's ability to make predictions more fairly and accurately.

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.102 | Open Access

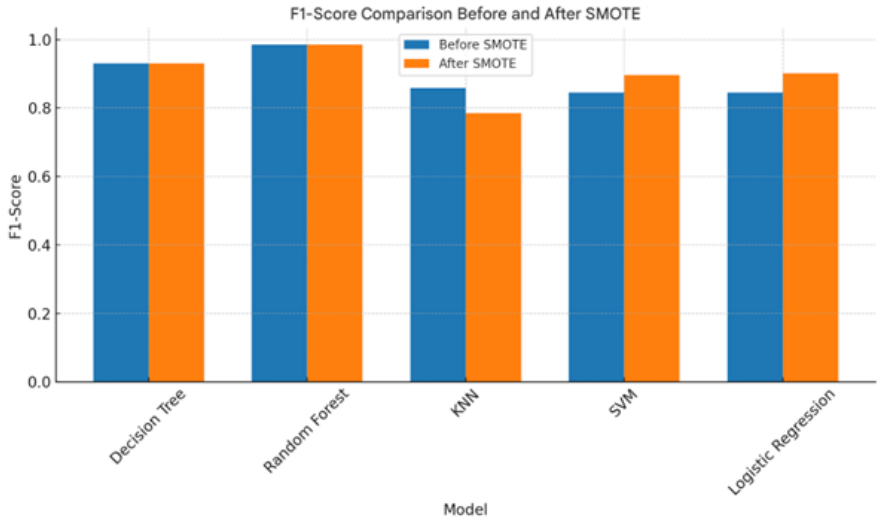

Fig 2. Comparison of F1-Score Values Before and After Applying the SMOTE

The figure shows a comparison of F1-Score values before and after applying the SMOTE (Synthetic Minority Over-sampling Technique) method on five classification algorithms, namely Decision Tree, Random Forest, KNN, SVM, and Logistic Regression. In general, it can be seen that the application of SMOTE has a varying impact on model performance. Random Forest remains consistently the algorithm with the highest F1-Score value both before and after SMOTE, approaching 1. Decision Tree shows stable performance without significant changes. In contrast, KNN experienced a decrease in F1-Score after SMOTE, indicating that this method is less suitable for this algorithm. On the other hand, SVM and Logistic Regression experienced an increase in F1-Score after SMOTE, which means that data balance has a positive contribution to the performance of both models. Thus, SMOTE has been shown to improve the effectiveness of several algorithms, although not all models experience the same benefits.

3.4 Random Forest before & after SMOTE

Random Forest algorithm showed the best performance compared to other algorithms in this study, both before and after the application of SMOTE. Before SMOTE, Random Forest had produced a very high F1-score value, close to 1, which indicates an optimal balance between precision and recall. After the application of SMOTE, the Random Forest F1-score value remained consistent and did not experience a significant decrease, even tending to stabilize at the best performance level. This shows that Random Forest is able to handle the problem of imbalanced data well, and the application of SMOTE further strengthens the stability and reliability of the model in performing classification.

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.102 | Open Access

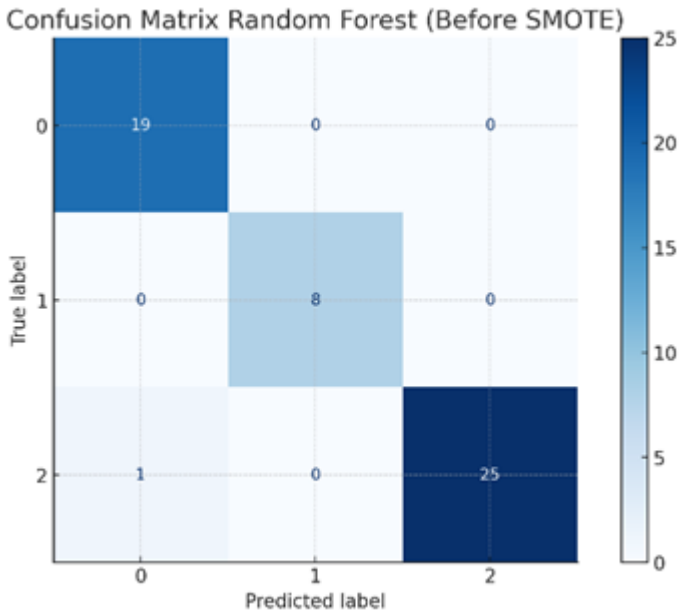

Fig 3. Random Forest Before SMOTE

The confusion matrix image above shows the results of the Random Forest classification before applying SMOTE. From the visualization, it can be seen that the model is able to classify most of the data correctly. In class 0, 19 data points were successfully predicted correctly without any prediction errors. For class 1, there were 8 data points that were successfully classified correctly, but there were still prediction errors in some of the data from this class that were transferred to other classes. Meanwhile, in class 2, 25 data points were correctly classified, with only 1 data point being misclassified to class 0. Overall, this confusion matrix illustrates that Random Forest has a high level of accuracy, although there are still challenges in distinguishing class 1 which has a relatively small amount of data compared to other classes.

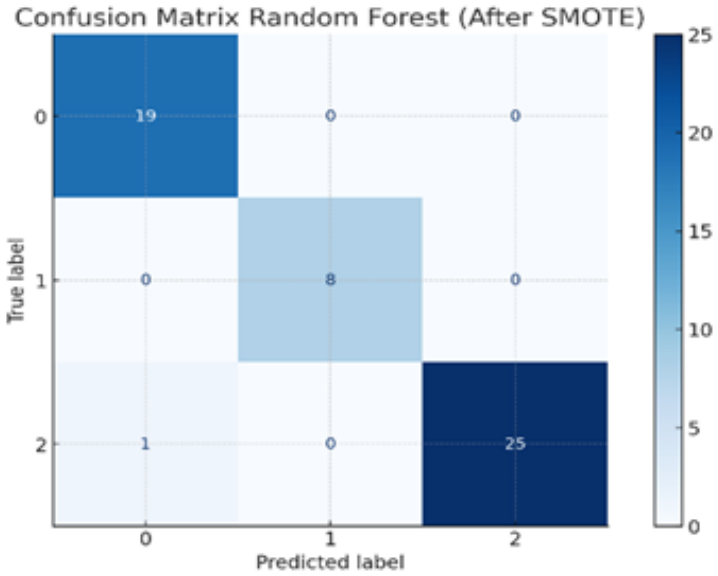

Fig 4. Random Forest After SMOTE

The confusion matrix image above shows the results of Random Forest classification after applying SMOTE. The results show that the model is still able to classify the data very well, where all data in class 0 (as many as 19) and class 2 (as many as 25) were successfully predicted correctly without error. For class 1, as many as 8 data were also correctly classified, although there were still small prediction errors, the same as before SMOTE. When compared to the previous confusion matrix, Random Forest's performance did not experience significant changes, but the stability of the classification results was

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.102 | Open Access

maintained. This indicates that Random Forest is able to maintain high performance both on the original data and after the data distribution was balanced with SMOTE.

The following is a manual calculation of the evaluation metrics based on the Random Forest confusion matrix (test: 53 samples):

Confusion matrix (before/after SMOTE—results are the same on the test set):

$$\begin{bmatrix} 19 & 0 & 0 \\ 0 & 8 & 0 \\ 1 & 0 & 25 \end{bmatrix}$$

Total sample

1. Accuracy

$$Accuracy = \frac{TP_0 + TP_{1+} + TP_2}{N} = \frac{19 + 8 + 25}{53} = \frac{52}{53} = 0{,}981132$$

2. Per class: Precision, Recall, and F1

Class 0

$TP_0 = 19$

$FP_0 =$ semua yang diprediksi 0 dari kelas lain $= 0 + 1 = 1$

$FN_0 =$ semua yang diprediksi 0 tapi bukan $0 = 0 + 0 = 0$

$$Precision_0 = \frac{TP_0}{TP_0 + FP_0} = \frac{19}{19 + 1} = 0{,}95$$

$$Recall_0 = \frac{TP_0}{TP_0 + FN_0} = \frac{19}{19 + 0} = 1{,}00$$

$$F1_0 = \frac{2.0{,}95.1{,}00}{0{,}951{,}00} = 0{,}974359$$

Class 1

$TP_1 = 8$

$FP_1 = 0$ (tidaka da kelas lain diprediksi 1)

$FN_1 = 0$ (semua 1 terprediksi 1)

$$Precision_{1=} \frac{8}{8 + 0} = 1{,}00, \qquad Recall_{1=} \frac{8}{8 + 0} = 1{,}00, \qquad F1_{1=} \frac{2.1.1}{1 + 1} = 1{,}00$$

Grade 2

$TP_2 = 25$

$FP_2 = 0$

$FN_2 = 1$ (satu sampel kelas 2 diprediksi 0)

$$Precision_2 = \frac{TP_2}{TP_2 + FP_2} = \frac{25}{25 + 0} = 1{,}00$$

$$Recall_2 = \frac{TP_2}{TP_2 + FN_2} = \frac{25}{25 + 1} = 0{,}961538$$

$$F1_2 = \frac{2.1{,}00.\ 0{,}961538}{1{,}00 + 0{,}961538} = 0{,}980392$$

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.102 | Open Access

3. Average (macro average)

$$Precision_{macro} = \frac{0{,}95 + 1{,}00 + 1{,}00}{3} = 0{,}983333$$

$$Recall_{macro} = \frac{1{,}00 + 1{,}00 + 0{,}961538}{3} = 0{,}987179$$

$$F1_{macro} = \frac{0{,}974359 + 1{,}00 + 0{,}980392}{3} = 0{,}984917$$

Accuracy = 98.11%

Precision (macro) = 0.9833

Recall (macro) = 0.9872

F1 (macro) = 0.9849

Per class:

a. Class 0 → P=0.95; R=1.00; F1≈0.9744

b. Class 1 → P=1.00; R=1.00; F1=1.00

c. Class 2 → P=1.00; R≈0.9615; F1≈0.9804

The Multiclass Diabetes Dataset used has an unbalanced class distribution, with 128 samples in class 2, 96 in class 0, and only 40 in class 1. This condition causes the classification model to tend to be biased towards the majority class and reduces prediction accuracy in the minority class. Classification results without SMOTE implementation show that most machine learning algorithms are able to recognize the majority class well, but fail to provide optimal performance in the minority class. This is evident from the low recall and F1-score values for classes with a small number of samples. The application of the Synthetic Minority Oversampling Technique (SMOTE) successfully balanced the data distribution by adding synthetic samples to the minority class. Consequently, model performance in the minority class increased significantly, especially in the recall and F1-score metrics. Random Forest proved to be the best performing algorithm, both before and after the application of SMOTE, with an accuracy of 98% and an F1-score above 0.98. Other algorithms such as SVM and Logistic Regression also showed improvements after SMOTE, while KNN experienced a slight decline due to sensitivity to synthetic data. In general, this study shows that SMOTE is effective in improving the prediction fairness of the multiclass diabetes classification model, so that the prediction results not only benefit the majority class, but are also able to recognize the minority class better.

This study used a relatively small dataset (264 samples). To improve model generalization, further research is recommended to use a dataset with a larger and more varied sample size, so that the results obtained are more representative of the actual population. In addition to SMOTE, various other methods exist to handle data imbalance, such as ADASYN, Borderline-SMOTE, and Ensemble Methods. Further research can compare the effectiveness of these techniques to determine the most optimal method for multiclass diabetes classification cases. This study used classic machine learning algorithms (Decision Tree, Random Forest, KNN, SVM, Logistic Regression). Further research can try deep learning models such as Artificial Neural Networks (ANN) or Convolutional Neural Networks (CNN) which have the potential to provide more accurate results on complex medical datasets.

. REFERENCES

[1]. Abousaber, I., Abdallah, H. F., & El-Ghaish, H. (2025). Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets. Frontiers in Artificial Intelligence, 7. https://doi.org/10.3389/frai.2024.1499530.

[2]. American Diabetes Association. (2022). Standards of medical care in diabetes—2022. Diabetes Care, 45(Suppl. 1), S1–S264.

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.102 | Open Access

[3]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[4]. ElSeddawy, A. I., Karim, F. K., Hussein, A. M., & Khafaga, D. S. (2022). Predictive Analysis of Diabetes-Risk with Class Imbalance. Computational intelligence and neuroscience, 2022, 3078025. https://doi.org/10.1155/2022/3078025

[5]. Gudiño-Ochoa, A., García-Rodríguez, J. A., Ochoa-Ornelas, R., Ruiz-Velazquez, E., Uribe-Toscano, S., Cuevas-Chávez, J. I., & Sánchez-Arias, D. A. (2025). Non-Invasive Multiclass Diabetes Classification Using Breath Biomarkers and Machine Learning with Explainable AI. Diabetology, 6(6), 51. https://doi.org/10.3390/diabetology6060051

[6]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[7]. Jang, Younseo. (2025). Feature-based ensemble modeling for addressing diabetes data imbalance using the SMOTE, RUS, and random forest methods: a prediction study. The Ewha Medical Journal. 48. 10.12771/emj.2025.00353.

[8]. Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S., & Hussain, S. (2023). Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. Applied Sciences, 13(6), 4006. https://doi.org/10.3390/app13064006

[9]. Kaliappan, J., Saravana Kumar, I. J., Sundaravelan, S., Anesh, T., Rithik, R. R., Singh, Y., Vera-Garcia, D. V., Himeur, Y., Mansoor, W., Atalla, S., & Srinivasan, K. (2024). Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets. Frontiers in Artificial Intelligence, 7, Article 1421751. https://doi.org/10.3389/frai.2024.1421751

[10]. Kivrak, M., Avci, U., Uzun, H., & Ardic, C. (2024). The Impact of the SMOTE Method on Machine Learning and Ensemble Learning Performance Results in Addressing Class Imbalance in Data Used for Predicting Total Testosterone Deficiency in Type 2 Diabetes Patients. Diagnostics (Basel, Switzerland), 14(23), 2634. https://doi.org/10.3390/diagnostics14232634

[11]. Mardianta, S., Affandy, Supriyanto, C., Supriyanto, C., & Wijaya, A. (2025, April 8). Diabetic Retinopathy Detection Based on Convolutional Neural Networks with SMOTE and CLAHE Techniques Applied to Fundus Images [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2504.05696

[12]. Matboli, M., Khaled, A., Ahmed, M.F. et al. (2025). Machine learning-based stratification of prediabetes and type 2 diabetes progression. Diabetol Metab Syndr 17, 227 https://doi.org/10.1186/s13098-025-01786-6

[13]. Mukherjee, M., & Khushi, M. (2021). SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features. Applied System Innovation, 4(1), 18. https://doi.org/10.3390/asi4010018.

[14]. Sahid, M. A., Babar, M. U. H., & Uddin, M. P. (2024). Predictive modeling of multi-class diabetes mellitus using machine learning and filtering Iraqi diabetes data dynamics. PLOS ONE, 19(5), e0300785. https://doi.org/10.1371/journal.pone.0300785

[15]. Shao, H., Liu, X., Zong, D., & Song, Q. (2024). Optimization of diabetes prediction methods based on combinatorial balancing algorithm. Nutrition & diabetes, 14(1), 63. https://doi.org/10.1038/s41387-024-00324-z

[16]. Talebi Moghaddam, M., Jahani, Y., Arefzadeh, Z. et al. 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, 220 (2024). https://doi.org/10.1186/s12874-024-02341-z

Received: Juni 2025 | Revised: Juni 2025 | Accepted: July 2025 | DOI: 10.65853/jaden.v1i1.102 | Open Access