

Enhanced diabetes prediction using SMOTE-Tomek and F1-optimized machine learning models

Chaitra Nayak Janasale¹, Rajanikanta Mohanty¹, Sarappadi Narasimha Prasad²

¹Department of Computer Science and Engineering, Faculty of Engineering and Technology, JAIN (Deemed-to-be University), Bengaluru, India

²Department of Electronics and Communications Engineering, School of Engineering, Dayananda Sagar University, Bengaluru, India

Article Info

Article history:

Received Apr 20, 2025

Revised Jul 21, 2026

Accepted Aug 12, 2026

Keywords:

Class imbalance

Diabetes prediction

Early diagnosis

Healthcare analytics

SMOTE-Tomek resampling

ABSTRACT

This study presents a systematic and effective strategy for the early prediction of diabetes, utilizing a hybrid machine learning framework designed to address prevalent issues found in conventional diagnostic systems. A significant challenge in medical datasets, particularly in the context of diabetes, is the issue of class imbalance, characterized by a disproportionate representation of non-diabetic cases compared to diabetic cases. The proposed framework addresses this issue by implementing the synthetic minority oversampling technique (SMOTE)-Tomek resampling in conjunction with a model selection strategy that emphasizes the F1-score for the diabetic class. This approach ensures a heightened focus on accurately identifying at risk individuals, rather than solely optimizing for overall accuracy. A variety of supervised learning models were assessed, comprising logistic regression (LR), decision trees (DT), support vector machines (SVM), random forest (RF), k-nearest neighbors (KNN), extreme gradient boosting (XGBoost), and artificial neural networks (ANN). A 10-fold cross-validation technique was utilized to ensure a robust evaluation of model performance. The evaluated models, ANN and XGBoost, demonstrate notable performance, with F1-scores of 0.87 and 0.84, respectively. The ANN model achieves the highest accuracy, recorded at 94.3%. SMOTE-Tomek has enhanced recall and decreased false negatives, thereby increasing the model's sensitivity to diabetic cases.

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



Corresponding Author:

Sarappadi Narasimha Prasad

Department of Electronics and Communication Engineering, School of Engineering

Dayananda Sagar University

Bengaluru 56011, India

Email: prasad.sn@dsu.edu.in

1. INTRODUCTION

Hyperglycaemia, or consistently high blood sugar levels, is a defining characteristic of diabetes mellitus, a complex, long-term metabolic disorder [1]. In recent years, diabetes has become one of the most significant global health challenges. According to the World Health Organization, the number of people suffering from diabetes continues to increase, with over 537 million individuals affected worldwide as of 2024 [2]. This growing prevalence not only threatens individual health but also puts immense pressure on healthcare systems around the globe. When left undiagnosed, diabetes can lead to a range of serious health complications. These include loss of vision [3], nerve damage [4], and kidney failure [5]. It can also increase the risk of developing cardiovascular diseases [6] and, in more severe cases, may result in limb amputations [7]. The widespread impact of the disease highlights the importance of early detection, effective

monitoring, and consistent treatment to reduce the risk of long-term health issues and to ease the burden on medical infrastructure.

To reduce the risk of diabetes-related complications, early diagnosis and timely intervention are essential [8]. However, commonly used diagnostic methods, such as the oral glucose tolerance test [9], fasting blood glucose test [10], and HbA1c levels [11], come with several drawbacks. While these tests are clinically valuable, they are expensive, especially when frequent monitoring is required. Moreover, their accuracy can be affected by a variety of factors, including a person's diet, stress levels, and individual health conditions, leading to inconsistent results. These limitations highlight the growing need for alternative diagnostic solutions that are not only accurate but also more cost-effective, efficient, and widely scalable. With the global rise in diabetes cases, accessible screening tools are critical for identifying high-risk individuals earlier and enabling quicker clinical responses.

Recent advancements in artificial intelligence (AI) [12], and more specifically in machine learning (ML) [13], are spreading light on exciting new possibilities for transforming the way medical diagnostics are approached. ML algorithms are particularly well-suited for healthcare applications because they can process and analyze large, complex datasets, detect subtle trends, and generate accurate predictive insights that might be missed by traditional methods. In the context of diabetes, ML offers a powerful tool for early prediction and risk assessment. By drawing on information from electronic health records, clinical metrics, and lifestyle factors, ML models can effectively identify individuals who may be at high risk before symptoms become severe. This not only facilitates earlier diagnoses but also supports the creation of personalized treatment plans, tailored to each patient's unique health profile. As a result, ML has the potential to significantly improve the speed and accuracy of diabetes detection, while also reducing the burden on healthcare systems.

This study focuses on building and evaluating effective predictive models for diabetes using a range of well-established supervised ML techniques. The models explored include logistic regression (LR), decision trees (DT), support vector machines (SVM), random forests (RF), k-nearest neighbors (KNN), extreme gradient boosting (XGBoost), and artificial neural networks (ANN). The goal is to assess how well these models perform in predicting diabetes by using a detailed dataset that includes a wide range of features appropriate to the condition. To ensure the models are both accurate and efficient, several important preprocessing steps were carried out. These included normalizing the data, encoding categorical variables, and using methods to address class imbalance, which is a common issue in medical datasets. By preparing the data properly, the models are better equipped to produce reliable results. Ultimately, this research aims to demonstrate how ML-driven intelligent systems can play a transformative role in the early detection of diabetes. Beyond improving diagnostic accuracy, such systems have the potential to lower healthcare costs, ease the global burden of the disease, and most importantly, enhance patient outcomes and quality of life by enabling timely and targeted interventions [13].

2. LITERATURE REVIEW

Longato *et al.* [14] adopted deep learning (DL) to predict major adverse cardiovascular events (MACE) in diabetics using Veneto administrative healthcare data. Trained their model using one year of pharmacy and hospital data from 214,676 diabetic patients. The predictive framework predicted death, heart failure, heart attacks, and strokes over one to five years. Due to neural network strengths, their approach required little data preprocessing, which increased model scalability. The study was hampered by claims data's complexity, the absence of clinical characteristics including patient weight and ethnicity, and the difficulty of developing good predictions from administrative datasets. Sharma *et al.* [15] developed a cloud-integrated, automated diabetes monitoring system to improve early identification and healthcare access in remote locations. They selected features using an extreme learning machine (ELM) classifier and principal component analysis (PCA). The solution was evaluated in standalone and cloud-based settings. The cloud implementation outperformed the solo system in accuracy and processing speed. However, the standalone version had lower processing power and slower classification times, demonstrating the benefits of cloud infrastructure in expanding healthcare monitoring technology.

Febrian *et al.* [16] carried out a comparative study to determine the most effective ML algorithm for the early diagnosis of diabetes, aiming to reduce health complications through timely intervention. Using the widely recognized Pima Indian diabetes dataset, the researchers evaluated models such as KNN, RF, SVM, and DT, focusing on both their accuracy and computational efficiency. While the findings showed that each algorithm had distinct advantages, the study also highlighted a key challenge, inconsistency in model performance across different scenarios, making it difficult to identify a single best model for all cases. In response to growing need for more personalized and accurate diabetes monitoring, Obeidat and Ammar [17] proposed an intelligent, patient-centric system that not only monitored blood glucose levels but also predicted insulin requirements using ANN. The system was tested on a small sample of 13 individuals and showed

promising results in real-time monitoring and prediction. However, the study noted limitations, such as the necessity for continuous data input and the challenge of managing individual differences in patients, which may limit how widely the system can be applied. Expanding the role of intelligent healthcare solutions, Marzouk *et al.* [18] introduced a secure, web-based diabetes management system that integrated ML algorithms with internet of things (IoT) technologies. Their approach used QR code-enabled cards to facilitate real-time data sharing between patients and healthcare providers. A variety of ML models, including DT, SVM, and ANN, were used to build a predictive engine for type-2 diabetes diagnosis and long-term patient monitoring.

Isaza and Diaz [19] explored the use of DL for detecting diabetes mellitus through foot thermography images. To overcome challenges related to limited dataset size and lack of diversity, they applied various data augmentation techniques on a publicly available dataset. The team utilized a ResNet50v2 convolutional neural network (CNN) for image classification and also experimented with transfer learning to boost the model's ability to generalize by incorporating knowledge from external datasets. Issues such as a small sample size, variability in thermographic images, and inconsistent data quality affected the model's robustness and overall accuracy, limiting its practical application. In a different approach, Dutta *et al.* [20] introduced a disease prediction model (DPM) aimed at early detection of type 2 diabetes (T2D) and hypertension. Their method used ensemble learning and addressed two common challenges in healthcare data: outliers and class imbalance. To manage these, the model incorporated isolation forest (iForest) for detecting outliers and synthetic minority oversampling technique (SMOTE)-Tomek for balancing the data. Tested across four different datasets, the model showed improved predictive performance. To support real-world use, the researchers also developed a mobile application for convenient, real-time deployment of the model, enhancing accessibility for both patients and healthcare providers. However, the study noted ongoing challenges in handling extreme outliers and guaranteeing dependable performance across imbalanced datasets, which could compromise model reliability.

Butt *et al.* [21] developed a diagnostic system employing ML algorithms for early detection and categorization of diabetes. Using data from several medical facilities, the study assessed LR, naïve Bayes, SVM, and RF performance. The models were evaluated using accuracy, precision, recall, F1-score, and Kappa index. The study found that naïve Bayes had the lowest accuracy (69%), while RF had the highest accuracy (99%). These findings stressed the relevance of dataset-based model selection and consistent performance. Rashtian *et al.* [22] investigated whether heart rate (HR) monitors could detect glucose abnormalities in diabetics at a lower cost than CGMs. They studied 550 people who wore HR monitors and CGMs for 10 days. Researchers found meaningful associations between HR patterns and glucose changes using canonical correlation analysis (CCA). While optimistic, the study relied on self-reported health information, could have found undiagnosed diabetes in healthy subjects, and needed more testing across more diverse and larger populations. Along with enhanced monitoring tools, Shokrehodaei *et al.* [23] developed a decision support system (DSS) that integrated general practitioners electronic health record (EHR) into a cloud-based platform to improve T2D management. The system focused on analyzing EHR data to evaluate the quality of care provided to patients with chronic illnesses using ML techniques. Designed as a DSS, it aimed to enhance patient outcomes by offering real-time insights and performance feedback to general practitioners. Additionally, the system incorporated data-driven assessments that could be tied to financial incentives, encouraging improved care delivery.

The studies reviewed underscore the growing role of ML and DL in improving both the early detection of diabetes and the development of more personalized approaches to disease management. These works consistently show that thoughtful preprocessing steps, such as data augmentation, managing outliers, and correcting class imbalances, play a crucial part in boosting model accuracy and making these systems more scalable for real-world use. Another important insight is the benefit of combining multiple data sources. By integrating physiological signals, such as glucose levels or thermographic images, with structured data from EHRs, predictive models become more robust and versatile, with greater potential to assist in clinical decision-making. Despite these promising developments, there are still significant challenges to address. Ensuring that models generalize well across different patient groups, preserving data quality, and building secure, real-time systems that can adapt to the needs of diverse populations remain key areas for future research and improvement. These challenges highlight the need for continuous refinement as intelligent healthcare systems move closer to everyday clinical use.

3. PROPOSED METHOD

Figure 1 presents a well-structured and modular pipeline designed for diabetes prediction, built on a multi-model machine learning architecture. What makes this approach stand out is its thoughtful integration of advanced preprocessing methods, ensemble-based performance evaluation, and a dynamic model selection strategy that prioritizes the F1-score of the diabetic class—ensuring the system is finely tuned to identify

Enhanced diabetes prediction using SMOTE-Tomek and F1-optimized ... (Chaitra Nayak Janasale)

at-risk individuals. The process begins with data acquisition, where a large and diverse dataset comprising 100,000 records is collected. Each entry includes nine key attributes: age, gender, body mass index (BMI), hypertension status, heart disease, smoking history, HbA1c levels, and blood glucose levels. Unlike many traditional models that depend heavily on raw statistical values, this study takes a more holistic approach by emphasizing demographic diversity and ensuring a balanced class distribution within the data. This results in a dataset that more accurately reflects real-world populations, thereby increasing the relevance and reliability of the model's predictions in practical healthcare settings. In the data preprocessing stage, standard normalization techniques are enhanced by applying SMOTE-Tomek resampling [24], a method specifically used to address class imbalance, algorithm steps are summarized in Algorithm 1. This technique not only balances the dataset by combining oversampling and undersampling but also helps eliminate noisy or borderline examples, resulting in a cleaner and more representative dataset. To manage categorical data efficiently, label encoding is used where appropriate, minimizing unnecessary feature expansion and keeping the model lightweight and efficient. For feature selection and extraction, a hybrid approach is adopted, combining correlation analysis with recursive feature elimination (RFE) [25], supported by cross-validation to improve reliability. This process ensures that only the most relevant features are retained for training, while also taking into account relationships between features and helping to prevent overfitting. These steps contribute to a more robust and streamlined model, capable of making accurate predictions while remaining generalizable across different data scenarios.

Algorithm 1. SMOTE-Tomek resampling

Step 1: Class Imbalance Detection Determine the number of samples in each class. Identify the minority class m such that $|m| < |M|$.

Step 2: SMOTE - Synthetic Minority Oversampling

- For each sample $x \in m$, identify its k nearest neighbors from the same class.
- Randomly select r neighbors from the k -nearest set.
- For each selected neighbor x_{nn} , generate a synthetic sample x_{new} using

$$x_{new} = x + \lambda \cdot (x_{nn} - x), \lambda \in [0, 1] \quad (1)$$

Step 3: Tomek links - Majority Class Undersampling

- For each pair (x_i, x_j) where $x_i \in M$ and $x_j \in m$, compute the nearest neighbor relationship.
- A Tomek link exists if x_i is the nearest neighbor of x_j and vice versa, and they belong to opposite classes.
- Remove such Tomek pairs—commonly discarding the majority class instance x_i —to reduce borderline noise.

The integration of SMOTE for oversampling and Tomek links for undersampling plays a crucial role in the preprocessing pipeline. This combined approach addresses class imbalance issues in medical datasets. Its primary function is twofold: first, it helps balance the class distribution by generating synthetic examples for the minority class, which in this case could be diabetic patients. By doing so, it reduces the inherent bias that machine learning models often develop toward the majority class, ensuring that underrepresented cases are given equal importance during training. This leads to a fairer and more accurate model that can better identify high-risk individuals. Secondly, the identification and removal of Tomek links—pairs of data points from distinct classes that are in close proximity—serves to eliminate ambiguous and overlapping samples that may adversely affect the model's accuracy. The application of SMOTE-Tomek technique in this study yields machine learning models that exhibit improved calibration and enhanced ability to accurately differentiate between diabetic and non-diabetic individuals. This results in enhanced sensitivity to minority class instances, thereby increasing the reliability and clinical significance of the system for early detection of diabetes. The model development phase incorporates an innovative aspect: the concurrent training of various machine learning models, such as LR, DT, SVM, RF, KNN, XGBoost, and ANN. Each model is subjected to 10-fold cross-validation to evaluate its generalizability. Instead of choosing the final model based on overall accuracy, a model selection process driven by class-specific F1-scores [26] is utilized, method steps are given in Algorithm 2. This mechanism selects the optimal classifier for identifying minority (diabetic) instances by emphasizing the balance between recall and precision, a critical factor in medical applications.

Algorithm 2. Class-specific F1-score based model selection

For each model $M_i \in M$:

- a. Apply k -fold cross-validation on dataset D
- b. For each fold:
 - i. Train model on $k-1$ folds
 - ii. Evaluate on the remaining fold
 - iii. Compute F1-score for class 1 (F1_class1)
- c. Calculate the mean F1_class1 over all folds

Select the model M^* such that: $M^* = \arg \max_{M_i \in M} (Mean F1_{class1(M_i)})$
 Return the selected model M^* for final training and prediction.

The model selection approach used in this study is carefully designed to prioritize the accurate identification of diabetic cases, which often represent a smaller portion of medical datasets. Rather than relying solely on overall accuracy, which can be misleading when working with imbalanced data, this strategy focuses on maximizing the F1-score specifically for the diabetic class. By doing so, it ensures a more balanced evaluation of both precision and recall, two metrics that are especially important in healthcare predictions, where missing a positive case can have serious consequences. This targeted focus helps the model become more sensitive to correctly identifying individuals with diabetes, resulting in fewer false negatives (FN). As a result, the system becomes not only more clinically relevant, but also more trustworthy for real-world medical use. By shifting the emphasis from general accuracy to class-specific performance, the model becomes better suited to meet the demands of accurate, early diagnosis in healthcare environments.

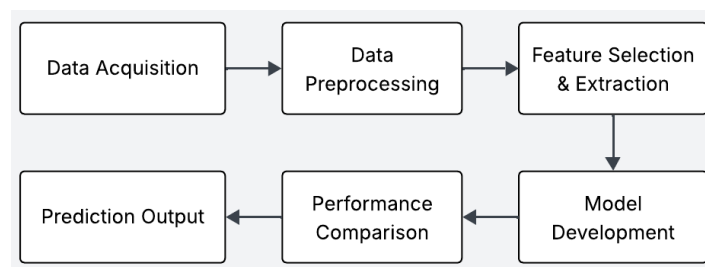


Figure 1. Block diagram of proposed method

After all models are evaluated, their performance results are brought together using weighted averaging, a method that gives more importance to models that not only perform well on balanced datasets but also show strength in accurately identifying at-risk individuals. This step is crucial in reducing bias toward the majority class, typically non-diabetic cases, and ensures that minority class predictions, like those for diabetes, are not overlooked. In the final phase, the selected model generates a probability-based risk score, which can be seamlessly integrated into clinical DSS. This allows healthcare providers to make more informed, data-driven decisions. Due to the modular structure of the pipeline, the system is designed with future expansion in consider. It can be easily adapted to work with real-time IoT health monitoring devices or integrated into cloud-based medical record systems, enabling continuous prediction, monitoring, and feedback. This flexibility makes it well-suited for practical use in modern, technology-driven healthcare environments.

4. PERFORMANCE EVALUATION

The results presented in this section highlight the effectiveness and innovation of the proposed approach for diabetes prediction. By combining advanced ML models with a class-specific F1-score-driven selection strategy and SMOTE-Tomek resampling, the system significantly improves the detection of diabetic cases, an area that often suffers due to class imbalance in medical datasets. The use of SMOTE-Tomek plays a vital role in this improvement. It not only generates synthetic samples to better represent the minority class but also removes overlapping data points that could blur the distinction between classes. This targeted resampling leads to noticeable gains in both recall and accuracy for diabetic predictions, ensuring that more true cases are correctly identified. SMOTE works by generating synthetic examples for the minority class, in this case, individuals with diabetes, helping to ensure these cases are better represented during model training. Meanwhile, Tomek links helps to clean the dataset by removing overlapping and noisy samples from the majority class, which can otherwise confuse the model and reduce prediction accuracy. The main advantage of using SMOTE-Tomek is that it improves the model's ability to identify diabetic cases more accurately, without favoring the majority class. This results in better class sensitivity and overall performance. As shown in Table 1, comparing results before and after applying the technique clearly demonstrates the model's performance improvement, particularly in the detection of diabetic cases (class 1). This makes SMOTE-Tomek a valuable step in preparing data for reliable and fair machine learning predictions in healthcare.

Once SMOTE-Tomek was applied, the dataset became more balanced by adding synthetic diabetic samples through SMOTE and removing Tomek links, which represent noisy. This adjustment had a clear and

positive impact on model performance, particularly in how well the models identified diabetic cases. The most significant improvement was seen in the recall for the diabetic class, with several models showing increased F1-scores - a key metric when evaluating performance on imbalanced datasets. Models like LR and DT, which often face challenges in recognizing minority class cases, showed marked gains. The introduction of synthetic examples gave the models more diabetic cases to learn from, while the removal of confusing majority class data helped reduce misclassifications.

Table 1. Model comparison without and with SMOTE-Tomek resampling

Model	Without SMOTE-Tomek (accuracy)	With SMOTE-Tomek (accuracy)
LR	89.5	92.0
DT	88.3	91.2
KNN	87.2	90.8
RF	89.7	93.5
XGBoost	90.0	94.1
ANN	91.2	94.3

To improve the accuracy of diabetes prediction, especially for the diabetic class (class 1), this study adopted a class-specific F1-score-based model selection strategy. This method is particularly valuable because it focuses on minimizing FNs, where diabetic individuals might otherwise be misclassified as non-diabetic. In a clinical context, such errors can lead to missed diagnoses and delayed treatment, making early and accurate identification critical. Rather than relying solely on overall accuracy, which can be misleading in imbalanced datasets, this approach prioritizes the F1-score for class 1, striking a balance between precision and recall. This ensures that the chosen model is not only identifying diabetic cases correctly but also doing so consistently, without being overly influenced by the majority class. As shown in Table 2, ANN and XGBoost achieved the highest F1-scores for the diabetic class, demonstrating superior performance in recognizing individuals at risk. This targeted evaluation method proves especially effective in handling the challenges posed by imbalanced medical data, leading to more reliable and clinically useful prediction outcomes.

Table 2. Model comparison based on class-specific F1-score for diabetic class

Model	Accuracy (%)	Precision (class 1)	Recall (class 1)	F1-score (class 1)
LR	92.0	0.87	0.61	0.72
DT	91.2	0.84	0.70	0.76
KNN	90.8	0.83	0.68	0.74
RF	93.5	0.88	0.76	0.81
XGBoost	94.1	0.90	0.79	0.84
ANN	94.3	0.91	0.83	0.87

ANN clearly had the best F1-score for the diabetic class from Table 2 at 0.87 and an accuracy of 94.3%. With an F1-score of 0.84 and a 94.1% accuracy, XGBoost came next. These findings show how well complex, non-linear models like ANN and XGBoost handle imbalanced datasets, enabling them to have more sensitivity for forecasting minority class occurrences. In comparison, LR achieved an acceptable overall accuracy of 92%, but its performance in identifying diabetic cases was less impressive. With a recall of just 0.61, the model struggled to detect a significant portion of diabetic individuals, leading to a reduced F1-score of 0.72. This suggests that although the model was effective at correctly classifying non-diabetic cases, it often failed to recognize those with diabetes. On the other hand, models like DT and KNN performed slightly better than LR, although they still lagged behind more advanced models such as ensemble techniques and DL algorithms. The use of a class-specific F1-score as a model selection criterion proves its value here by ensuring the focus remains on accurately identifying diabetic patients.

In order to prove the strength of these findings, the findings were compared with the available literature. The sensitivity enhancement noted with use of SMOTE-Tomek is consistent with Dutta *et al.* [20], who also showed that the prediction of diseases is not reliable without addressing the issue of class imbalance. They, however, studied ensemble learning in their research, whereas we build upon it by incorporating a class-specific F1-score selection strategy. This will overcome the weakness that Febrian *et al.* [16] pointed out, which is that different ML algorithms tend to exhibit variable performance in different contexts. The ANN model 94.3% accuracy, 0.87 F1-score can be useful in reducing the high false-negative values typical of most accuracy-driven methods, including the ones reported by Butt *et al.* [21], by focusing the F1-score on

the diabetic class instead of the overall accuracy. This comparison proves that the SMOTE-Tomek would be more clinically safe and effective in diagnosis when used in combination with F1-optimization as compared to the traditional accuracy-focused procedures.

4.1. Prediction error analysis

Especially in sectors like healthcare, where early diagnosis can greatly affect patient outcomes, prediction error analysis is a vital component of evaluating the efficacy of ML models. Using confusion matrix-based measures, the ML models employed for diabetes prediction in this work were thoroughly assessed with an emphasis on error analysis. The primary error kinds examined were false positives (FP), in which non-diabetic patients are wrongly categorized as diabetic, and FN, in which diabetic patients are wrongly labeled as non-diabetic. In healthcare, these mistakes are especially significant: FPs could cause needless therapies and anxiety for non-diabetic people and FNs could cause delayed medical action for those who really require it. Examining the confusion matrix for the best-performing model, ANN, after applying the SMOTE-Tomek resampling technique assisted to better comprehend and visualize these issues. Trained on a balanced dataset, where synthetic examples were generated for the minority class and overlapping or unclear cases were eliminated from the majority class, examining the confusion matrix showed the model misclassified roughly 12% of non-diabetic people as diabetes (FPs). Though this percentage is somewhat low, it emphasizes the need to prevent needless diagnosis for healthy people. More importantly, over 22% of diabetic cases were wrongly categorized as non-diabetic (FNs). Undiagnosed diabetes patients could have delays in getting required therapy, which could be quite worrying since they could develop major health problems.

AUC value of 0.91 provides more understanding of the model's capacity to differentiate between non-diabetic and diabetic individuals. Still, there is space for development, especially in terms of lowering FNs. The fairly high proportion of false negatives draws attention to the need to improve the model to more accurately identify early-stage diabetes and minimize the risk of missed diagnosis. Further research on prediction results across several demographic categories and risk factors, including age, BMI, HbA1c levels, and hypertension, helped to clarify these mistakes. This demographic analysis helps identify whether some groups were more likely to be misclassified. Age-based studies indicated that older patients, especially those between 60 and 80, had a greater prevalence of FNs. This tendency probably reflects the changing character of diabetes in older people, where symptoms could be more mild or unusual. For example, older people may have other health issues, which could complicate the model's ability to precisely identify early indicators of diabetes.

5. CONCLUSION

This study introduces a comprehensive and innovative ML framework aimed at enhancing the early prediction of diabetes. This approach integrates comprehensive data preprocessing techniques with a focused evaluation strategy to tackle significant challenges encountered in medical prediction tasks. The framework's core strength lies in its implementation of SMOTE-Tomek resampling, which addresses class imbalance effectively. This is particularly relevant in healthcare datasets, where diabetic cases are frequently underrepresented. This technique enhances the representation of the minority class while simultaneously minimizing noise and overlap, resulting in cleaner and more reliable training data. The framework incorporates a class-specific F1-score as a criterion for model selection, enhancing its effectiveness. This approach prioritizes the accurate identification of diabetic patients over overall accuracy, which may not provide a reliable measure in imbalanced datasets. The enhanced emphasis on precision and recall for the diabetic class contributes to the development of more sensitive and clinically relevant prediction tools, thereby improving the likelihood of early detection and timely intervention. The proposed approach underwent a comprehensive evaluation utilizing a real-world dataset comprising 100,000 patient records, yielding promising results, especially with the ANN and XGBoost models. Among the two models evaluated, the ANN demonstrated superior performance, attaining an overall accuracy of 94.3% and an F1-score of 0.87 specifically for the diabetic class. The results demonstrate the system's capability in precisely identifying individuals who are at an elevated risk for diabetes. An extensive error analysis indicated that older patients experienced a higher rate of misclassification, suggesting the potential advantage of integrating age-specific features into forthcoming predictive models. This analysis highlights the necessity of enhancing prediction methodologies to more effectively account for variations among distinct age demographics. The framework not only enhances classification metrics but also effectively addresses critical clinical issues, specifically by minimizing FNs. This facilitates earlier diagnosis and enables timely medical intervention. This research introduces a scalable and intelligent DSS that enhances the accuracy of diabetes detection while establishing a foundation for more personalized and data-driven healthcare solutions.

FUNDING INFORMATION

Authors state no funding involved.

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
Chaitra Nayak Janasale	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			✓
Rajanikanta Mohanty		✓		✓		✓		✓		✓		✓		
Sarappadi Narasimha Prasad				✓	✓		✓			✓		✓	✓	

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] M. Cazzola, P. Rogliani, J. Ora, L. Calzetta, D. Lauro, and M. G. Matera, "Hyperglycaemia and chronic obstructive pulmonary disease," *Diagnostics*, vol. 13, no. 21, Nov. 2023, doi: 10.3390/diagnostics13213362.

[2] S. Goyal and V. Vanita, "The rise of type 2 diabetes in children and adolescents: an emerging pandemic," *Diabetes/Metabolism Research and Reviews*, vol. 41, no. 1, Jan. 2025, doi: 10.1002/dmrr.70029.

[3] S. B. Bressler, P. H. Scanlon, and E. Pearce, "Why is continued vision loss still a problem in some patients with diabetic retinopathy, despite treatment?," *JAMA Ophthalmology*, vol. 140, no. 4, pp. 308–309, Mar. 2022, doi: 10.1001/jamaophthalmol.2022.0135.

[4] S. A. Eid *et al.*, "New perspectives in diabetic neuropathy," *Neuron*, vol. 111, no. 17, pp. 2623–2641, May 2023, doi: 10.1016/j.neuron.2023.05.003.

[5] E. Rafiee, "Photonic crystal-based biosensor for diagnosis of kidney failure and diabetes," *Plasmonics*, vol. 19, no. 1, pp. 439–445, Aug. 2024, doi: 10.1007/s11468-023-02014-5.

[6] L. M. G.-Peña, E. D. Abel, and R. O. Pereira, "Mitochondrial dynamics, diabetes, and cardiovascular disease," *Diabetes*, vol. 73, no. 2, pp. 151–161, Jan. 2024, doi: 10.2337/dbi23-0003.

[7] B. T. Rodrigues, V. N. Vangaveti, R. Urkude, E. Biro, and U. H. Malabu, "Prevalence and risk factors of lower limb amputations in patients with diabetic foot ulcers: a systematic review and meta-analysis," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 16, no. 2, Jan. 2022, doi: 10.1016/j.dsx.2022.102397.

[8] S. Thipsawat, "Early detection of diabetic nephropathy in patients with type 2 diabetes mellitus: a review of the literature," *Diabetes and Vascular Disease Research*, vol. 18, no. 6, Nov. 2021, doi: 10.1177/14791641211058856.

[9] M. Lages, R. Barros, P. Moreira, and M. P. Guarino, "Metabolic effects of an oral glucose tolerance test compared to the mixed meal tolerance tests: a narrative review," *Nutrients*, vol. 14, no. 10, May 2022, doi: 10.3390/nu14102032.

[10] P. P.-Ros, E. N.-Flores, I. J.-Rochina, F. M. M.-Arnau, and O. Cauli, "Changes in salivary amylase and glucose in diabetes: a scoping review," *Diagnostics*, vol. 11, no. 3, Mar. 2021, doi: 10.3390/diagnostics11030453.

[11] J. Chen, D. Yin, and K. Dou, "Intensified glycemic control by HbA1c for patients with coronary heart disease and type 2 diabetes: a review of findings and conclusions," *Cardiovascular Diabetology*, vol. 22, no. 1, Jun. 2023, doi: 10.1186/s12933-023-01875-8.

[12] A. Ahmed, S. Aziz, A. A.-Alrazaq, F. Farooq, and J. Sheikh, "Overview of artificial intelligence-driven wearable devices for diabetes: scoping review," *Journal of Medical Internet Research*, vol. 24, no. 8, Aug. 2022, doi: 10.2196/36010.

[13] M. A. Makroum, M. Adda, A. Bouzouane, and H. Ibrahim, "Machine learning and smart devices for diabetes management: systematic review," *Sensors*, vol. 22, no. 5, Feb. 2022, doi: 10.3390/s22051843.

[14] E. Longato, G. P. Fadini, G. Sparacino, A. Avogaro, L. Tramontan, and B. D. Camillo, "A deep learning approach to predict diabetes' cardiovascular complications from administrative claims," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 9, pp. 3608–3617, Mar. 2021, doi: 10.1109/JBHI.2021.3065756.

[15] S. K. Sharma *et al.*, "A diabetes monitoring system and health-medical service composition model in cloud environment," *IEEE Access*, vol. 11, pp. 32804–32819, Jan. 2023, doi: 10.1109/ACCESS.2023.3258549.

[16] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," *Procedia Computer Science*, vol. 216, pp. 21–30, Jan. 2022, doi: 10.1016/j.procs.2022.12.107.

[17] Y. Obeidat and A. Ammar, "A system for blood glucose monitoring and smart insulin prediction," *IEEE Sensors Journal*, vol. 21, no. 12, pp. 13895–13909, Apr. 2021, doi: 10.1109/JSEN.2021.3070706.

- [18] R. Marzouk, A. S. Alluhaidan, and S. A. El Rahman, "An analytical predictive models and secure web-based personalized diabetes monitoring system," *IEEE Access*, vol. 10, pp. 105657–105673, Jan. 2022, doi: 10.1109/ACCESS.2022.3211264.
- [19] A. A.-Isaza and M. Z.-Diaz, "Detection of diabetes mellitus with deep learning and data augmentation techniques on foot thermography," *IEEE Access*, vol. 10, pp. 59564–59591, Jan. 2022, doi: 10.1109/ACCESS.2022.3180036.
- [20] A. Dutta *et al.*, "Early prediction of diabetes using an ensemble of machine learning models," *International Journal of Environmental Research and Public Health*, vol. 19, no. 19, Sep. 2022, doi: 10.3390/ijerph191912378.
- [21] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine learning based diabetes classification and prediction for healthcare applications," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–17, Sep. 2021, doi: 10.1155/2021/9930985.
- [22] H. Rashtian, S. S. Torbaghan, S. Rahili, M. Snyder, and N. Aghaeepour, "Heart rate and CGM feature representation diabetes detection from heart rate: learning joint features of heart rate and continuous glucose monitors yields better representations," *IEEE Access*, vol. 9, pp. 83234–83240, Jan. 2021, doi: 10.1109/ACCESS.2021.3085544.
- [23] M. Shokrehodaie, D. P. Cistola, R. C. Roberts, and S. Quinones, "Non-invasive glucose monitoring using optical sensor and machine learning techniques for diabetes applications," *IEEE Access*, vol. 9, pp. 73029–73045, Jan. 2021, doi: 10.1109/ACCESS.2021.3079182.
- [24] D. S. Assyifa and A. Luthfiarta, "SMOTE-Tomek re-sampling based on random forest method to overcome unbalanced data for multi-class classification," *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 9, no. 2, pp. 151–160, Jul. 2024, doi: 10.25139/inform.v9i2.8410.
- [25] M. Awad and S. Fraihat, "Recursive feature elimination with cross-validation with decision tree: feature selection method for machine learning-based intrusion detection systems," *Journal of Sensor and Actuator Networks*, vol. 12, no. 5, Sep. 2023, doi: 10.3390/jsan12050067.
- [26] N. P. Walter, J. Fischer, and J. Vreeken, "Finding interpretable class-specific patterns through efficient neural search," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, pp. 9062–9070, Mar. 2024, doi: 10.1609/aaai.v38i8.28756.

BIOGRAPHIES OF AUTHORS



Chaitra Nayak Janasale    obtained her master's degree in Data Engineering and Cloud Computing from REVA University, Bengaluru, in 2014. She is currently pursuing a Ph.D. in Computer Science and Engineering at Jain University, Bengaluru, India, with a research focus on early-stage prediction of diabetes. Her areas of expertise include artificial intelligence and machine learning. She is an assistant professor with over seven years of teaching experience, has organized more than ten technical talks, and has actively served on organizing committees for several national and international conferences. Additionally, she has frequently served as the Master of Ceremony for numerous academic events. She has filed an Indian patent titled "Conversion of handwritten text to digitized form using machine learning" (application no. 202141019038). She can be contacted at email: chaitranayak1012@gmail.com.



Dr. Rajanikanta Mohanty    is a seasoned academician with over 33 years of teaching and mentoring experience in higher education. He earned his M.Tech. from Utkal University and completed his Ph.D. at SoA University. Throughout his career, he has been actively involved in mentoring students across B.Tech., MCA, and M.Tech. programs, reflecting his strong commitment to academic excellence and student development. He has held key administrative roles such as head of department, vice principal, and principal at reputed institutions. He has also contributed as a visiting faculty member at several esteemed government and private organizations. Under his co-guidance, one Ph.D. scholar was awarded the degree in January 2024, while another is expected to submit the thesis in April 2025. He is currently guiding five Ph.D. scholars, demonstrating his ongoing dedication to research supervision. His scholarly contributions include numerous publications in Scopus and SCI-indexed journals, alongside several patents. His research interests lie in software engineering, the internet of things, cloud computing, and other emerging technologies. He can be contacted at email: m.rajanikanta@jainuniversity.ac.in.



Dr. Sarappadi Narasimha Prasad    is professor in Department of Electronics and Electronics Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, Karnataka, India. His total experience is 22 years, completed graduation from Mangalore University, Post-graduation from Visvesvaraya Technological University and doctorate from Jain University. More than 80 journals/conferences in profile and presently guiding 8 research scholars. His area of interest is AI, embedded systems, and signal processing. He can be contacted at email: prasad.sn@dsu.edu.in.