IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, September 2025 Impact Factor: 7.67

Improving Diabetes Diagnosis through Ensemble Machine Learning Models and Data Imbalance Handling Methods

Rasila Rohitdas Jawalkar[1] and Dr. Vilas Shivaji Gaikwad[2]
[1]Research Scholar, Department of Computer Science
[2]Research Guide, Department of Computer Science
Vikrant University, Gwalior (M.P.)

Abstract: *Diabetes mellitus is a chronic metabolic disorder that requires early and accurate diagnosis to prevent severe health complications. However, traditional diagnostic approaches often struggle with limited accuracy, especially when dealing with imbalanced clinical datasets where non-diabetic cases significantly outnumber diabetic ones. This study proposes an advanced framework for improving diabetes diagnosis by integrating ensemble machine learning models with effective data imbalance handling techniques..*

Keywords: Diabetes Diagnosis, Ensemble Learning, Machine Learning

I. INTRODUCTION

Diabetes mellitus has emerged as one of the most pressing global health challenges of the 21st century, affecting hundreds of millions of individuals worldwide and placing a substantial burden on healthcare systems. Characterized by chronic hyperglycemia resulting from defects in insulin secretion, insulin action, or both, diabetes can lead to severe complications such as cardiovascular disease, kidney failure, neuropathy, and vision impairment if not diagnosed and managed effectively. Early and accurate diagnosis is therefore critical in preventing disease progression and reducing long-term health risks. However, traditional diagnostic approaches often rely on limited clinical indicators, manual interpretation, and threshold-based decision-making, which may not fully capture the complex and multifactorial nature of the disease.

In recent years, advancements in artificial intelligence (AI) and machine learning (ML) have opened new avenues for enhancing medical diagnosis. Machine learning models have demonstrated significant potential in analyzing large-scale healthcare data, identifying hidden patterns, and providing predictive insights that surpass conventional statistical methods. Among these, ensemble machine learning models—techniques that combine multiple individual models to improve predictive performance—have gained particular attention for their robustness, accuracy, and generalization capabilities. By leveraging the strengths of different algorithms, ensemble methods such as bagging, boosting, and stacking can reduce variance, minimize bias, and achieve superior diagnostic outcomes compared to single-model approaches.

Despite these advantages, one of the major challenges in applying machine learning to diabetes diagnosis is the issue of data imbalance. In many real-world medical datasets, the number of non-diabetic cases significantly outweighs the number of diabetic cases. This imbalance can lead to biased models that favor the majority class, resulting in poor detection of minority class instances—often the very cases that require the most attention. For example, a model trained on imbalanced data may achieve high overall accuracy while failing to correctly identify patients with diabetes, thereby limiting its clinical utility.

Addressing data imbalance is therefore a crucial step in developing reliable diagnostic systems. Various data-level and algorithm-level techniques have been proposed to mitigate this issue. Data-level methods include resampling

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, September 2025 Impact Factor: 7.67

techniques such as oversampling the minority class, undersampling the majority class, and generating synthetic samples using approaches like the Synthetic Minority Over-sampling Technique (SMOTE). These methods aim to balance the class distribution and provide the model with sufficient representative examples of both classes. On the other hand, algorithm-level strategies involve modifying the learning process itself, such as incorporating cost-sensitive learning, adjusting class weights, or designing specialized loss functions that penalize misclassification of minority class instances more heavily.

The integration of data imbalance handling methods with ensemble machine learning models presents a promising solution for improving diabetes diagnosis. Ensemble techniques can be combined with resampling strategies to create more balanced training datasets, thereby enhancing the model's ability to detect diabetic cases. For instance, boosting algorithms like AdaBoost and Gradient Boosting can be adapted to focus more on misclassified instances, while bagging-based approaches such as Random Forest can be enhanced through balanced sampling techniques. Additionally, hybrid models that incorporate both data preprocessing and ensemble learning have shown significant improvements in classification performance, particularly in terms of sensitivity (recall) and F1-score, which are critical metrics in medical diagnosis.

Another important aspect of improving diabetes diagnosis through machine learning is the selection and utilization of relevant features. Diabetes is influenced by a wide range of factors, including age, body mass index (BMI), glucose levels, blood pressure, genetic predisposition, lifestyle habits, and more. Feature selection techniques can help identify the most informative attributes, reduce dimensionality, and eliminate noise, thereby improving model efficiency and interpretability. When combined with ensemble learning, these techniques can further enhance predictive accuracy by ensuring that each base learner focuses on meaningful patterns within the data.

Moreover, the increasing availability of electronic health records (EHRs), wearable device data, and large-scale medical datasets has created new opportunities for developing data-driven diagnostic systems. However, these datasets often come with challenges such as missing values, noise, and heterogeneity, which must be addressed through appropriate preprocessing techniques. Data cleaning, normalization, and imputation play a vital role in ensuring the quality and reliability of the input data, ultimately influencing the performance of machine learning models.

The application of ensemble machine learning models with data imbalance handling methods is not only limited to improving diagnostic accuracy but also contributes to the development of decision support systems for healthcare professionals. Such systems can assist clinicians in making informed decisions by providing risk predictions, identifying high-risk patients, and suggesting personalized treatment plans. This can lead to more efficient resource allocation, early intervention, and improved patient outcomes.

Furthermore, the interpretability of machine learning models is an important consideration in the healthcare domain. While ensemble models are often perceived as complex and less transparent, recent advancements in explainable AI (XAI) have made it possible to understand and interpret model predictions. Techniques such as feature importance analysis, SHAP (Shapley Additive Explanations), and LIME (Local Interpretable Model-agnostic Explanations) can provide insights into how different features contribute to the diagnosis, thereby increasing trust and acceptance among healthcare practitioners.

Improving diabetes diagnosis through the integration of ensemble machine learning models and data imbalance handling methods represents a significant advancement in the field of medical informatics. By addressing the limitations of traditional diagnostic approaches and overcoming challenges associated with imbalanced datasets, these techniques offer a powerful and effective solution for early and accurate detection of diabetes. As research in this area continues to evolve, the development of more sophisticated models, improved data preprocessing methods, and enhanced interpretability tools will further strengthen the role of machine learning in transforming healthcare and combating the global burden of diabetes.

CHALLENGES IN DIABETES DIAGNOSIS USING MACHINE LEARNING

The primary challenges include:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, September 2025 Impact Factor: 7.67

Class imbalance: Unequal distribution between diabetic and non-diabetic samples.
Feature overlap: Clinical features often overlap between classes.
Noise and missing data: Medical datasets may contain incomplete or noisy records.
Model bias: Standard classifiers tend to favor the majority class.
These challenges necessitate advanced approaches such as ensemble learning and imbalance handling techniques to improve diagnostic accuracy.

ENSEMBLE MACHINE LEARNING MODELS

Ensemble learning combines multiple base learners to improve overall performance. Common ensemble techniques include:
Bagging (e.g., Random Forest)
Boosting (e.g., Gradient Boosting, XGBoost)
Voting classifiers
Stacking ensembles
Ensemble methods reduce variance and bias while improving generalization ability. In diabetes prediction, ensemble models have shown superior performance compared to single classifiers.
For example, studies integrating voting-based ensembles with decision trees, random forest, and logistic model trees achieved high predictive performance with improved AUC scores when combined with SMOTE balancing techniques.

DATA IMBALANCE HANDLING METHODS

A. Oversampling Techniques

SMOTE (Synthetic Minority Over-sampling Technique) generates synthetic samples for the minority class.
Variants include Borderline-SMOTE and ADASYN.
SMOTE improves minority class representation and reduces bias in classification models.

B. Under sampling Techniques

Random Undersampling (RUS)
Tomek Links
Edited Nearest Neighbors (ENN)
These methods reduce majority class samples to balance the dataset but may discard useful information.

C. Hybrid Techniques

Combination of oversampling and undersampling (e.g., SMOTE + RUS)
These approaches balance datasets while preserving informative samples.

D. Cost-Sensitive Learning

Assigns higher misclassification costs to minority class.
Useful when resampling is not desirable.

INTEGRATION OF ENSEMBLE LEARNING AND IMBALANCE HANDLING

Combining ensemble models with resampling techniques enhances diabetes prediction performance. Key strategies include:
Applying SMOTE before training ensemble classifiers
Using balanced random forests
Combining SMOTE with voting or stacking ensembles
Feature-based ensemble learning with resampling
A feature-based ensemble model integrating SMOTE and random undersampling with random forest classifiers showed improved accuracy and AUC, outperforming several baseline models and demonstrating better minority class detection.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, September 2025 Impact Factor: 7.67

COMPARATIVE TABLE OF METHODS

Approach	Technique	Advantages	Limitations	Reported Outcome
Single Classifier	Logistic Regression / SVM	Simple, interpretable	Poor performance on imbalanced data	Low sensitivity to minority class
Oversampling + Single Model	SMOTE + Random Forest	Improves minority class learning	Risk of overfitting	Improved recall and AUC
Undersampling + Ensemble	RUS + Random Forest	Reduces majority bias	Loss of information	Moderate performance improvement
Hybrid Resampling + Ensemble	SMOTE + RUS + Voting Ensemble	Balanced dataset, robust predictions	Computational cost	High accuracy (~0.87) and AUC (~0.92)
Stacking Ensemble	Multiple base learners + meta-model	Captures complex patterns	Complex implementation	High ROC-AUC and generalization
Balanced Random Forest	Random Forest with class balancing	Handles imbalance internally	May require tuning	Competitive performance

EVALUATION METRICS FOR IMBALANCED DATA

Accuracy alone is not sufficient for imbalanced datasets. Common evaluation metrics include:

Precision

Recall (Sensitivity)

F1-score

ROC-AUC

PR-AUC

ROC-AUC is particularly useful for evaluating discrimination ability across thresholds. Ensemble models combined with SMOTE often achieve high AUC values, indicating strong classification performance.

DISCUSSION

Improving diabetes diagnosis using ensemble machine learning models, combined with effective data imbalance handling techniques, has become a critical area of research in healthcare analytics. Diabetes datasets are often highly imbalanced, where the number of non-diabetic cases significantly outweighs diabetic cases. This imbalance can lead to biased models that achieve high overall accuracy but perform poorly in identifying minority class instances, which in this case are the actual diabetic patients. Therefore, addressing both model performance and data distribution is essential for reliable diagnosis.

Ensemble learning methods, such as Random Forest, Gradient Boosting, and AdaBoost, have shown strong potential in improving predictive accuracy. These models work by combining multiple base learners to produce a more robust and generalized prediction system. For instance, Random Forest reduces variance by aggregating multiple decision trees, while Gradient Boosting focuses on minimizing errors iteratively. This collaborative approach helps capture complex patterns in medical data, such as nonlinear relationships between glucose levels, BMI, age, and insulin levels, which are often missed by single models.

However, even the most sophisticated ensemble models can struggle when trained on imbalanced datasets. This is where data imbalance handling techniques play a crucial role. Methods such as oversampling, under sampling, and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, September 2025 Impact Factor: 7.67

hybrid approaches are widely used. Oversampling techniques like SMOTE (Synthetic Minority Oversampling Technique) generate synthetic samples of the minority class, helping balance the dataset without losing important information. On the other hand, under sampling reduces the number of majority class samples, though it may risk discarding useful data. Hybrid methods attempt to combine the strengths of both approaches to achieve better results.

Another important aspect is feature selection and preprocessing. High-quality input data enhances the effectiveness of both ensemble models and imbalance handling techniques. Removing irrelevant or redundant features, normalizing data, and handling missing values can lead to more accurate predictions. Furthermore, cross-validation strategies, especially stratified sampling, ensure that the imbalance ratio is maintained during training and testing, leading to more reliable evaluation. However, challenges remain:

- Risk of overfitting due to synthetic data
- Increased computational complexity
- Need for optimal parameter tuning
- Data privacy concerns in clinical datasets

Future research may explore deep ensemble models, hybrid sampling with generative models, and explainable AI (XAI) to improve interpretability in clinical decision-making.

II. CONCLUSION

Improving diabetes diagnosis through the integration of ensemble machine learning models and effective data imbalance handling methods represents a significant advancement in modern healthcare analytics. Diabetes, being a chronic and often asymptomatic condition in its early stages, requires timely and accurate diagnosis to prevent severe complications. Traditional diagnostic approaches, while clinically valuable, may fall short in handling large-scale, complex, and imbalanced datasets commonly found in medical records. This is where machine learning, particularly ensemble techniques, offers transformative potential.

Ensemble models, such as Random Forest, Gradient Boosting, and Voting Classifiers, combine the strengths of multiple algorithms to produce more robust and reliable predictions. By aggregating diverse learners, these models reduce the risk of overfitting and improve generalization across unseen data. Their ability to capture non-linear relationships and interactions between features makes them particularly suitable for medical diagnosis, where multiple factors such as glucose levels, BMI, age, and genetic predisposition interact in complex ways.

Furthermore, the integration of these approaches enhances model interpretability and reliability two crucial factors in healthcare applications. Clinicians are more likely to trust and adopt machine learning systems that demonstrate consistent performance across diverse patient populations and provide explainable insights. Ensemble models, when paired with feature importance analysis, can also help identify key risk factors, contributing to better clinical decision-making.

REFERENCES

[1]. Alghamdi, M., Al-Mallah, M., Keteyian, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach. *PLoS ONE, 12*(7), e0179805. https://doi.org/10.1371/journal.pone.0179805

[2]. Sampath, P., Elangovan, G., Ravichandran, K. (2024). Robust diabetic prediction using ensemble machine learning models with SMOTE. *Scientific Reports, 14*, 28984.

[3]. Jang, Y. (2025). Feature-based ensemble modeling for addressing diabetes data imbalance using SMOTE and RUS. *Ewha Medical Journal*. https://doi.org/10.12771/emj.2025.00353

[4]. Kivrak, M., Avci, U., Uzun, H. (2024). Impact of SMOTE on machine learning performance in diabetes-related imbalance datasets. *Diagnostics, 14*(23), 2634. https://doi.org/10.3390/diagnostics14232634

[5]. Khan, A. A., Chaudhari, O., & Chandra, R. (2023). A review of ensemble learning and data augmentation models for class imbalanced problems. *arXiv preprint*.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, September 2025 Impact Factor: 7.67

[6]. Islam, M. N., Rimon, M. M. H. (2025). An improved ensemble-based machine learning model with feature optimization for early diabetes prediction. *arXiv preprint*.

[7]. Khokhar, P. B., Pentangelo, V., Palomba, F., & Gravino, C. (2025). Towards transparent and accurate diabetes prediction using machine learning and XAI. *arXiv preprint*.