Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ28428

Enhancing Type 2 Diabetes Prediction via SMOTEENN and Weighted Voting Classifiers: Balancing Recall and Accuracy in Imbalanced Medical Data

Yuchen Chen

Fryeburg Academy, Fryeburg, USA
744715336@qq.com

Abstract. Diabetes is a chronic disease with significant health and economic impacts worldwide. Early prediction of type 2 diabetes is critical for timely intervention and prevention of severe complications. This study evaluates multiple machine learning classifiers—Logistic Regression, Random Forest, XGBoost, and AdaBoost—along with two configurations of a Voting Classifier, to identify patients at risk of diabetes using clinical and demographic data. To address class imbalance, the SMOTEENN technique was applied, combining oversampling with noise removal. Models were assessed on Accuracy, Recall, Precision, and Macro-F1 score, with a primary focus on recall for the positive (diabetes) class, given its significance in clinical screening. Results indicate that Random Forest achieved the highest accuracy (0.81), whereas the weighted Voting Classifier—with increased weight assigned to XGBoost—achieved the highest recall (0.87), though at the expense of overall accuracy. These findings underscore the trade-off between recall and precision in diagnostic modeling. They also suggest that model choice should be context-dependent: recall-optimized models for high-risk screening, and balanced models for general population screening.

Keywords: Diabetes prediction, Machine learning, SMOTEENN, Voting Classifier, Recall

1. Introduction

Type 2 diabetes mellitus (T2DM) is a major global health challenge. It affects hundreds of millions of people worldwide. Moreover, early detection is important because it helps prevent serious complications such as heart disease, kidney failure, and nerve damage. Machine learning (ML) has become an important tool in healthcare. It can identify high-risk individuals using clinical and demographic data.

Class imbalance is a common problem in medical prediction, where the number of negative cases is usually much larger than the number of positive ones. As a result, many models focus too much on the majority class: they predict negatives well but often fail to detect positives, which leads to low recall for the minority class. In diabetes research, this means that patients with the disease may be overlooked.

DOI: 10.54254/2755-2721/2026.TJ28428

Earlier studies have shown the same issue in other health domains. Chawla et al. reported that traditional classifiers lose sensitivity when trained on imbalanced data [1]. They proposed SMOTE as a way to improve recall for minority cases. Fernández et al. reviewed imbalance learning methods and noted that oversampling with noise reduction can make models more robust [2]. However, this may also reduce precision. In diabetes prediction, Sisodia and Sisodia found that ensemble methods improved accuracy but still struggled to reach high recall [3]. In short, recall remains a major challenge in medical screening, and missing cases can delay treatment and cause serious harm to patients.

To address this gap, the present study investigates both individual classifiers and ensemble strategies for diabetes prediction under class imbalance. Specifically, we apply the SMOTEENN technique to balance the dataset while reducing noise. We then compare the performance of Logistic Regression, Random Forest, XGBoost, and AdaBoost. Beyond single models, we also test two Voting Classifier configurations: one with equal weights and another with increased weight for XGBoost. This design allows us to see if ensemble methods can improve recall for positive cases while still keeping reasonable precision and macro-F1 scores. By examining the trade-offs between accuracy, recall, and precision, this study provides guidance for choosing models in different clinical settings.

2. Methods

2.1. Dataset

The dataset used in this study was the Behavioral Risk Factor Surveillance System (BRFSS) 2015 Diabetes Health Indicators dataset, a large-scale health survey conducted by the U.S. Centers for Disease Control and Prevention (CDC). The BRFSS is one of the world's largest continuously conducted health surveys, collecting data on health-related risk behaviors, chronic health conditions, and the use of preventive services from U.S. adults.

The 2015 dataset includes more than 400,000 survey responses, of which a subset was extracted for diabetes-related prediction tasks. The target variable indicates whether the respondent reported having been diagnosed with diabetes by a healthcare professional (positive = 1, negative = 0). Given the nature of survey-based collection, the dataset is inherently imbalanced, with negative cases substantially outnumbering positive cases.

The predictor variables encompass a wide range of demographic, behavioral, and clinical health indicators, including:

- Demographics: age group, gender, education level, and income category.
- Behavioral factors: smoking status, alcohol consumption, physical activity, and dietary habits.
- Health indicators: body mass index (BMI), general health status (self-reported), high blood pressure, and high cholesterol.

The Kaggle version of the dataset was already numerically encoded (binary, ordinal, or continuous values), making it directly suitable for model training.

2.2. Data preprocessing and imbalance handling

To prepare the data for model training, continuous variables such as BMI were standardized to a zero mean and unit variance, ensuring comparability across models sensitive to scale. Missing or implausible values, though limited in the Kaggle version, were addressed by median imputation for continuous features and mode imputation for categorical features.

DOI: 10.54254/2755-2721/2026.TJ28428

A substantial class imbalance was present, with negative (non-diabetic) cases far exceeding positive (diabetic) cases. To address this issue, the SMOTEENN (Synthetic Minority Oversampling Technique combined with Edited Nearest Neighbors) method was applied. SMOTE generated synthetic minority samples to improve the representation of positive cases, while ENN removed noisy or ambiguous samples from both classes. This hybrid approach both increased sensitivity to minority cases and improved overall dataset robustness.

2.3. Models evaluated

To evaluate predictive performance, we selected a diverse set of machine learning models that represent both linear and non-linear classifiers, as well as ensemble approaches commonly applied in medical data mining.

Logistic Regression. We used Logistic Regression as a baseline model. It is simple and easy to explain, which makes it a common choice in clinical prediction tasks. To handle imbalance, we applied the class_weight='balanced' option so that the minority (diabetes-positive) class received more weight during training. This model served as a benchmark for evaluating more advanced methods.

Random Forest. Random Forest builds many decision trees and combines their results. This approach works well even when the data has noise and can capture non-linear patterns. In our study, we trained 200 trees (n_estimators=200) and enabled class balancing. We selected Random Forest because it often performs well in medical prediction tasks and is less likely to overfit compared to a single decision tree [4].

XGBoost. XGBoost is a boosting algorithm that builds trees in sequence, with each new tree correcting the errors of the previous ones. It is widely used in healthcare prediction tasks because it can capture complex interactions. In this study, we used 200 trees with a learning rate of 0.1 and a maximum depth of 5. We also adjusted the scale_pos_weight parameter to address class imbalance. XGBoost was included because it provides strong predictive power while remaining efficient to train [5].

AdaBoost. AdaBoost combines several weak learners, often shallow decision trees, to form a stronger model. Each new learner focuses more on the mistakes of earlier ones. We used 200 estimators with a learning rate of 0.1. AdaBoost is known to improve performance in moderately imbalanced datasets by adaptively giving more weight to difficult cases.

Voting Classifier. In addition to individual models, we used Voting Classifiers that combine predictions from multiple algorithms. We tested two versions: one where each model had equal weight, and another where XGBoost was given a higher weight (2.0). The purpose of this design was to see if combining models could improve recall for positive cases while still keeping a balance with precision and accuracy.

2.4. Evaluation protocol

Data was split into training (80%) and test (20%) sets using stratified sampling to preserve class ratios.

Evaluation metrics:

- Accuracy = (TP+TN)/(Total)
- Recall (Positive Class) = TP/(TP+FN)
- Precision (Positive Class) = TP/(TP+FP)
- Macro-F1 = Mean of F1 scores across classes

DOI: 10.54254/2755-2721/2026.TJ28428

Given the clinical setting, recall for the positive class was prioritized over accuracy.
All experiments were conducted in Python 3.10 using scikit-learn 1.3.0 and XGBoost 1.7.6.

3. Results

Table 1. Performance of different models on the BRFSS 2015 diabetes dataset

Model	Accuracy	Recall (Positive Class)	Precision (Positive Class)	Macro F1
Logistic Regression	0.73	0.78	0.31	0.63
Random Forest	0.81	0.48	0.38	0.66
XGBoost	0.73	0.79	0.31	0.63
AdaBoost	0.8	0.57	0.36	0.66
Voting Classifier (Equal Weights)	0.71	0.8	0.29	0.62
Voting Classifier (Weighted, XGBoost Higher)	0.67	0.87	0.27	0.59

As shown in Table 1, Random Forest achieved the highest accuracy but relatively low recall. XGBoost exhibited strong recall with moderate accuracy. The weighted Voting Classifier achieved the highest recall but showed the lowest accuracy and precision.

4. Discussion

The results illustrate a clear trade-off between accuracy and recall. Random Forest's superior accuracy (0.81) is offset by its low recall (0.48), making it less suitable for high-risk medical screening. Conversely, the weighted Voting Classifier maximized recall (0.87) at the cost of accuracy (0.67) and precision (0.27).

In a clinical context, false negatives (missed diagnoses) are more critical than false positives, particularly in high-risk populations where early detection significantly improves prognosis. Under these circumstances, the recall-focused weighted Voting Classifier is preferable. However, for general population screening—where minimizing false alarms is also important—XGBoost or an equal-weight Voting Classifier offers a better balance between recall and precision.

These findings are consistent with prior studies highlighting the recall–precision trade-off in imbalanced medical data [2]. Notably, adjusting ensemble weights offers a simple yet effective mechanism for tailoring performance to clinical priorities.

5. Limitations

• Dataset size and feature diversity were limited to the available records, which may affect generalizability.

• No external validation dataset was used, so results may not fully translate to other populations.

• The analysis did not include deep learning models, which may capture complex nonlinearities in larger datasets.

6. Conclusion

This study compared several machine learning classifiers for predicting type 2 diabetes, with a focus on recall for the positive class. We looked at both individual models and ensemble methods. The weighted Voting Classifier reached the highest recall (0.87), which makes it useful in high-risk

DOI: 10.54254/2755-2721/2026.TJ28428

screening where missing a diagnosis is unacceptable. In contrast, XGBoost and the equal-weight Voting Classifier provided a better balance between recall and accuracy, which may be more practical for general screening.

Our findings highlight how different models behave under class imbalance and why the choice of model depends on the clinical context. Random Forest offered the highest accuracy but suffered from low recall. XGBoost performed well in both recall and overall balance. The weighted Voting Classifier maximized recall but gave up accuracy and precision. These comparisons give practical guidance for choosing models depending on whether recall or accuracy is more important.

However, despite these insights, there are some limits to this study. The dataset came only from the BRFSS 2015 survey, which may affect generalizability. We did not test on an external dataset, so the results may not hold in other populations. We also did not include deep learning methods, which could capture more complex relationships in larger datasets.

Future work will explore hyperparameter tuning and the use of additional clinical and lifestyle features. We will also test external datasets for validation. Another step will be trying advanced ensemble methods and deep learning to see if they can further improve recall and make the models more useful in practice.

References

[1] 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

[2] Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research, 61, 863–905. https: //doi.org/10.1613/jair.1.11192

[3] Sisodia, D. S., & Sisodia, D. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578–1585. https: //doi.org/10.1016/j.procs.2018.05.122

[4] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https: //doi.org/10.1023/A: 1010933404324

[5] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https: //doi.org/10.1145/2939672.2939785