A Comparative Analysis of Machine Learning Algorithms for the Early Prediction of Diabetes with an Evaluation of Class-Imbalance Handling

Ayorinde Oduroye, Temilade Opanuga, Esther Tosin Akanbi, Adeoye Abimbola Oluwadare, Ezekiel Obiruo Oyivbi, Ademola Francis Adebowale, Theophilus Bamise Ajala*

Department of Computer Science, College of Computing and Information Sciences, Caleb University, KM 15, Ikorodu-Itoikin Road, Imota, P.M.B. 21238, Lagos State, Nigeria

***Corresponding Author**

DOI: https://doi.org/10.51584/IJRIAS.2026.11080001

Received: 16 August 2026; Accepted: 21 August 2026; Published: 26 August 2026

ABSTRACT

Although early detection of diabetes significantly lowers its harm, machine learning models for screening are often assessed based on overall accuracy, a metric that is deceptive given the high-class imbalance that characterizes clinical data. In this work, we compared and evaluated the performance of five algorithms: logistic regression, naive Bayes, support vector machine (SVM), decision tree, and extreme gradient boosting (XGBoost) to early predict diabetes in many patients. Diabetes affected about 13.9% of the 253,680 records that were looked at in the 2015 Behavioral Risk Factor Surveillance Systems Diabetes Health Indicators dataset and to deal with this problem, people followed a process called CRISP-DM and they also used something called the Synthetic Minority Oversampling Technique or SMOTE for short. They used these things to retrain each model with the data, which was not balanced. Each model was then looked at using a few different measures, including the F1-score, accuracy, precision, recall and the area, under the ROC curve to see how well each Diabetes model was working. The Diabetes models were evaluated to see how well they were doing. Ten-fold cross-validation and a held-out test set were both used to validate the results. In the absence of imbalance handling, most models diagnosed less than one in five cases of diabetes with an accuracy of roughly 86%. While the decision tree improved somewhat and XGBoost was essentially unaffected, applying SMOTE increased the recall of the linear and probabilistic models from below 0.17 to above 0.76 at the expense of accuracy and precision. The best predictors were found to be high blood pressure, overall health, and high cholesterol. The study comes to the conclusion that headline accuracy is an unreliable guide in imbalanced medical prediction, that imbalance handling can change a model's practical usefulness, and that this benefit is strongly algorithm-dependent, meaning that the decision to resample should be based on the algorithm and the screening priorities rather than being applied consistently.

Keywords: Diabetes, detection, Synthetic Minority Oversampling Technique (SMOTE), machine learning, comparative analysis, XGBoost, Logistic regression, support vector machine (SVM)

INTRODUCTION

Diabetes mellitus is one of the most prevalent chronic diseases worldwide, and its global burden continues to grow. The International Diabetes Federation estimates that the number of adults living with diabetes has risen steeply over the past two decades and is projected to reach several hundred million more by 2045, with much of the increase concentrated in low- and middle-income countries (Sun et al., 2022). Most of the damages caused by type 2 diabetes result from its diagnosis at a very late stage because this form of diabetes tends to occur silently after many years of development, and a great number of individuals develop complications like heart disease, kidney problems, and visual impairment before being diagnosed.

The role of early diagnosis here comes into play. With an early diagnosis, through lifestyle changes and proper treatment, the disease's progression can be slowed and the possibility of complication avoided. This clinical fact

has led to the development of predictive testing since it becomes possible to identify individuals who have high risks of developing the disease based on regularly obtained health measures before the disease manifests itself. Machine learning is a good tool for doing this because clinical outcomes are mostly based on non-linear interaction between variables (Kavakiotis et al., 2017).

There have been many studies where machine learning techniques have been used for diabetes diagnosis with the help of logistic regression, naïve Bayes, support vector machines, decision trees, and gradient boosting ensemble algorithms. It should be noted that the results obtained in such studies tend to be very good; however, at the same time, the field is not yet mature enough because there is no consensus on the algorithm that works better since each study uses its own dataset and different measures for evaluating the outcomes.

The other problem is that of class imbalance. Most often than not, there will be many more patients who do not suffer from a particular ailment compared to those who actually suffer from the disease. The classifier could then use its training set to predict all cases as belonging to the majority class and thereby attain very good accuracy rates but at the expense of missing out on the minority cases which are crucial for the screening process. This could be very risky since the cost of a false negative is always much higher than a false positive. One technique that solves this problem is known as SMOTE (Chawla et al., 2002), but their effect on diabetes prediction is not consistently reported or systematically compared across algorithms.

Both gaps are filled by this study. Using a sizable, actual public health dataset, it compares five machine learning algorithms for early diabetes prediction and investigates the impact of resolving class imbalance by SMOTE on predictive accuracy, paying special attention to the identification of positive instances The specific goals include preprocessing and exploring the dataset, building predictive models using five algorithms, applying SMOTE and comparing performance with and without imbalance handling, evaluating and comparing the models using accuracy, precision, recall, F1-score, and the area under the ROC curve, and identifying the health indicators that most strongly contribute to prediction and suggesting the best modelling strategy for early screening.

Accordingly, the study is directed by three research questions: which algorithm on the dataset predicts diabetes the best; how SMOTE influences each algorithm's predictive performance, especially the identification of positive cases; and which indicators are most important for accurate prediction.

Significance of the Study

This research has multiple benefits. It reduces the uncertainty that occurs when results are derived from studies utilizing diverse data and methodologies by giving researchers a methodical, like-for-like comparison of five popular algorithms on a single huge dataset. It provides evidence on a component that significantly affects screening performance but is frequently disregarded by analysing imbalance handling openly rather than as an afterthought. The results provide useful advice on which modelling techniques are most effective for early diabetes detection and on the trade-offs associated with optimizing for the detection of positive cases for practitioners and health-technology developers. The most significant improvements in patient outcomes and cost savings can be achieved by early intervention, which is supported by work that increases the reliability of early screening instruments for the larger objective of preventive healthcare. Additionally, the study offers a repeatable methodology that may be used with region-specific datasets when they become available, such as those from Nigeria and other African contexts.

Related Work

Machine Learning in Healthcare and Diabetes Prediction

Machine learning has expanded rapidly across healthcare, dominating prediction, diagnosis, and management tasks, with supervised learning the most common paradigm (Kavakiotis et al., 2017). For diabetes specifically, the appeal lies in combining routine indicators such as blood pressure, body mass index, age, and general health status into a single predictive model. Nevertheless, the reliability of such models strongly relies on the quality of data, preprocessing, the selection of an algorithm, and, most importantly, the way evaluation is carried out.

Various comparative analyses in terms of the prediction of diabetes yield quite inconsistent results. In particular, some suggest support vector machine and random forest as the best performers, whereas others, who use ensemble and bagging techniques in conjunction with cross-validation, argue that random forest provides the highest accuracy of about 97% (Yadu et al., 2024). Other researchers, comparing several algorithms, claim that a neural network produces the highest accuracy, followed by logistic regression and support vector machine; it should be noted that earlier research proved superiority of decision-tree techniques compared to SVM (Wang et al., 2023). Similar is the case with deep learning and probabilistic techniques (Chou et al., 2023). There are several reasons why results vary; first of all, studies use different datasets of various sizes and quality, different preprocessing, tweaking of models, and publication of various measures, which prefer accuracy to recall. Hence, it can be concluded that there is no common opinion concerning the most effective algorithm, and mentioned accuracy of the late 1980s-early 1990s cannot be compared directly.

Class Imbalance and Its Handling

Class imbalance is known to be one of the factors affecting the effectiveness of screening. A classification algorithm trained on an unbalanced dataset may attain high overall accuracy by predicting the majority class at the expense of poor prediction of the minority class of interest. Therefore, the performance of the algorithm must be assessed based on its ability to identify cases from the minority class rather than accuracy (Salmi et al., 2024). This can be done by balancing the classes through various resampling algorithms. The most famous of these is SMOTE.

Importantly, the impact of the SMOTE technique is not always positive for all measures. It has been found that use of SMOTE may lead to lowering the accuracy of the model overall but significantly raising the level of sensitivity, or detection of positive examples. This is exactly the type of trade-off that will be measured in the current experiment using the five different algorithms. A literature analysis of a decade of research on imbalanced medical data demonstrates that performance of medical prediction models should be assessed primarily through measures related to minority class performance, including recall and F1-measure, not accuracy (Salmi et al., 2024).

The Five Algorithms

This work will assess five classifiers from different families of algorithms used in supervised learning. One such classifier is the logistic regression classifier, which employs a linear model for estimating the probability of a binary output based on the weighted sum of input features; the classifier is preferred because it is simple and interpretable and, more importantly, usually competes with complex classifiers (Wang et al., 2023), thus emphasizing that complexity does not necessarily result in better results. The naive Bayes classifier is also a probabilistic classifier, relying on Bayes' theorem and an independence assumption; despite its simplicity, it is efficient and works well but usually underperforms ensemble classifiers on complex data (Chou et al., 2023).

Although the decision tree's transparency is useful in clinical situations, individual trees are prone to overfitting and instability. The decision tree classifies data by recursively splitting on feature values, yielding understandable if-then rules. It serves as the conceptual basis for the ensemble that follows as well as a stand-alone interpretable model. Introduced by Chen and Guestrin (2016), Extreme Gradient Boosting (XGBoost) builds decision trees sequentially, with each new tree correcting the errors of its predecessors while a regularization term limits overfitting. Although its outputs are less directly interpretable than those of a single tree or a logistic model, it frequently achieves state-of-the-art results on tabular data and is a leading candidate in diabetes prediction.

METHODOLOGY

Research Design and Framework

Based on secondary data, the study uses an experimental, quantitative research methodology. It is experimental in the machine-learning sense that it trains competing models under controlled settings and compares their results on the same held-out data, and it is quantitative since it depends on numerical measurement of model

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VIII August 2026

performance. There is no collection of source data. This paper follows CRISP-DM process – an approach used to structure the research into problem understanding, data understanding, data preparation, modeling, evaluation, and conclusion. The use of such an approach makes methodology transparent and reproducible, as well as highlights the principle of evaluation that is key in imbalanced learning, stating that performance of models in medical screening should primarily be evaluated based on performance on the minority class (Salmi et al., 2024).

Dataset

Diabetes Health Indicators data set that was used in the current study was obtained from 2015 Behavioral Risk Factor Surveillance System (BRFSS) – the annual telephone survey about health related issues that is being carried out among the adult US population. 253,680 records with a binary diabetes label and 21 predictor factors covering health conditions, behaviors, and demographics make up the binary version used here. The data contained no missing values. A substantial class difference was found when the target variable was analyzed: 218,334 records (86.1%) did not have diabetes, while 35,346 records (13.9%) did. The primary disadvantage is that there may be self-report bias because the variables are self-reported rather than clinically measured. Therefore, rather than being diagnostic tools, the models are better viewed as risk-screening tools. However, the data provides a credible and well-documented foundation for assessment because it is derived from a sizable, real national survey.

Data Preprocessing

To get ready for modeling, the raw data underwent pre-processing. There were no missing records when the dataset was examined. Duplicate records were kept as legitimate separate observations. Duplicate records naturally occur in survey data when different respondents provide identical answers across the survey items. To prevent a single variable from controlling distance- or gradient-based algorithms, features measured on various scales were standardized. This was done for scale-sensitive algorithms, such as logistic regression, naive Bayes, and support vector machines, but it was not necessary for tree-based methods, which are scale-invariant. To ensure that reported performance represents behaviour on unseen data, the dataset was split into training and testing subsets using an 80:20 stratified split, with stratification maintaining the class proportions in both subsets. The test set was kept out and used exclusively for final evaluation.

Handling Class Imbalance

Because diabetic cases form the minority in the data, class imbalance was addressed explicitly using SMOTE (Chawla et al., 2002), which generates synthetic minority-class examples by interpolating between existing ones. Crucially, SMOTE was only used on the training set following the train-test split; it was never used on the test set. If it had been used prior to splitting, it would have leaked artificial information into the test data and inflated performance. In order to directly quantify the impact of imbalance management on each method by comparing the two situations, each model was trained twice: once on the original imbalanced training data and once on the SMOTE-balanced training data.

Validation, Metrics, and Tools

Two complementary validation strategies were used. Ten-fold stratified cross-validation was applied during model development to obtain robust estimates that do not depend on a single partition, dividing the training data into ten folds and averaging performance across them. For final reporting, each trained model was evaluated on the held-out test set that was never used in training or cross-validation. Because the data is imbalanced, overall accuracy is uninformative, since a classifier predicting the majority class for every record would achieve about 86% accuracy while detecting no diabetes at all; accuracy is therefore reported only for completeness. The metrics of substance are precision, recall (sensitivity), F1-score, and the area under the ROC curve. Feature-importance analysis was used to identify the indicators contributing most strongly to prediction. The analysis was implemented in Python using the scikit-learn library (Pedregosa et al., 2011), the imbalanced-learn library for SMOTE, and the XGBoost library for the gradient-boosting model.

Definition of Key Terms

The following definitions of the key terminology used in this study are provided for clarification. In the supervised learning job of classification, a model places each record into one of a predetermined category and in this case, diabetic or non-diabetic. When one class has significantly more instances than another due to unequal representation of the categories, this is known as class imbalance. By creating artificial examples of the minority class instead of replicating preexisting ones, SMOTE is a technique that resolves class imbalance. The percentage of real positive instances that a model accurately detects is known as recall, or sensitivity. This is an important screening parameter since missed cases can be expensive. Precision, which reflects the burden of false alarms, is the percentage of instances that are positive despite being projected to be positive. By calculating the harmonic mean of precision and recall, the F1-score offers a single, balanced metric. Area under the curve of the ROC (AUC), where a value of 0.5 means chance and 1.0 is for perfect separation, gives us the ability of the classifier to differentiate between classes over all decision thresholds.

RESULTS

Data Preparation

Following the procedure described above, the 253,680 records were divided into a training set of 202,944 records and a test set of 50,736 records, maintaining the 13.9% diabetic proportion in both. The test set was left unaltered to offer an objective assessment, and SMOTE was then applied solely to the training set, increasing the number of minority-class training instances to match the majority class and creating a balanced training set. The test data was subjected to the standardization that was fitted to the training data. The performance with and without imbalance handling is reported in the next subsections, which are followed by cross-validation and feature-importance findings.

Performance Without Imbalance Handling

Table 1 shows the performance of the five algorithms that were trained on the initial, unbalanced data and assessed on the held-out test set.

Table 1. Performance of models without SMOTE

Model	Acc.	Prec.	Recall	F1	AUC
Logistic Regression	0.862	0.517	0.158	0.242	0.819
Naive Bayes	0.772	0.320	0.564	0.408	0.780
Support Vector Machine	0.862	0.518	0.164	0.249	0.819
Decision Tree	0.862	0.519	0.162	0.247	0.808
XGBoost	0.863	0.524	0.177	0.264	0.820

Recall refers to detection of the positive (diabetic) class.

The class-imbalance issue is directly revealed by the results. While four of the five models which include logistic regression, support vector machine, decision tree, and XGBoost achieved excellent accuracy of about 86%, their recall was incredibly poor, ranging from 0.158 to 0.177. Practically speaking, these seemingly accurate algorithms missed over 80% of the very people a screening tool is meant to identify, detecting less than one in five real cases of diabetes; their high accuracy is attained by predicting the majority (non-diabetic) class for nearly all cases. Because its probabilistic decision mechanism is less biased toward the majority class and thus discovered more true cases at the expense of accuracy and precision, Naive Bayes reacted differently, exhibiting lower accuracy (0.772) but noticeably noticeably higher recall (0.564).
In conclusion, this analysis demonstrates why just accuracy is not enough and why the recall and F1 score are

the more relevant measures in this case. Area Under the ROC Curve turned out to be quite close in all cases between 0.780 and 0.820, meaning that models have equal ranking properties despite having extremely different behavior on the decision boundary. The underlying discriminatory power of the models appears to be quite similar as well, since their ROC curves cluster around 0.80-0.82; the modest ceiling denotes the limited predictive signal carried by self-reported indicators.

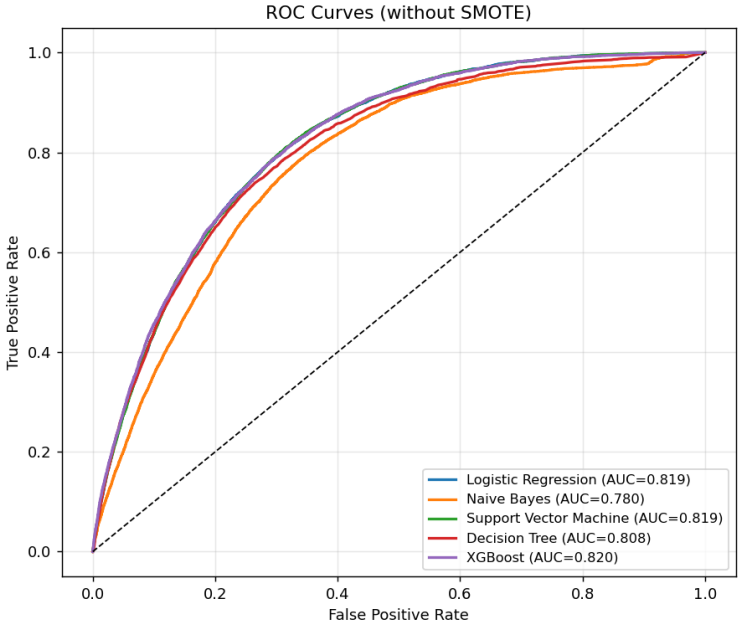

Figure 1. ROC curves for the five models (without SMOTE).

Performance With SMOTE

Table 2 reports the performance of the same five algorithms after SMOTE was applied to the training data.

Table 2. Performance of models with SMOTE

Model	Acc.	Prec.	Recall	F1	AUC
Logistic Regression	0.733	0.312	0.761	0.443	0.819
Naive Bayes	0.724	0.297	0.717	0.420	0.783
Support Vector Machine	0.732	0.311	0.761	0.442	0.819
Decision Tree	0.822	0.390	0.492	0.435	0.805
XGBoost	0.863	0.527	0.177	0.265	0.821

Compare with Table 1 to observe the effect of imbalance handling on each model.

The use of SMOTE resulted in a remarkable improvement in performance for all models. Recall increased for logistic regression from 0.158 to 0.761, for the support vector machine from 0.164 to 0.761, and for naive Bayes from 0.564 to 0.717. After the balance, these models were capable of detecting almost three-quarters of true cases of diabetes, and there was an entirely new quality in their applicability for diabetes screening. The improvement was paid for by a decrease in accuracy and precision of models due to the increase in false positives. For the discussed models, the F1-score, which is a measure balancing precision and recall, improved significantly – for example, for logistic regression, from 0.242 to 0.443.

Two models exhibited distinct behaviors that warrant attention. The decision tree showed a more gradual increase, with recall increasing to 0.492 and accuracy remaining higher at 0.822. SMOTE had little effect on XGBoost; its recall stayed at 0.177. The strongest ensemble method's ability to withstand oversampling is an important discovery. While the linear and probabilistic models are much more responsive to the rebalanced training distribution, it appears that XGBoost's internal handling of the objective and its regularization already

account for majority-class dominance in a way that leaves little for SMOTE to change at the default threshold.

Crucially, the model's area under the ROC curve remained essentially intact between the two situations, indicating that SMOTE modified the models' operating point trading precision for recall at the default decision threshold rather than changing their fundamental capacity to rank instances.

Cross-Validation

Using the F1-score as the summary metric, ten-fold stratified cross-validation was carried out on the training data to ensure that the results were not an artifact of a single train-test split. The averaged findings, displayed in Table 3, were in line with the results of the held-out test: XGBoost and naive Bayes had the greatest cross-validated F1 on the initial unbalanced data, whereas the other models had lower values. The small standard deviations indicate stable, reproducible performance across folds.

Table 3. Ten-fold cross-validation F1-scores (original training data)

Model	Mean F1	Std. dev.
Logistic Regression	0.245	0.011
Naive Bayes	0.413	0.006
Support Vector Machine	0.250	0.009
Decision Tree	0.246	0.029
XGBoost	0.271	0.010

F1-score reflects balanced performance on the minority class.

Feature Importance

The feature importances of the XGBoost model were analyzed in order to detect the most effective predictions. The ten strongest indicators are shown in Figure 2. Hypertension was responsible for almost 45% of the model's importance, thus becoming the largest indicator. Other indicators included age, heart disease, body mass index, high cholesterol, general health condition, and alcohol abuse. Such a sequence correlates well with the known clinical information about the risk factors of type 2 diabetes, which include connections between hypertension, poor lipid profile, and general health condition. Besides being an adequate basis for choosing indicators for screening, such a correspondence between learned indicators and known risk factors ensures the model's adequacy.

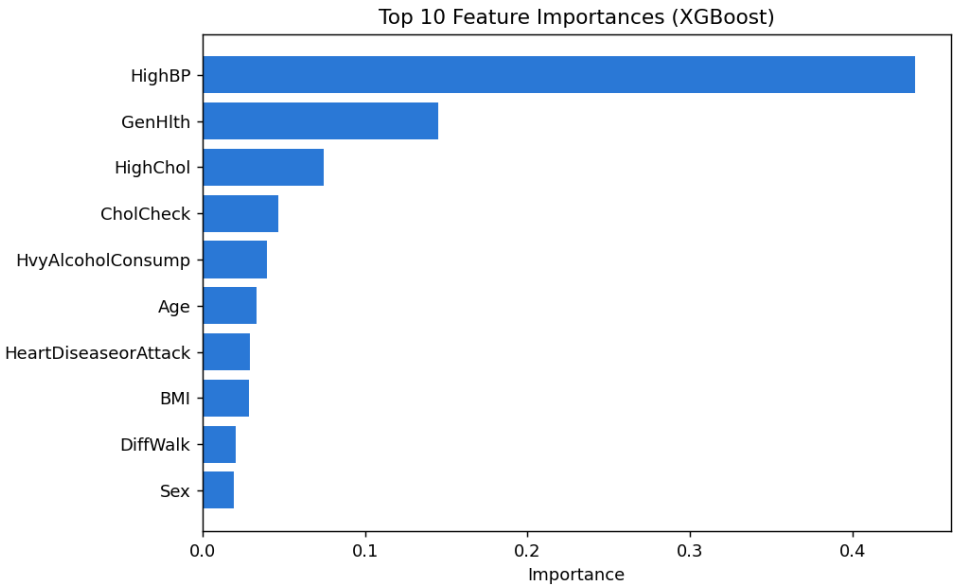

Figure 2. Top ten feature importances (XGBoost).

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VIII August 2026

DISCUSSION

All three research questions presented in the research are answered by the results obtained from the research. Regarding the question of what algorithm works better for the prediction of diabetes, the answer may vary depending on the metric. XGBoost shows some superiority in terms of the overall accuracy, but it is misleading since the high accuracy for unbalanced dataset means that the prediction is just majority class prediction. None of the models shows good performance without class imbalance handling using the metric that is important for screening - that is, positive case detection. Logistic regression and support vector machine have the best recall and the best combination of recall and precision after SMOTE.

The results offer an unambiguous and nuanced answer concerning the influence of SMOTE on model performance. While sacrificing accuracy and precision, SMOTE greatly increased the recall of linear and probabilistic models from less than 0.17 to more than 0.76 in the most successful cases. The F1-scores prove that such a trade-off is worth making for screening purposes. Minimal for XGBoost and significant for the decision tree, the influence is, therefore, irrelevant for some algorithms and highly valuable for others – an important point that should be taken into account but is often ignored in research that only presents aggregate results. This result confirms the literature on imbalanced learning, according to which resampling often increases sensitivity at the cost of accuracy (Salmi et al., 2024), and adds to that literature by showing how such an increase works on five algorithms using a real, large dataset.

According to clinical understanding of diabetes risk factors, high blood pressure, overall health, and high cholesterol appeared as the leading predictors. This gives the models face validity and shows which regularly gathered indicators are most useful for screening. When considered collectively, the results support the main thesis of the paper, which is that headline accuracy in unbalanced medical prediction is an unhelpful guide, that imbalance handling can change a model's practical usefulness, and that its advantage depends on the algorithm.

The results can be read alongside prior comparative studies, whose disagreements this study helps to explain. Reports that random forest or support vector machine is the strongest performer, sometimes exceeding 97% accuracy (Yadu et al., 2024), and reports favouring neural networks or logistic regression (Wang et al., 2023), are difficult to reconcile in the abstract, but the present findings suggest that much of the apparent disagreement stems from differences in how imbalance is handled and which metric is foregrounded. A study reporting high accuracy without reporting recall may be describing a model that performs as poorly on minority detection as the 86%-accuracy models in Table 1, whose recall fell below 0.18. When recall and F1-score are foregrounded, as the imbalanced-learning literature recommends (Salmi et al., 2024), the ranking of models changes substantially, and the effect of resampling becomes visible. This reinforces the value of a like-for-like comparison on a single dataset with a consistent evaluation protocol, which is what this study provides.

There is a clear practical implication of having developed the screening tool. For instance, screening that identifies a greater number of people for further confirmation tests is often acceptable when the cost of failing to diagnose is high. In such instances, where the objective is to maximize the detection of the true cases, the improved recall from SMOTE technique is worth the associated reduction in precision, assuming the number of false positives does not become unmanageable for the screening process. From the results, one can conclude that there is no contribution of resampling at the default threshold, when the chosen model is XGBoost, and that simply adjusting the threshold is a better approach for increasing recall compared to generating fake data. When implementing the model, one needs to consider the relationship between false negatives and false positives from the confusion matrix, while assessing the performance of the models based on recall and F1-score of the positive class instead of accuracy.

These findings are subject to some limitations, which need to be pointed out. Even though the dataset is large and genuine, it suffers from self-reporting bias due to the fact that the data is based on self-reported survey answers. Its predictive ceiling, which is represented by AUC values around 0.82, is limited by the characteristics of the indicators. The data's direct generalizability to Nigerian or other African patients is

limited because it comes from a population in the United States. The study used default or slightly adjusted model settings and assessed a single imbalance-handling method, SMOTE, at default decision thresholds. These

decisions maintain the comparison's control and interpretability while allowing for future research to compare a broader range of imbalance-handling strategies, including ADASYN, under sampling, hybrid approaches, and cost-sensitive learning; investigate decision-threshold optimization in conjunction with resampling, since changing the threshold may achieve SMOTE's recall gains without synthetic data; incorporate new algorithms and hyperparameter tweaking; and, most significantly, apply the pipeline to locally gathered clinical data from African and Nigerian populations so that the models and risk factors may be validated in the deployment environment. Since a screening advice is more likely to be trusted and followed when the elements influencing it are transparent, a related development would include model-explanation approaches so that physicians may evaluate individual predictions. Lastly, compared to the retrospective evaluation employed here, prospective evaluation—which compares a model's predictions with later confirmed diagnoses across time—would offer more convincing proof of practical utility.

Ethical and Deployment Considerations

The application of the current findings is shaped by the fact that a screening model's mistakes have actual repercussions since it influences decisions that have an impact on individuals. When a person who does not have diabetes gets marked as one, leading to the unnecessary hassle of additional tests and expense, we have what is known as the false positive. On the other hand, when a person with diabetes is not marked as one, leading to the missed opportunity of intervention, we get a false negative. It can thus be said that the balance between recall and precision outlined above is essentially an ethical concern, and the correct operational point will depend on both the resources that can be made available for follow-ups and the costs involved with both kinds of errors. When a predictive model is used for assigning an invasive test, it may well have different settings than one that is used for a population screen.

The data itself is another factor to consider. The models used here should not be applied to a clinical setting or another country without revalidation because they were trained on self-reported survey responses from a single national population. When a model whose training distribution is significantly different from the population it is applied to is used, there is a danger that the population's risk will be consistently mis-estimated, and the direction of this inaccuracy is not necessarily predictable. Because of this, the study identifies validation on local, clinically assessed data as a prerequisite for any practical application and presents its models as a scientific demonstration and a replicable template rather than as a deployable clinical tool.

CONCLUSION

Using a sizable public dataset, this work created and compared five machine learning models for the early prediction of diabetes, paying particular emphasis to the impact of class-imbalance handling. Five algorithms were trained on the Diabetes Health Indicators dataset using a CRISP-DM technique, both with the natural class distribution retained and after applying SMOTE. They were then assessed using criteria selected to prioritize the detection of positive instances. The risk of depending on accuracy for unbalanced data was highlighted by the fact that most models, in the absence of imbalance handling, achieved roughly 86% accuracy yet identified less than one in five diabetic cases. At the expense of accuracy and precision, using SMOTE increased the recall of logistic regression and the support vector machine to roughly 0.76 and that of naive Bayes to 0.72; the decision tree improved somewhat, but XGBoost was essentially unaffected. The best indicators were high blood pressure, overall health, and high cholesterol.

There are three conclusions that follow. First, recall and F1-score for the positive class are the proper screening criteria in unbalanced medical prediction; total accuracy is an unreliable and potentially deceptive indicator of a model's utility. Second, imbalance handling via SMOTE can change a model's usefulness by turning seemingly accurate but inefficient classifiers into ones that identify most genuine cases, but at a precision cost that must be balanced against the advantage of identifying more cases. Third, and perhaps most notably, the advantage of imbalance handling is highly algorithm-dependent: it was insignificant for XGBoost, moderate for the decision tree, and decisive for the linear and probabilistic models. Therefore, a general recommendation to use SMOTE is too straightforward; the choice should be based on the screening priorities and the algorithm. The workflow is repeatable and can be used with clinical data relevant to a certain location, such as data from Nigeria and other African settings, when it becomes available.

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. 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). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785
3. Chou, C.-Y., Hsu, D.-Y., & Chou, C.-H. (2023). Predicting the onset of diabetes with machine learning methods. Journal of Personalized Medicine, 13(3), Article 406. https://doi.org/10.3390/jpm13030406
4. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
6. Salmi, M., Atif, D., Oliva, D., Abraham, A., & Ventura, S. (2024). Handling imbalanced medical datasets: Review of a decade of research. Artificial Intelligence Review, 57(10), Article 273. https://doi.org/10.1007/s10462-024-10884-2
7. Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., Stein, C., Basit, A., Chan, J. C. N., Mbanya, J. C., Pavkov, M. E., Ramachandaran, A., Wild, S. H., James, S., Herman, W. H., Zhang, P., Bommer, C., Kuo, S., Boyko, E. J., & Magliano, D. J. (2022). IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Research and Clinical Practice, 183, Article 109119. https://doi.org/10.1016/j.diabres.2021.109119
8. Wang, S., Chen, R., Wang, S., Kong, D., Cao, R., Lin, C., Feng, W., Wang, Y., & Xu, H. (2023). Comparative study on risk prediction model of type 2 diabetes based on machine learning theory: A cross-sectional study. BMJ Open, 13(8), Article e069018. https://doi.org/10.1136/bmjopen-2022-069018
9. Yadu, S., Chandra, R., & Sinha, V. K. (2024). Comparing different machine learning techniques in predicting diabetes on early stage. Engineering Proceedings, 62(1), Article 20. https://doi.org/10.3390/engproc2024062020