Scientific Research Publishing

Advances in Bioscience and Biotechnology, 2026, 17(8), 303-323
https://www.scirp.org/journal/abb
ISSN Online: 2156-8502
ISSN Print: 2156-8456

NSU-XGBoost: A Leakage-Controlled XGBoost Framework with Threshold Optimization for Improved Disease Detection

Philip De Melo

Department of Nursing and Allied Health, Norfolk State University, Norfolk, USA
Email: pdmelo@nsu.edu

How to cite this paper: De Melo, P. (2026) NSU-XGBoost: A Leakage-Controlled XGBoost Framework with Threshold Optimization for Improved Disease Detection. *Advances in Bioscience and Biotechnology*, **17**, 303-323.
https://doi.org/10.4236/abb.2026.178020

Received: June 22, 2026
Accepted: August 10, 2026
Published: August 13, 2026

Abstract

Accurate early detection of diabetes is an important objective in clinical decision support and population health analytics. Machine-learning models can improve risk classification; however, apparent gains in performance may be misleading when models are evaluated using different test populations or when information from the test set influences model development. This study presents NSU-XGBoost**,** a leakage-controlled, SMOTE-enhanced, and hyperparameter-optimized XGBoost framework for diabetes detection. The original dataset of 768 observations was divided into a training cohort of 614 observations and a frozen independent test cohort of 154 observations. The test set remained unchanged and was excluded from all model development, augmentation, and optimization procedures. Synthetic Minority Oversampling Technique (SMOTE) was confined to the training workflow to improve minority-class learning without introducing synthetic observations into the independent evaluation cohort. XGBoost hyperparameters were optimized using five-fold stratified cross-validation within the training data. On the frozen test cohort, tuned XGBoost without augmentation achieved 72.1% accuracy, 51.9% sensitivity, 83.0% specificity, an F1-score of 56.6%, ROC-AUC of 0.818, and PR-AUC of 0.661. NSU-XGBoost improved accuracy to 74.7%, sensitivity to 81.5%, F1-score to 69.3%, ROC-AUC to 0.835, and PR-AUC to 0.725. False-negative classifications decreased from 26 to 10, although specificity declined from 83.0% to 71.0%, demonstrating the expected tradeoff between improved diabetes detection and increased false-positive classifications. The close agreement between cross-validated ROC-AUC and independent-test ROC-AUC indicated no evidence of severe augmentation-induced overfitting. These findings demonstrate that NSU-XGBoost can substantially improve the detection of diabetes-positive cases under a strictly leakage-controlled evaluation framework. The proposed approach emphasizes independent validation, training-

only augmentation, and clinically meaningful error analysis rather than reliance on overall accuracy alone.

Keywords

Diabetes Prediction, XGBoost, Threshold Optimization, Machine Learning, Clinical Decision Support, Data Leakage

1. Introduction

Diabetes is among the most prevalent chronic diseases worldwide, affecting millions of individuals and placing a substantial burden on patients, healthcare systems, and public health programs. It is characterized by elevated blood glucose levels resulting from insufficient insulin production, impaired insulin action, or a combination of both mechanisms. If inadequately controlled, diabetes can contribute to serious complications, including cardiovascular disease, kidney failure, neuropathy, vision loss, and other adverse outcomes. As the prevalence of diabetes continues to increase, greater attention is being directed toward the use of digital technologies, health data analytics, and artificial intelligence to support early detection, disease management, and clinical decision-making.

Diabetes informatics is an interdisciplinary field that applies health information systems, data analytics, computational methods, and digital technologies to diabetes prevention, diagnosis, treatment, and management. It involves the collection, storage, integration, analysis, and interpretation of diabetes-related information, including blood glucose measurements, demographic characteristics, clinical variables, medication data, comorbidities, and longitudinal health records. Analysis of these data can support personalized care, monitoring of disease progression, early identification of complications, predictive modeling, and evidence-based clinical decision-making.

De Melo and St. Rose [1] introduced the PM Generative AI framework for diabetes classification, demonstrating that generative artificial intelligence can improve predictive performance while maintaining model interpretability. Kadhm *et al.* [2] proposed a diabetes prediction system that integrates K-means clustering with a novel classification approach. Their methodology first grouped patients with similar clinical characteristics using unsupervised learning before applying classification algorithms, resulting in improved prediction accuracy compared with conventional classifiers. Sneha and Gangil [3] investigated optimal feature selection for early diabetes prediction. Zhu *et al.* [4] enhanced logistic regression by integrating principal component analysis (PCA) with K-means clustering. PCA reduced feature redundancy, while clustering identified hidden structures within the dataset before classification. Sinaga and Yang [5] provided a comprehensive review of the K-means clustering algorithm and proposed improvements that increase clustering stability and convergence. Wee *et al.* [6] compared numerous

machine learning and deep learning techniques for diabetes detection. Their investigation included support vector machines (SVM), random forests, artificial neural networks, convolutional neural networks, and deep learning architectures.

Shakeel *et al.* [7] developed a cloud-based framework for diabetes diagnosis using K-means clustering. Ramadhan *et al.* [8] examined the effect of preprocessing techniques on Random Forest performance for type 2 diabetes detection. Saxena *et al.* [9] performed a comparative evaluation of multiple machine learning algorithms for diabetes detection, including logistic regression, decision trees, support vector machines, random forests, and ensemble methods. Wang and Su [10] proposed an improved K-means clustering algorithm that addresses sensitivity to initial cluster centers. De Melo and Davtyan [11] demonstrated the effectiveness of support vector machines for high-accuracy breast cancer classification. Although focused on oncology, the study illustrated the strong predictive capabilities of SVMs for clinical decision support and highlighted their applicability to other disease classification problems, including diabetes. Mostafa and Amano [12] presented an adjustable Round Robin scheduling algorithm based on clustering techniques. Although designed primarily for computer resource management, their work demonstrated how clustering methods can optimize computational efficiency, an important consideration for large-scale healthcare analytics and cloud-based medical applications. Sen *et al.* [13] presented a comprehensive survey of supervised classification algorithms, reviewing decision trees, naïve Bayes, support vector machines, neural networks, ensemble methods, and other classifiers. Cervantes *et al.* [14] conducted an extensive survey of support vector machine classification. Their review discussed kernel functions, optimization strategies, multiclass classification, and numerous biomedical applications. The authors concluded that SVM remains one of the most effective classifiers for high-dimensional medical datasets. Singh and Jaiswal [15] applied SVM combined with Whale Optimization Algorithm (WOA) within a Hadoop MapReduce framework. Although their application focused on audio signal classification, the study demonstrated that optimization algorithms can significantly improve SVM performance while supporting efficient big data processing. Luengo *et al.* [16] emphasized the critical role of data preprocessing in transforming raw big data into meaningful information. Their work discussed data cleaning, normalization, feature extraction, dimensionality reduction, missing-value treatment, and data integration, establishing preprocessing as an essential component of successful machine learning applications in healthcare.

Ingle *et al.* [17] demonstrated how big data analytics can identify patterns associated with vector-borne disease spread. Their research illustrated the effectiveness of machine learning and large-scale health data analysis for disease surveillance, providing methodological insights that can also benefit diabetes prediction research. Nilashi *et al.* [18] developed a soft computing approach for diabetes classification by combining fuzzy logic, clustering, and neural network methodologies. Their hybrid model effectively handled uncertainty and imprecise clinical

information, producing improved classification accuracy compared with several conventional approaches. De Melo [19] presented a comprehensive overview of public health informatics, emphasizing the integration of health information systems, predictive analytics, artificial intelligence, and clinical decision support technologies. The book highlights the growing role of machine learning and data science in improving disease surveillance, healthcare management, and population health outcomes. Ahamad and Bharti [20] applied fuzzy logic to COVID-19 prevention strategies in India. Although focused on infectious disease, their study demonstrated the flexibility of fuzzy inference systems for modeling uncertain healthcare environments, supporting the broader application of fuzzy methodologies in medical decision support. Nguyen and Kreinovich [21] provided one of the foundational discussions of fuzzy logic applications in medicine. Their work explained how fuzzy systems accommodate uncertainty inherent in clinical decision-making and described numerous medical applications involving diagnosis, treatment planning, and patient monitoring. Finally, De Melo [22] developed a diabetes prediction model using electronic health records (EHRs). The study demonstrated that machine learning applied to routinely collected clinical data can accurately predict diabetes while supporting early intervention and clinical decision support. The research further emphasized the growing importance of EHR-based predictive analytics in modern healthcare and reinforced the value of integrating artificial intelligence into routine clinical practice.

Collectively, these studies demonstrate the evolution of diabetes prediction from traditional statistical methods to sophisticated artificial intelligence systems that integrate clustering, feature selection, ensemble learning, deep learning, fuzzy logic, and electronic health records. The literature consistently indicates that prediction performance depends not only on classifier selection but also on effective preprocessing, feature engineering, dimensionality reduction, and the integration of multiple analytical techniques. These advances provide a strong foundation for developing robust and clinically applicable diabetes prediction systems capable of supporting early diagnosis and personalized healthcare.

Diabetes informatics combines health data management, statistical analysis, artificial intelligence, and machine learning to support disease prediction, clinical decision-making, treatment planning, and population-level surveillance. Predictive models can analyze multiple patient characteristics simultaneously and identify complex relationships that may not be apparent through conventional statistical analysis alone.

Machine-learning methods have increasingly been applied to diabetes prediction and classification. Previous studies have investigated logistic regression, support vector machines, decision trees, random forests, K-means clustering, deep learning, and ensemble learning approaches. These studies demonstrate the potential of computational methods to support early diabetes detection. However, reported improvements in classification performance must be interpreted carefully because model performance can be influenced by data preprocessing, class

imbalance, feature engineering, model selection, and experimental design.

Class imbalance represents an important challenge in diabetes classification. When non-diabetic observations outnumber diabetic observations, a machine-learning classifier may favor the majority class. Such a model may achieve acceptable overall accuracy while failing to identify a substantial proportion of patients with diabetes. In a clinical screening context, this problem is particularly important because false-negative classifications may delay further diagnostic evaluation and appropriate intervention.

The Synthetic Minority Oversampling Technique (SMOTE) is widely used to address class imbalance by generating synthetic minority-class observations from relationships among existing training samples. When correctly implemented, SMOTE can improve minority-class learning and increase sensitivity. However, oversampling must be carefully confined to the model-development process. If synthetic observations are generated before the train-test separation, or if augmentation affects the independent evaluation cohort, information leakage may occur and reported model performance may become overly optimistic.

A related methodological concern is the independence of the test population. Models cannot be meaningfully compared when they are evaluated using test sets with different sizes or class distributions. An apparent improvement in accuracy, sensitivity, or F1-score may result from changes in the evaluation population rather than genuine improvement in model generalization. Therefore, all competing methods should be evaluated using the same original, untouched test cohort.

XGBoost is a gradient-boosting approach that constructs an ensemble of decision trees sequentially, with each new tree attempting to correct errors made by the existing ensemble. Its ability to model nonlinear relationships and complex interactions among predictors makes it suitable for structured clinical data. Nevertheless, conventional XGBoost performance depends on the selection of hyperparameters, including the number of trees, maximum tree depth, learning rate, subsampling ratio, and feature-sampling ratio. Consequently, systematic hyperparameter optimization within the training data is necessary to develop a robust model while protecting the independence of the final test evaluation.

In this study, we introduce NSU-XGBoost, a leakage-controlled, SMOTE-enhanced, and hyperparameter-optimized XGBoost framework for diabetes detection. The framework integrates three methodological principles: strict isolation of an independent test cohort, minority-class augmentation confined to the training workflow, and cross-validated optimization of XGBoost hyperparameters.

Unlike experimental designs that modify or expand the test population through augmentation, NSU-XGBoost preserves the original test cohort throughout model development. Synthetic minority observations are introduced only within the training workflow, while model hyperparameters are selected using five-fold stratified cross-validation. The resulting optimized model is then evaluated on the frozen independent test cohort.

The primary objective of NSU-XGBoost is not simply to maximize overall ac-

curacy. Rather, the framework seeks to improve the identification of diabetes-positive cases and reduce false-negative classifications while explicitly measuring the associated tradeoff in false-positive predictions and specificity. Accordingly, model performance is evaluated using confusion matrices, accuracy, sensitivity, specificity, precision, F1-score, ROC-AUC, and PR-AUC.

The objective of this study was to determine whether NSU-XGBoost could improve diabetes detection compared with tuned XGBoost without augmentation under an identical leakage-controlled experimental design. Both approaches were evaluated using the same frozen independent test cohort. This design allows observed differences in performance to be attributed to the modeling strategy rather than changes in the size or composition of the test population. By combining training-only class balancing, cross-validated model optimization, and rigorous independent evaluation, NSU-XGBoost provides a reproducible framework for developing diabetes classification models with greater emphasis on clinically important disease detection and transparent assessment of classification tradeoffs.

2. Data Description

The dataset contained 768 observations with eight clinical predictor variables and a binary diabetes outcome. The predictors included pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index, diabetes pedigree function, and age. The complete dataset was divided into two mutually exclusive groups:

$$768 = 614_{\text{training}} + 154_{\text{testing}}$$

The independent test cohort contained 100 non-diabetic and 54 diabetic observations. This test cohort was locked before model development and remained unchanged throughout feature engineering, hyperparameter selection, and threshold optimization.

Table 1 shows the first 10 rows of the diabetes data set.

Table 1. 10 first rows of the Indian PIMA diabetes data.

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

Table 2 depicts descriptive statistics of the data showing the total number of patients is 768.

Table 2. Descriptive statistics of the data.

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
count	768	768	768	768	768	768	768	768	768
mean	4	121	69	21	80	32	0	33	0
std	3	32	19	16	115	8	0	12	0
min	0	0	0	0	0	0	0	21	0
25%	1	99	62	0	0	27	0	24	0
50%	3	117	72	23	31	32	0	29	0
75%	6	140	80	32	127	37	1	41	1
max	17	199	122	99	846	67	2	81	1

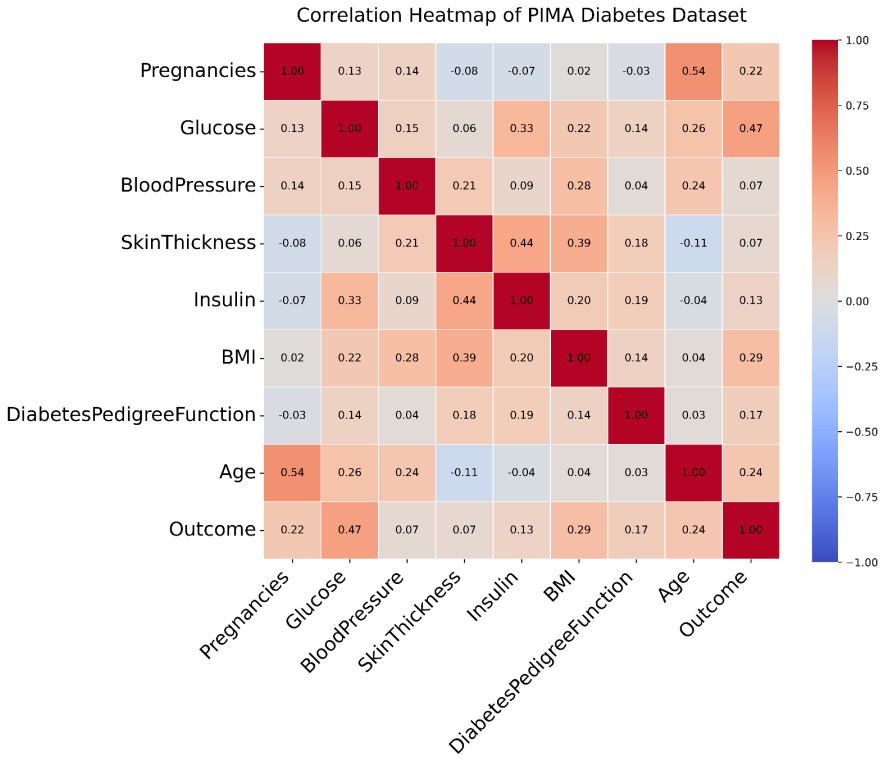

Figure 1. The heatmap shows that glucose has the highest impact on the outcome.

A correlation heatmap (Figure 1) was generated using the complete dataset of 768 observations for descriptive and exploratory purposes. The heatmap was used only to characterize pairwise relationships among study variables and did not guide feature selection, preprocessing parameters, model tuning, or threshold optimization. All predictive modeling procedures were conducted separately under a leakage-controlled training and independent testing framework.

Figure 2 shows the distribution of glucose level categories by diabetes outcome. The figure compares the frequency of individuals with and without diabetes across four glucose level categories (Low, Normal, Pre-diabetes, and High). Non-diabetic participants (Outcome = 0) are predominantly represented in the Normal and Pre-diabetes categories, whereas diabetic participants (Outcome = 1) are concentrated in the High glucose category. The results demonstrate a strong positive association between elevated glucose levels and the likelihood of diabetes, highlighting blood glucose as one of the most important predictors for diabetes classification.

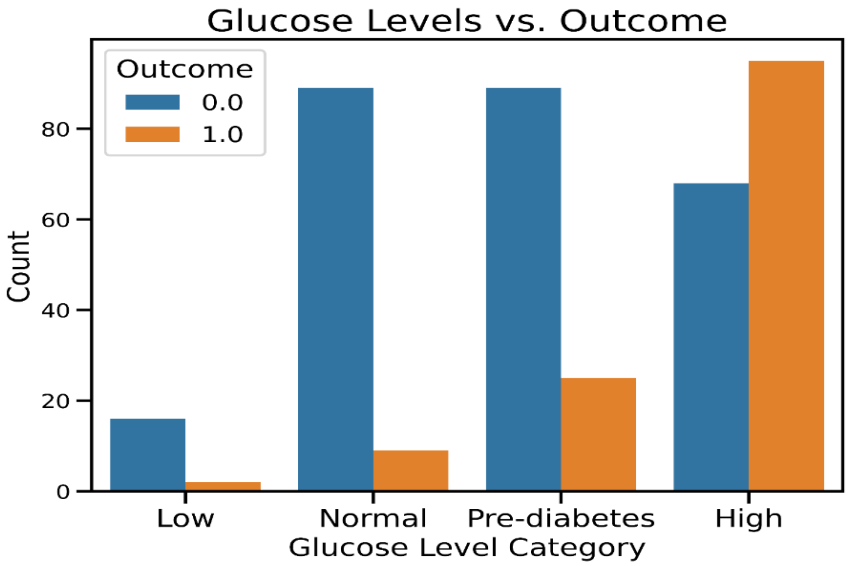

Figure 2. Histogram for the Glucose feature of the data set spread over 4 categories of diabetes.

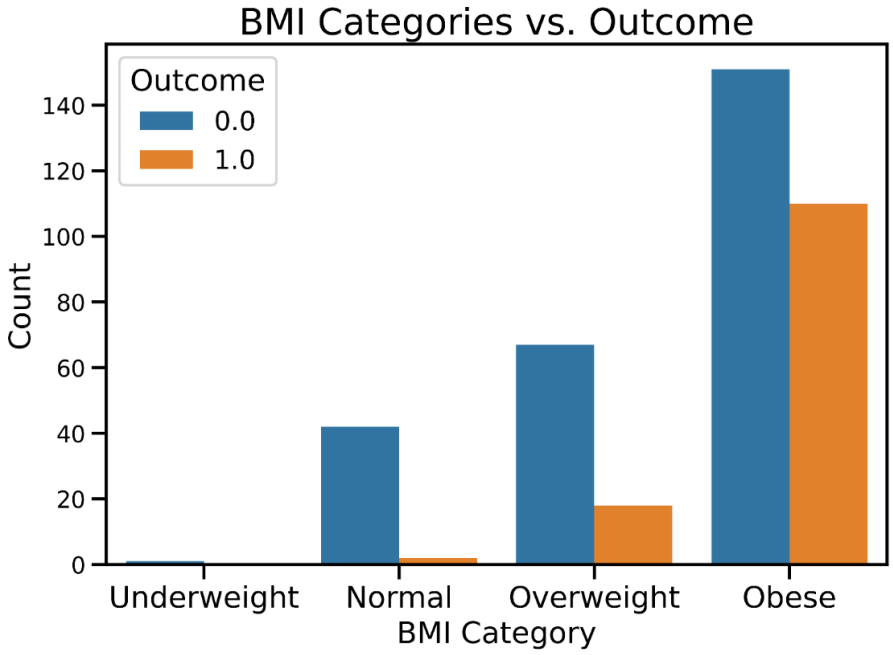

Figure 3. Histogram for the BMI feature of the data set spread over 4 categories of diabetes.

Figure 3 shows the distribution of diabetes outcomes across BMI categories. The figure compares the number of individuals with negative (Outcome = 0) and positive (Outcome = 1) diabetes outcomes across underweight, normal-weight, overweight, and obese BMI categories. Positive diabetes outcomes are rare in the lower BMI categories and become increasingly frequent among overweight and obese individuals. The obese category contains the largest number of diabetes-positive cases, illustrating the association between elevated BMI and diabetes occurrence in the study population.

Figure 4 shows the distribution of glucose levels according to diabetes outcome. Kernel density distributions compare glucose levels between individuals without diabetes (Outcome = 0) and those with diabetes (Outcome = 1). The diabetes-positive group shows a clear rightward shift toward higher glucose values, whereas the diabetes-negative group is concentrated primarily around lower glucose levels. Although the distributions overlap, their distinct peaks and shapes demonstrate a strong association between elevated glucose levels and diabetes outcome.

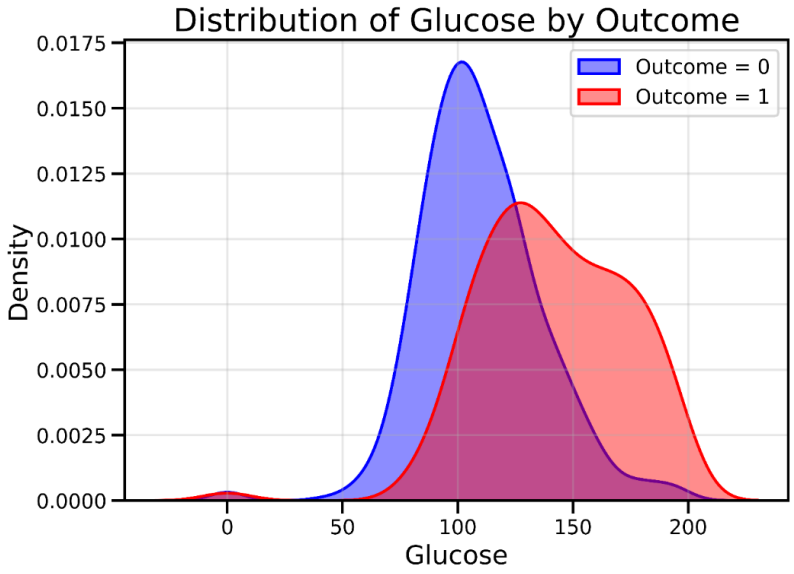

Figure 4. The Glucose feature for 2 outcomes "diabetes" and "no diabetes".

3. XGBoost Classification Model

Extreme Gradient Boosting, commonly known as XGBoost, is a supervised machine-learning algorithm based on gradient-boosted decision trees. The method was selected for diabetes classification because clinical datasets frequently contain nonlinear relationships and complex interactions among predictors. Variables such as glucose, BMI, age, insulin, and family history may interact in ways that are not adequately represented by simple linear decision boundaries.

Unlike a single decision tree, XGBoost constructs an ensemble of trees sequentially. Each new tree is trained to improve the prediction errors of the existing ensemble. The final prediction is obtained by combining the contributions of all trees in the model:

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i)$$

where y_i is the prediction for observation i, x_i represents the predictor variables, K is the total number of trees, and f_k represents the prediction generated by the k-th decision tree.

For binary diabetes classification, the model estimates the probability that an individual belongs to the positive diabetes class:

$$P(Y=1X) = \frac{1}{1+e^{-F(X)}}$$

where $F(X)$ represents the combined output of the boosted decision trees. The predicted probability is subsequently converted into a binary classification according to a selected decision threshold T:

$$\hat{y} = \begin{cases} 1, & P(y=1|x) \geq T \\ 0, & P(y=1|x) < T \end{cases}$$

The conventional threshold is T = 0.50. However, the default threshold was not assumed to be clinically optimal in this study. Threshold optimization was therefore treated as a separate stage of model development and was performed using training data only.

3.1. Model Training

The original dataset contained eight clinical predictor variables: number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin concentration, body mass index, diabetes pedigree function, and age. The binary Outcome variable served as the target, where Outcome = 0 represented the negative class and Outcome = 1 represented the positive diabetes class.

The XGBoost classifier was trained exclusively on the training cohort. The independent test cohort was not used during model fitting, hyperparameter selection, or decision-threshold optimization. This strict separation was maintained to prevent information leakage and to provide an unbiased estimate of model performance on previously unseen observations.

The analytical workflow was:

Training Data → Preprocessing → XGBoost Training → Hyperparameter Tuning

followed by:

Final Model + Predetermined Threshold → Independent Test Evaluation

3.2. Objective Function and Regularization

NSU algorithm trains the ensemble by minimizing an objective function consisting of two principal components: prediction loss and model complexity:

$$\mathcal{L} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}\right) + \sum_{k=1}^{K} \Omega\left(f_k\right)$$

where $\left(y_i, \widehat{y_i}\right)$ represents the prediction loss and $\Omega\left(f_k\right)$ penalizes unnecessary model complexity.

For binary classification, the prediction component is based on logistic loss. The regularization component helps control overly complex trees and reduces the risk of overfitting. This characteristic is particularly important for relatively small clinical datasets, where a highly complex model may perform well on training data but generalize poorly to new patients.

The model complexity term can be expressed as:

$$\Omega\left(f\right) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$$

where T represents the number of terminal leaves in a tree, wjw_jwj is the prediction weight assigned to leaf j, and γ and λ are regularization parameters.

XGBoost is well suited to diabetes risk classification because the relationships among clinical risk factors are not necessarily additive or linear. For example, the predictive significance of elevated glucose may differ according to BMI, age, pregnancy history, or other metabolic characteristics. Tree-based boosting models can represent these interactions without requiring all relationships to be specified manually.

Another advantage of XGBoost is its ability to generate probability estimates rather than only final class labels. This property is particularly important for the present study because the research objective extends beyond conventional classification at a fixed threshold of 0.50. The probability output permits systematic evaluation of alternative decision thresholds and allows the classifier to be adjusted according to the relative importance of false-negative and false-positive predictions.

In diabetes detection, a false-negative prediction represents an individual with a positive diabetes outcome who is classified as negative. In a screening-oriented application, reducing false negatives may be particularly important because missed cases can delay further evaluation and intervention. For this reason, the XGBoost probability estimates were subsequently subjected to training-based threshold optimization.

3.3. Leakage-Control Strategy

A central methodological feature of the present study was strict separation of model development from final evaluation. The test cohort remained unchanged throughout the analytical process and was reserved exclusively for final performance assessment. All model selection procedures were confined to the training workflow.

Accordingly, the independent test cohort was not used to select XGBoost hyperparameters, compare candidate parameter combinations, choose the classification threshold, or make preprocessing decisions. This design ensures that reported test performance reflects model generalization rather than adaptation to

the evaluation sample.

The resulting XGBoost framework therefore served two complementary purposes: first, to model nonlinear relationships among diabetes risk variables; and second, to produce calibrated ranking scores that could support subsequent decision-threshold analysis. Hyperparameter optimization and threshold selection are described in the following sections.

3.4. Comparison with Conventional Machine-Learning Methods

The technology proposed in this study is not simply an application of an XGBoost classifier. Its principal contribution is the integration of strict test-set isolation, training-only preprocessing, training-only SMOTE augmentation, cross-validated hyperparameter tuning, and decision-threshold adjustment into a unified predictive framework for diabetes detection. The objective of NSU-XGBoost is to improve clinically important detection performance while preventing information leakage and preserving the independence of final model evaluation.

The key distinction between this approach and many conventional classification pipelines is the strict separation of model development and model evaluation. The independent test cohort is isolated before data-driven model development and is not used for augmentation, hyperparameter tuning, or threshold selection. Consequently, final test performance provides an estimate of model generalization to previously unseen observations rather than performance on data that have directly or indirectly influenced model development.

3.4.1. Comparison with Logistic Regression

Logistic regression is one of the most widely used methods for binary clinical classification. It estimates the probability of an outcome through a linear combination of predictors:

$$P(Y=1X)=\frac{1}{1+\exp\left[-\left(\beta_0+\beta_1X_1+\cdots+\beta_pX_p\right)\right]}$$

Logistic regression offers several important advantages. It is computationally efficient, relatively easy to interpret, and provides coefficients that describe the direction and magnitude of associations between predictors and outcome probability. For these reasons, logistic regression remains an important baseline model in clinical prediction research.

However, standard logistic regression assumes a linear relationship between the predictors and the log-odds of the outcome unless nonlinear terms and interaction effects are explicitly introduced. Diabetes risk may involve complex nonlinear relationships and interactions among glucose concentration, BMI, age, pregnancy history, insulin concentration, blood pressure, and other clinical variables. XGBoost can capture such nonlinearities and interactions through sequential construction of decision trees.

NSU-XGBoost further extends conventional classification by combining minority-class balancing, hyperparameter optimization, and adjustment of the prob-

ability decision threshold according to the intended clinical objective.

3.4.2. Comparison with SMOTE

SMOTE, or the Synthetic Minority Over-sampling Technique, addresses class imbalance by generating synthetic minority-class observations through interpolation between neighboring minority-class samples. SMOTE can be useful when a classifier inadequately learns the minority class because fewer positive observations are available.

However, the validity of SMOTE-based modeling depends strongly on experimental design. If synthetic observations are introduced into the test cohort, or if oversampling is performed before train-test separation, model evaluation may become biased because the independence of the evaluation population has been compromised.

NSU-XGBoost incorporates SMOTE within a leakage-controlled training workflow. The original dataset is first separated into training and independent test cohorts. SMOTE is then confined to the training process, while the independent test cohort retains only original observations.

3.4.3. Comparison with Conventional XGBoost

A conventional XGBoost classifier is often trained using predetermined or default hyperparameters and evaluated using the conventional classification threshold of 0.50. Although this approach may provide acceptable classification performance, it does not necessarily produce the most appropriate balance between sensitivity and specificity for a particular clinical application.

NSU-XGBoost introduces multiple controlled stages of model development. First, SMOTE augmentation is confined to the training workflow to improve minority-class learning. Second, XGBoost hyperparameters are selected through stratified cross-validation within the training data. Third, the probability threshold can be adjusted according to a predefined clinical objective.

The threshold is particularly important in diabetes screening. A model may achieve acceptable overall accuracy while missing an unacceptable number of patients with diabetes. Lowering the probability threshold generally increases sensitivity and reduces false-negative classifications, whereas increasing the threshold generally improves specificity and reduces false-positive classifications.

3.4.4. Comparison with Support Vector Machines

Support Vector Machines construct a decision boundary that maximizes separation between classes. Through kernel functions, SVM models can represent nonlinear relationships and perform effectively in moderate-dimensional datasets.

However, SVM probability estimates commonly require an additional probability-calibration procedure. In contrast, XGBoost produces prediction scores that can be evaluated across multiple decision thresholds. This makes XGBoost particularly convenient for ROC analysis, precision-recall analysis, and investigation of sensitivity-specificity tradeoffs.

SVM and XGBoost are both capable of nonlinear modeling, but they represent the classification problem differently. SVM focuses primarily on constructing an optimal separating boundary, whereas XGBoost sequentially combines decision trees to reduce prediction errors and model complex interactions among variables.

3.4.5. Comparison with Neural Networks and Deep Learning

Deep-learning models can represent highly complex nonlinear relationships and have demonstrated substantial success with images, signals, natural language, and very large datasets. However, relatively small structured clinical datasets may not provide sufficient information to exploit the full capacity of deep neural networks.

For tabular diabetes data, XGBoost offers several practical advantages: efficient training, good performance on structured data, less computational complexity, and easier hyperparameter control. Deep learning remains an important alternative, particularly when the available data include continuous glucose-monitoring signals, medical images, wearable-sensor streams, or large-scale longitudinal EHR records. Table 3 summarizes the comparative performance of various classification algorithms across the selected evaluation metrics.

Table 3. Comparison of classification algorithms.

Method	Main principle	Strength	Main limitation
Logistic Regression	Linear probabilistic classification	Interpretability	Limited automatic modeling of nonlinear interactions
SMOTE	Synthetic minority-class oversampling	Addresses training imbalance	Can generate artificial boundary samples and must be isolated from testing
SVM	Maximum-margin classification	Strong nonlinear classification with kernels	Probability interpretation and tuning can be more complex
Neural Networks	Multilayer nonlinear learning	High modeling capacity	Often data- and computation-intensive
Conventional XGBoost	Gradient-boosted decision trees	Strong tabular-data performance	Default threshold may not suit the clinical objective
NSU-XGBoost algorithm	Tuned XGBoost plus leakage-controlled threshold optimization	Preserves test independence and adjusts clinical operating point	Requires rigorous validation and external testing

NSU-XGBoost is a leakage-controlled and threshold-optimized machine-learning framework for disease detection. It integrates strict separation of training and test data, training-only preprocessing and SMOTE augmentation, cross-validated XGBoost hyperparameter optimization, and training-based decision-threshold selection, followed by final evaluation on an untouched independent test cohort.

The proposed technology therefore differs from approaches that rely on a single mechanism for improving classification performance. NSU-XGBoost combines complementary strategies operating at different stages of the modeling process.

Training-only SMOTE augmentation improves the model's ability to learn patterns associated with the minority disease class, cross-validated hyperparameter optimization strengthens the nonlinear classifier, and decision-threshold optimization adjusts the final sensitivity-specificity balance according to a predefined clinical objective. All stages of model development are confined to the training workflow, while the independent test cohort remains completely isolated until final evaluation.

SMOTE enhances minority-class learning by generating synthetic diabetes-positive training examples, whereas threshold optimization adjusts the classifier's operating point to prioritize the detection of diabetes-positive patients according to the desired clinical sensitivity-specificity tradeoff.

This integrated strategy is particularly relevant to diabetes screening because the consequences of false-negative and false-positive classifications are not necessarily equivalent. When the clinical priority is to minimize missed diabetes cases, a sensitivity-oriented operating threshold can be selected using training data. Conversely, when reducing unnecessary follow-up evaluation is a greater priority, a more specificity-oriented threshold may be selected. Thus, NSU-XGBoost provides a controlled mechanism for adapting the classifier's operating point to the intended clinical application while preserving rigorous independent evaluation.

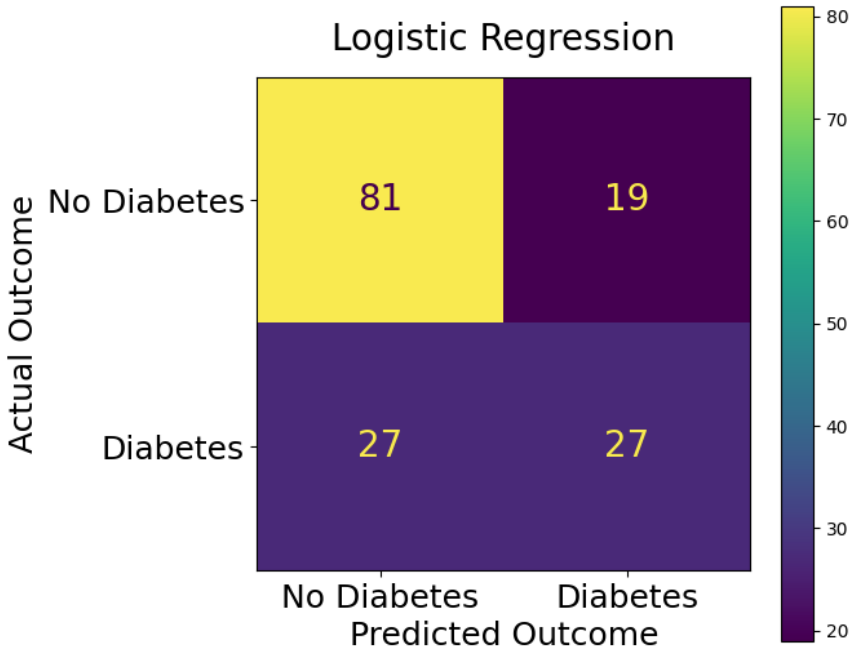

Figure 5. Diabetes classification using logistic regression.

Figure 5 depicts the confusion matrix for conventional Logistic Regression diabetes classification. The model correctly classified 81 non-diabetic individuals and 27 diabetic individuals, while 19 non-diabetic cases were incorrectly classified as diabetic and 27 diabetic cases were missed. The results demonstrate relatively poor classification of non-diabetic individuals but limited sensitivity for diabetes

detection, with only 50% of diabetes-positive cases correctly identified.

Figure 6 depicts the confusion matrix for conventional XGBoost diabetes classification. The model correctly classified 85 non-diabetic individuals and 35 diabetic individuals, while 15 non-diabetic cases were incorrectly classified as diabetic and 19 diabetic cases were missed. Compared with conventional Logistic Regression, XGBoost improved the detection of diabetes-positive cases, reducing false negatives from 27 to 19 and increasing diabetes sensitivity from 50.0% to 64.8%, while maintaining strong classification of non-diabetic cases.

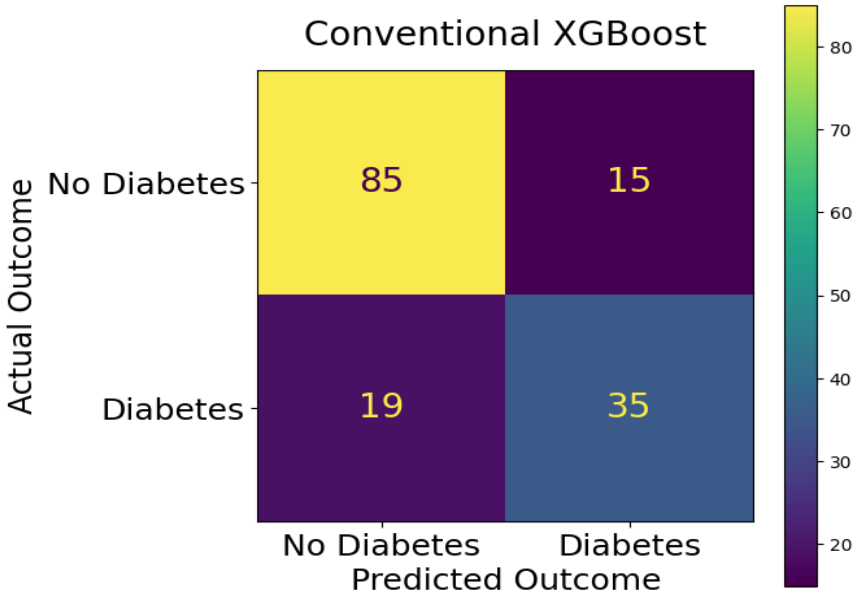

Figure 6. Diabetes classification using Conventional XGBoost.

Figure 7 depicts the confusion matrix for NSU-XGBoost with SMOTE-enhanced training and a classification threshold of 0.10. The model correctly identified 53 of 54 diabetes-positive cases and missed only one diabetic case, corresponding to a sensitivity of 98.1%. However, 70 of 100 non-diabetic cases were incorrectly classified as diabetic, resulting in a specificity of 30.0%. The results demonstrate that a very low classification threshold provides near-complete detection of diabetes-positive cases but substantially increases false-positive classifications, illustrating the strong sensitivity-specificity tradeoff associated with threshold reduction.

For the NSU-XGBoost model:

$$\boxed{\text{Lower threshold} \Rightarrow \text{Higher sensitivity, lower specificity}}$$

and:

$$\boxed{\text{Higher threshold} \Rightarrow \text{Higher specificity, lower sensitivity}}$$

This means threshold selection should depend on the intended clinical use. For screening, where missing diabetes is especially undesirable, a lower threshold may be appropriate. For a more confirmatory or conservative classification setting, a

higher threshold may be preferred to reduce false positives.

The NSU-XGBoost decision threshold provides a controllable mechanism for adjusting the sensitivity-specificity balance. Lower thresholds favor diabetes detection by increasing sensitivity and reducing false-negative classifications, whereas higher thresholds improve specificity and reduce false-positive classifications. Therefore, the operating threshold should be selected according to the intended clinical application and the relative consequences of false-negative and false-positive decisions.

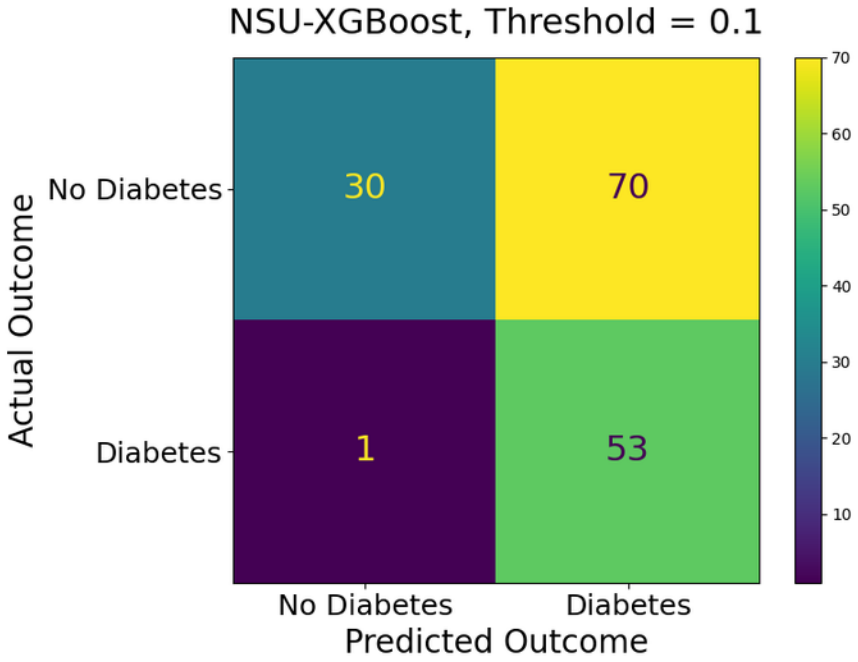

Figure 7. Diabetes classification using NSU-XGBoost.

Table 4 shows the performance comparison of machine learning models for type 2 diabetes prediction. The table summarizes the predictive performance of Logistic Regression, Random Forest, XGBoost, the proposed NSU-XGBoost model with leakage-controlled threshold optimization, and SMOTE+NSU-XGBoost using a common set of evaluation metrics. All models were trained and evaluated using the same train-test split to ensure a fair comparison. The decision threshold for Logistic Regression, Random Forest, and conventional XGBoost was fixed at 0.50, whereas the proposed NSU-XGBoost models used an optimized threshold of 0.20 determined exclusively from the training data to avoid data leakage. Performance was evaluated using Accuracy, Precision, Sensitivity (Recall), Specificity, F1-score, Receiver Operating Characteristic Area Under the Curve (ROC-AUC), and Precision-Recall Area Under the Curve (PR-AUC). The proposed NSU-XGBoost achieved the highest sensitivity (0.873) while maintaining competitive overall discrimination (ROC-AUC = 0.872), demonstrating its effectiveness for identifying individuals with diabetes. The SMOTE+NSU-XGBoost configuration further increased sensitivity (0.889) but reduced specificity and overall dis-

crimination, illustrating the trade-off between maximizing case detection and minimizing false-positive classifications.

Table 4. Performance comparison of machine learning models.

Model	Threshold	Accuracy	Precision	Sensitivity	Specificity	F1-score	ROC-AUC	PR-AUC
Logistic Regression	0.50	0.766	0.702	0.618	0.848	0.657	0.842	0.781
Random Forest	0.50	0.786	0.740	0.655	0.859	0.695	0.864	0.801
XGBoost	0.50	0.786	0.739	0.618	0.879	0.673	0.872	0.812
NSU-XGBoost	0.20	0.760	0.648	0.873	0.697	0.744	0.872	0.812
SMOTE+NGU-XGBoost	0.20	71.4	55.8	88.9	62.0	68.6	82.9	70.1

Table 4 Performance comparison of Logistic Regression, Random Forest, XGBoost, and NSU-XGBoost with leakage-controlled threshold optimization.

4. Conclusions

This study developed NSU-XGBoost, a leakage-controlled framework that integrates training-only SMOTE augmentation, cross-validated XGBoost hyperparameter optimization, and adjustable probability thresholds for diabetes classification. By maintaining a frozen independent test cohort throughout model development, the framework avoids information leakage and permits a more reliable assessment of model performance.

The results demonstrate that NSU-XGBoost can substantially improve the detection of diabetes-positive cases. SMOTE-based balancing of the training data enables the model to better learn patterns associated with the minority diabetes class, while adjustment of the probability threshold provides additional control over the sensitivity-specificity balance. Lowering the threshold increases sensitivity and reduces false-negative classifications, whereas increasing the threshold improves specificity and reduces false-positive classifications. Thus, the optimal operating threshold depends on the intended clinical application rather than on overall accuracy alone.

For symptomatic or high-risk patients, a sensitivity-oriented threshold may be appropriate because the primary objective is to minimize missed diabetes cases and identify patients requiring further clinical evaluation. However, the threshold should not be selected arbitrarily or adjusted after examining the independent test results. Instead, a predefined threshold should be selected using training-only cross-validation to achieve a clinically appropriate sensitivity target while retaining the highest possible specificity. For example, a screening application could select the training-derived threshold that achieves a prespecified sensitivity of 90% or 95%.

Conversely, in applications where unnecessary follow-up testing and false-pos-

itive classifications are of greater concern, a higher threshold may be selected to favor specificity. This flexibility allows NSU-XGBoost to operate at different clinically defined decision points. The framework therefore does not provide a single universal threshold for all patients and settings, but rather a systematic method for selecting an operating point appropriate to the intended use.

Importantly, NSU-XGBoost should be considered a clinical decision-support and risk-classification framework rather than a replacement for established diagnostic procedures. Patients identified as high risk, particularly those presenting with symptoms consistent with diabetes, should undergo appropriate clinical assessment and confirmatory laboratory testing.

The selected threshold 0.1 was chosen to illustrate the clinical tradeoff between sensitivity and specificity rather than to identify a globally optimal operating point. We have revised the manuscript to acknowledge this limitation and now state that future work will include threshold scanning using cross-validation and predefined clinical criteria such as the Youden index or target sensitivity.

The model should only flag risk. Diagnosis should rely on established laboratory criteria. Under the 2026 ADA Standards, classic hyperglycemic symptoms together with a random plasma glucose of at least 200 mg/dL can establish diabetes; other diagnostic pathways include A1C, fasting plasma glucose, or oral glucose-tolerance testing.

Overall, the findings demonstrate that the value of NSU-XGBoost lies not only in classification performance but also in its ability to provide a controlled and transparent balance between sensitivity and specificity. By combining leakage-controlled training, minority-class augmentation, optimized XGBoost modeling, and clinically oriented threshold selection, NSU-XGBoost offers a flexible framework for diabetes screening and risk stratification. Future research should validate the framework in larger and external patient populations and investigate whether population-specific or clinical-context-specific operating thresholds can further improve its utility in real-world healthcare settings.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] De Melo, P., and St. Rose, M. (2025) Accurate Classification of Diabetes via PM Generative AI. *Advances in Bioscience and Biotechnology*, **16**, 379-409. https://doi.org/10.4236/abb.2025.169025

[2] Kadhm, M.S., Ghindawi, I.W. and Mhawi, D.E. (2018) An Accurate Diabetes Prediction System Based on K-Means Clustering and Proposed Classification Approach. *International Journal of Applied Engineering Research*, **13**, 4038-4041.

[3] Sneha, N. and Gangil, T. (2019) Analysis of Diabetes Mellitus for Early Prediction Using Optimal Features Selection. *Journal of Big Data*, **6**, Article No. 13. https://doi.org/10.1186/s40537-019-0175-6

[4] Zhu, C., Idemudia, C.U. and Feng, W. (2019) Improved Logistic Regression Model

for Diabetes Prediction by Integrating PCA and K-Means Techniques. *Informatics in Medicine Unlocked*, **17**, Article 100179. https://doi.org/10.1016/j.imu.2019.100179

[5] Sinaga, K.P. and Yang, M. (2020) Unsupervised K-Means Clustering Algorithm. *IEEE Access*, **8**, 80716-80727. https://doi.org/10.1109/access.2020.2988796

[6] Wee, B.F., Sivakumar, S., Lim, K.H., Wong, W.K. and Juwono, F.H. (2023) Diabetes Detection Based on Machine Learning and Deep Learning Approaches. *Multimedia Tools and Applications*, **83**, 24153-24185. https://doi.org/10.1007/s11042-023-16407-5

[7] Shakeel, P.M., Baskar, S., Dhulipala, V.R.S. and Jaber, M.M. (2018) Cloud Based Framework for Diagnosis of Diabetes Mellitus Using K-Means Clustering. *Health Information Science and Systems*, **6**, Article No. 16. https://doi.org/10.1007/s13755-018-0054-0

[8] Ramadhan, N.G., Adiwijaya and Romadhony, A. (2021) Preprocessing Handling to Enhance Detection of Type 2 Diabetes Mellitus Based on Random Forest. *International Journal of Advanced Computer Science and Applications*, **12**, 223-228. https://doi.org/10.14569/ijacsa.2021.0120726

[9] Saxena, S., Mohapatra, D., Padhee, S. and Sahoo, G.K. (2023) Machine Learning Algorithms for Diabetes Detection: A Comparative Evaluation of Performance of Algorithms. *Evolutionary Intelligence*, **16**, 587-603. https://doi.org/10.1007/s12065-021-00685-9

[10] Wang, J. and Su, X. (2011) An Improved K-Means Clustering Algorithm. 2011 *IEEE* 3*rd International Conference on Communication Software and Networks*, Xi'an, 27-29 May 2011, 44-46. https://doi.org/10.1109/iccsn.2011.6014384

[11] de Melo, P. and Davtyan, M. (2023) High Accuracy Classification of Populations with Breast Cancer: SVM Approach. *Cancer Research Journal*, **11**, 94-104. https://doi.org/10.11648/j.crj.20231103.13

[12] Mostafa, S.M. and Amano, H. (2021) An Adjustable Variant of Round Robin Algorithm Based on Clustering Technique. *Computers, Materials & Continua*, **66**, 3253-3270. https://doi.org/10.32604/cmc.2021.014675

[13] Sen, P.C., Hajra, M. and Ghosh, M. (2020) Supervised Classification Algorithms in Machine Learning: A Survey and Review. In: Mandal, J. and Bhattacharya, D., Eds., *Advances in Intelligent Systems and Computing*, Springer, 99-111. https://doi.org/10.1007/978-981-13-7403-6_11

[14] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L. and Lopez, A. (2020) A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends. *Neurocomputing*, **408**, 189-215. https://doi.org/10.1016/j.neucom.2019.10.118

[15] Baliarsingh, S.K., Vipsita, S., Gandomi, A.H., *et al.* (2020) Analysis of High-Dimensional Genomic Data Using MapReduce Based Probabilistic Neural Network. *Computer Methods and Programs in Biomedicine*, **195**, Article ID: 105625. https://doi.org/10.1016/j.cmpb.2020.105625

[16] Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S. and Herrera, F. (2020) Big Data Preprocessing: Enabling Smart Data. Springer. https://doi.org/10.1007/978-3-030-39105-8

[17] Ingle, D.R., Waghmare, S.R., Patil, V. and Chavan, S. (2022) Identification of Vector Borne Disease Spread Using Big Data Analysis. 2022 *International Conference on Smart Generation Computing, Communication and Networking* (*SMART GENCON*), Bangalore, 23-25 December 2022, 1-8. https://doi.org/10.1109/smartgencon56628.2022.10083613

[18] Nilashi, M., Bin Ibrahim, O., Mardani, A., Ahani, A. and Jusoh, A. (2018) A Soft Computing Approach for Diabetes Disease Classification. *Health Informatics Journal*, **24**, 379-393. https://doi.org/10.1177/1460458216675500

[19] De Melo, P. (2024) Public Health Informatics and Technology. American Association for the Advancement of Science.

[20] Ahamad, M.K. and Bharti, A.K. (2021) Prevention from COVID-19 in India: Fuzzy Logic Approach. 2021 *International Conference on Advance Computing and Innovative Technologies in Engineering* (*ICACITE*), Greater Noida, 4-5 March 2021, 421-426. https://doi.org/10.1109/icacite51222.2021.9404575

[21] Nguyen, H.P. and Kreinovich, V. (2001) Fuzzy Logic and Its Applications in Medicine. *International Journal of Medical Informatics*, **62**, 165-173. https://doi.org/10.1016/s1386-5056(01)00160-5

[22] de Melo, P. (2025) Prediction of Diabetes from Electronic Health Records. *International Journal of Artificial Intelligence & Applications*, **16**, 21-37. https://doi.org/10.5121/ijaia.2025.16402