ADBA Computer Science
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8544
RESEARCH ARTICLE
Vol.1 / No.1 / 2024 / pp.26-35
https://doi.org/10.69882/adba.cs.2024075

Optimizing Diabetes Prediction: Addressing Data Imbalance with Machine Learning Algorithms

Khalid Hani Abushahla[*,1] **and Muhammed Ali Pala**[α,2]
[*]Biomedical Engineering Department, Graduate Education Institute, Sakarya University of Applied Sciences, 54050, Sakarya, Turkiye, [α]Biomedical Technologies Application and Research Center & Electrical and Electronics Engineering, Faculty of Technology, Sakarya University of Applied Sciences, 54050, Sakarya, Turkiye.

ABSTRACT

Imbalanced datasets pose significant challenges in various fields including the classification of medical conditions such as diabetes. This study investigates six methodologies for handling imbalanced diabetes datasets aiming to enhance classification performance through diverse preprocessing techniques. The methodologies are evaluated using multiple models: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, SVM, KNN, Naive Bayes, XGBoost, LightGBM, and CatBoost. The preprocessing techniques include simple implementation, data standardization, normalization, standardization with K-Fold cross-validation, and two variations incorporating the SMOTE oversampling technique. The effectiveness of each methodology is assessed based on accuracy, precision, recall, and F1 scores across different classifiers. Results indicate that standardization combined with K-Fold cross-validation consistently enhances model performance. Additionally, the integration of the SMOTE technique significantly improves results, especially for Gradient Boosting and SVM classifiers. Among the tested models, CatBoost demonstrated exceptional performance in handling imbalanced datasets, achieving an accuracy of 95.18%, precision of 91.10%, recall of 95.52%, and an F1 score of 93.26%. This study underscores the importance of tailored preprocessing techniques in improving the classification of imbalanced medical datasets, highlighting their potential to enhance predictive accuracy in critical applications.

KEYWORDS
Machine learning
Imbalanced dataset
Diabetes classification
Ensemble learning

INTRODUCTION

Diabetes is a long-term metabolic disease characterized by hyperglycemia (elevated blood glucose levels) due to deficiencies in the function of insulin secretion or both that damages the heart, blood vessels, eyes, kidneys, nerves, and heart over time (Association 2009). The most common kind of diabetes, Type 2, usually appears in adulthood as a result of either insufficient or resistant insulin production. Over the past 30 years, its prevalence has skyrocketed globally across all income categories. The hallmark of type 1 diabetes, also known as juvenile or insulin-dependent diabetes, is insufficient insulin production by the pancreas. For those with diabetes, having affordable access to treatment—in particular insulin—is essential. By 2025, the global goal is to stop the rise in diabetes and obesity. Approximately 422 million people worldwide suffer from diabetes, most of whom live in low- and middle-income countries. The disease is directly responsible for 1.5 million fatalities per year. Over the past few decades, there has been a steady increase in both the number of cases and the incidence of diabetes. Therefore, a tool that can help physicians identify this fatal disease earlier and halt its course is desperately needed (Abdulhadi and Al-Mousa 2021).

Machine learning techniques offer immense potential to enhance medical research and clinical care, particularly as providers increasingly utilize electronic health records. Two areas ready to benefit from the application of ML in the medical field are diagnosis and outcome prediction (Shivahare *et al.* 2024). This encompasses the potential identification of high-risk scenarios for

Manuscript received: 30 May 2024,
Revised: 27 June 2024,
Accepted: 30 June 2024.

[1] 22500305002@subu.edu.tr
[2] pala@subu.edu.tr

medical emergencies, such as relapse or transitioning into another disease state (Sidey-Gibbons and Sidey-Gibbons 2019). Recent successes include predicting the progression from pre-diabetes to type 2 diabetes using routinely-collected electronic health record data (Anderson *et al.* 2016).

Though in machine learning and AI, class imbalance in datasets is a common issue in real-world dataset analysis, particularly in industries like healthcare, finance, and telecommunications. It can lead to negative effects if incorrectly classified minority cases are identified. Two strategies have been developed in research: external techniques to rebalance distributions before training and internal algorithms to manage imbalance directly. The research aims to provide solid solutions for handling class imbalance in real-world data analysis situations (Ramyachitra and Manikandan 2014).

Several studies have explored predicting diabetes using various datasets and criteria resulting in varying accuracies and performance levels. Chang et al. (2022) evaluated interpretable machine learning models within the Internet of Medical Things (IoMT) using the Pima Indians diabetes dataset with random forest outperforming Naïve Bayes and J48 decision tree across multiple metrics (Chang *et al.* 2023). Naz & Ahuja (2020) focused on predicting diabetes onset achieving high accuracy rates with Deep Learning showing the highest accuracy (Naz and Ahuja 2020). Rajni & Amandeep (2019) introduced the RB-Bayes framework combining methods to improve prediction accuracy emphasizing early detection (Rajni and Amandeep 2019). Bhoi et al. (2021) employed multiple machine learning algorithms with Logistic Regression emerging as the top performer (Bhoi *et al.* 2021). Patra & Khuntia (2021) introduced the sdknn classifier showing significant improvement over conventional techniques (Patra and Khuntia 2021). Miao (2021) developed prediction models highlighting glucose, insulin, and BMI's correlations with diabetes and the Support Vector Classifier's potential (Miao 2021). Mousa et al. (2023) examined machine-learning models for diabetes diagnosis with LSTM performing best in capturing temporal dependencies (Mousa *et al.* 2023).

The prediction of diabetes using various machine learning models has been a topic of extensive research yielding diverse levels of accuracy and performance. Numerous algorithms have been used in studies ranging from interpretable models like Naive Bayes and random forest to deep learning strategies and ensemble frameworks that combine several techniques. Even though research shows notable improvements, the problem of class imbalance in diabetes datasets continues to be a major barrier to predictive accuracy. This study aims to address the problem of imbalanced diabetes datasets by investigating six methodologies to enhance classification performance through diverse preprocessing techniques. We evaluate the effectiveness of these methodologies using multiple models including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, SVM, KNN, Naive Bayes, XGBoost, LightGBM, and CatBoost. The preprocessing techniques explored include simple implementation, data standardization, normalization, standardization with K-Fold cross-validation, and two variations incorporating the SMOTE oversampling technique. The models' effectiveness is assessed based on accuracy, precision, recall, and F1 scores. This study underscores the importance of tailored preprocessing techniques in improving the classification of imbalanced medical datasets, highlighting their potential to enhance predictive accuracy in critical applications.

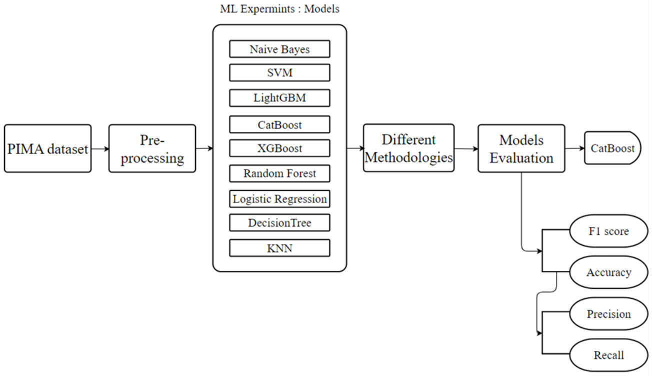

Figure 1 Graphical abstract of the proposed method

METHODOLOGY AND MATERIAL

In our study, we employ a comprehensive methodology to address the challenge of imbalanced datasets. Through detailed exploration, we delve into various preprocessing techniques, validation strategies, and data splitting methodologies to confront this issue head-on. Our approach involves precise experimentation to analyze the methods and differences between these approaches, revealing their strengths and weaknesses. By examining the results, we pointed out the most effective approach, one that not only mitigates data imbalance effects but also maximizes predictive performance. Our dedication to methodical exploration ensures optimal results and a deeper understanding of the underlying dynamics within our datasets.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial first step in the knowledge discovery process, where data scientists use a series of analysis operations (such as filtering, aggregating, and visualizing data) to interactively explore unknown datasets (Milo and Somech 2020). Before formal modeling, graphical representations, and visualizations, EDA seeks to conduct preliminary data investigations to find patterns, evaluate presumptions, and test hypotheses. Explanatory and comparative charts are a common feature of data visualizations, which help to clearly convey both concrete and abstract concepts. When issues are identified and resolved, the accuracy of diabetes diagnosis can be increased. Key aspects and hidden trends in the data can be summarized to help detect difficulties (**?**). Through Table 1, Figure 2, and Figure 3, some substantial insights of the PIMA dataset were illustrated with understandable visuals.

Understanding and Visualizing Data

This dataset includes data from studies on diabetes among women who identify as Pima Indian and who live in Phoenix, Arizona, USA. The women in the dataset are 21 years of age and older. It has eight different numeric variables and 768 entries (**?**). The selected target variable classes are labeled as follows: 1 denotes a positive diabetes test, while 0 denotes a negative test. The name of the dataset, descriptions of the data types, and corresponding roles are shown in Table 1. EDA clarified the dataset and showed that it included 268 individuals with diabetes and 500 values of sample patients who were not diabetic. This makes the imbalance in the dataset its primary challenge. For reference, refer to Figure 3.

Table 1 Dataset description

Main Criteria	Data Type	Input/Target	Notes
Outcome	Categorical	Target	0: No diabetes / 1: Diabetes
Pregnancies	Numerical	Input	-
Glucose	Numerical	Input	-
Blood Pressure	Numerical	Input	-
Skin Thickness	Numerical	Input	-
Insulin	Numerical	Input	-
BMI	Numerical	Input	-
Diabetes Pedigree Function	Numerical	Input	-
Age	Numerical	Input	-

Table 2 Statistical summary of the dataset

	Count	Mean	Std	Min	25%	50%	75%	Max
Pregnancies	768	3.845	3.37	0	1	3	6	17
Glucose	768	120.895	31.973	0	99	117	140.25	199
Blood Pressure	768	69.105	19.356	0	62	72	80	122
Skin Thickness	768	20.536	15.952	0	0	23	32	99
Insulin	768	79.799	115.244	0	0	30.5	127.25	846
BMI	768	31.993	7.884	0	27.3	32	36.6	67.1
Diabetes Pedigree Function	768	0.472	0.331	0.078	0.244	0.372	0.626	2.42
Age	768	33.241	11.76	21	24	29	41	81
Outcome	768	0.349	0.477	0	0	0	1	1

Pre-processing of the Data

Pre-processing is the primary prerequisite for working with datasets. Firstly, outliers in the dataset are addressed using Z-score calculation, where data points exceeding a specified threshold are replaced with NaN values. Subsequently, missing values and zero values in specific columns are handled. The only columns which are excluded from the zero values checking are the "Pregnancies" because it's a true value that many women haven't been pregnant before, and the "Outcome" column because 0 there demonstrates no diabetes diagnosis. Initially, missing values are identified and replaced with the mean of their respective columns. Then zero values in selected columns are replaced with the mean value as well. Then the data was split into 2 splits: training with 80% of the data and testing with 20%. Finally, a confirmation of the replacements is provided. Cumulatively, these processes guarantee that the dataset is free of outliers, NaNs, missing data, and zero values, making it appropriate for activities involving machine learning.

K-Fold Cross-Validation

To assess the generalization performance of the models, K-Fold cross-validation is employed with a predefined number of folds (k=5) (Murugan *et al.* 2023). This technique divides the dataset into k subsets, with each subset serving as a testing set and the remaining data as the training set. The procedure is repeated k times, and the average performance metrics across all folds are determined, assuring the models' durability and dependability beyond the original train-test split (Sohil *et al.* 2013).

Imbalanced Dataset and Solution

Classifiers are designed to categorize objects based on their attributes. However, in practical scenarios, datasets often exhibit class imbalance, where certain classes have significantly fewer instances compared to others. This class imbalance poses a challenge for traditional classification algorithms as they tend to be less accurate in predicting minority classes. This phenomenon

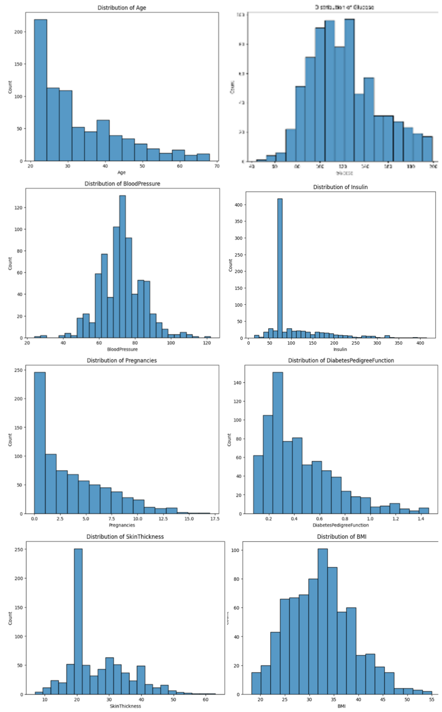

Figure 2 Distribution of the 6 input columns values [Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age]

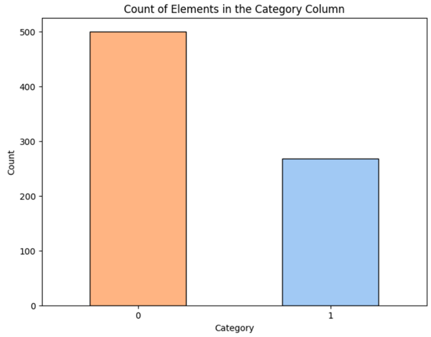

Figure 3 Outcome column values distribution [classification goal]

is known as the class imbalance problem. To address this issue, various techniques have been proposed, including over and under-sampling methods (Pradipta *et al.* 2021). Examples of that are SMOTE (Synthetic minority oversampling approach), ADASYN (Adaptive Synthetic Sampling Method), Borderline-SMOTE, and Safe-Level SMOTE (Gosain and Sardana 2017). These techniques aim to rebalance the dataset by adjusting the class distribution, thereby enhancing the performance of classification algorithms on imbalanced data (Pradipta *et al.* 2021). To address the issue of imbalanced dataset problem, the SMOTE technique was utilized.

SMOTE (Synthetic Minority Over-sampling Technique) introduced by Chawla et al. in 2002 addresses class imbalance by generating synthetic samples in the minority class rather than simply replicating existing samples, thus avoiding overfitting. To further enhance accuracy and mitigate overfitting, the SMOTE algorithm was refined. This method creates artificial minority instances along the line segments connecting minority samples and their 'k' nearest neighbors within the minority class. The 'k' nearest neighbors are randomly selected based on the desired oversampling rate. However, a limitation of SMOTE is its tendency to oversimplify the minority class space without considering the majority class, potentially leading to increased overlap between classes (Gosain and Sardana 2017).

Feature Scaling of the Dataset

MinMaxScaler and StandardScaler are both preprocessing techniques commonly used in machine learning (D.K. *et al.* 2019). MinMaxScaler rescales features to a predetermined range, typically between 0 and 1 (Powers 2020), whereas StandardScaler transforms features to have a mean of 0 and a standard deviation of 1. The primary distinction between them lies in their treatment of outliers and the resulting shape of the distribution.MinMaxScaler may distort data in the presence of outliers, whereas StandardScaler exhibits less sensitivity to them (de Amorim *et al.* 2023). In this study, both feature scalling tools were employed across various experiments to assess their impact on model training and the performance achieved when using normalized data.

Proposed Machine Learning Models

In this study, ten different machine learning models were employed to analyze the dataset and their outcomes were compared. While ensemble models have shown promising performance in prior research, this study explored and compared various ensemble models beside conventional ones in the domain of machine learning experimentation.

Logistic regression (LR) is a model that predicts the probability of a binary outcome by assessing the odds of the event occurring versus not occurring using predictor variables (y = 0 or 1). It employs the natural logarithm of these odds as a regression function. Odds ratios quantify the impact of predictors on the outcome, with the exponential of the regression coefficients providing these ratios. While logistic regression doesn't have a straightforward formula for estimation like linear regression, it involves iterative processes to converge on the best estimates (LaValley 2008).

A Decision Tree is generated from a collection of labeled training examples, each described by a set of attribute values paired with a class label. Given the expansive search options, decision-tree learning generally follows a greedy, top-down, and recursive approach commencing with the full training dataset and an unfilled tree. It selects an attribute that optimally divides the training data as the root split, subsequently segregating the data into distinct subsets based on the attribute's values. This process repeats recursively for each subset until all instances within a subset share the same class label (Su 2024).

Random Forest is created by combining different tree predictors so that every tree in the forest is dependent on the values of a random vector that is randomly sampled and has the same distribution for every tree. As the number of trees in a forest increases, the generalization error converges a.s. to a limit. The strength of each individual tree in the forest and the correlation between them determine the generalization error of a forest of tree classifiers. Each node can be split using a random feature selection process, which produces error rates that are more resilient to noise but still compare favorably to Adaboost (Rigatti 2017). Internal estimates track correlation, inaccuracy, and strength and are used to illustrate how the number of features employed in the splitting changes. The importance of each variable is also determined by internal estimations. These concepts also apply to regression (Breiman 2001).

Support Vector Machines (SVM) is a powerful algorithm based on Vapnik-Chervonenkis theory designed for supervised learning classification problems. It aims to find the optimal separating surface or hyperplane between two classes using kernel functions and slack variables for noisy data. SVM maximizes margins, the separation between the decision boundary and support vectors to maximize confidence in predictions and generalization ability, ensuring robustness and good generalization to new data (Bhavsar and Panchal 2024).

K-Nearest Neighbor (KNN) is an algorithm that is a straightforward yet effective machine learning method utilized for both classification and regression tasks. It operates by grouping data into coherent clusters or subsets and classifying new input based on its similarity to previously trained data. Essentially, the input is assigned to the class with the most nearest neighbors. While KNN is widely used due to its simplicity and effectiveness, it also possesses several weaknesses. To address these shortcomings, modified versions of the KNN algorithm have been developed through prior research efforts. These variants aim to enhance efficiency by mitigating the limitations of the original KNN approach (Taunk *et al.* 2019).

The basis of the Naive Bayes classifier is a probabilistic approach. Under the presumption that the existence of one feature in a class is unrelated to the existence of another feature in the same class, it applies Bayes' theorem. To estimate the probabilities of a particular category, one uses the joint probabilities of terms and categories. This independence assumption makes it possible to study each term's parameters separately, which speeds up calculation. A set of conditional probabilities and a structural model make up the Bayesian network (Kumari *et al.* 2021).

Gradient Boosting is a fundamental ensemble learning technique developed by Jerome H. Friedman in the late 1990s. It involves iteratively improving predictive models by training weak learners like decision trees to rectify errors from previous models. By focusing on the residuals or gradients of the loss function from the previous model, it reduces prediction errors and assembles an ensemble of models each refining the previous model's predictive accuracy.

XGBoost, an evolution of Gradient Boosting, was developed by Tianqi Chen in 2014 and quickly gained prominence for its efficiency and scalability. Building upon the principles of Gradient Boosting, XGBoost introduces advanced regularization techniques, parallel and distributed computing capabilities, and a comprehensive set of hyperparameters. By optimizing the model's architecture and training process, XGBoost significantly enhances performance while mitigating overfitting. Its versatility and robustness have made it a staple in data science competitions and real-world applications alike (Chen and Guestrin 2016).

LightGBM, a cutting-edge gradient boosting framework, emerged from the labs of Microsoft in 2016, engineered by Guolin Ke et al. Unlike traditional approaches, LightGBM employs novel tree-growing algorithms like Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to enhance speed and efficiency. By prioritizing the most informative data points during tree construction and leveraging histogram-based algorithms, LightGBM achieves unparalleled performance on large-scale datasets. With native support for categorical features and a rich set of hyperparameters, LightGBM empowers users to build high-quality models with minimal computational resources (Ke *et al.* 2016).

CatBoost developed by Yandex researchers in 2017, CatBoost revolutionizes gradient boosting with its intrinsic handling of categorical features. Created by Daniil Osokin and others, CatBoost automates categorical data encoding, sparing users from tedious preprocessing tasks. Employing advanced regularization techniques like ordered boosting and dynamic tree level regularization, CatBoost effectively combats overfitting while preserving predictive accuracy. Furthermore, its GPU-accelerated training and built-in visualization tools make it a formidable choice for practitioners seeking both performance and interpretability in their models (Prokhorenkova *et al.* 2019).

Evaluation Metrics To assess the models, we employed various metrics commonly used in machine learning evaluations such as the confusion matrix and its derived metrics: Accuracy, Precision, Recall, and F1 Score (Arias-Duart *et al.* 2023). Additionally, the ROC graph was displayed alongside the confusion matrix (Salih and Abdulazeez 2021).

RESULTS AND DISCUSSION

In the study, the methodology tackles the challenge of dealing with imbalanced datasets head-on. Various preprocessing techniques, validation strategies, and data splitting methodologies have been implemented to address this issue comprehensively. Through rigorous experimentation, we've meticulously examined the nuances and disparities between these methods, meticulously dissecting both their strengths and weaknesses.

By scrutinizing the results meticulously, we've identified the most effective approach, one that not only mitigates the effects of data imbalance but also maximizes predictive performance. Our dedication to methodical exploration ensures that we not only achieve optimal results but also gain a deeper understanding of

Table 3 Evaluation Metrics Definition, Formulas, and Ideal Values

Metric	**Formula**	**Definition**	**Ideal Situation**
Confusion Matrix	Table of [TP, TN, FP, FN]	An error matrix is another name for a confusion matrix. It facilitates our analysis of each categorization model's performance. It provides a clear picture of the efficiency of your classification method (Duvva 2024).	The ideal confusion matrix has values only along the diagonal (Duvva 2024).
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Proportion of correct predictions out of total predictions (Luque *et al.* 2019).	1.0
Precision	$\frac{TP}{TP+FP}$	Proportion of true positive predictions out of all positive predictions made by the model (Luque *et al.* 2019).	1.0
Recall	$\frac{TP}{TP+FN}$	Proportion of true positive predictions out of all actual (Luque *et al.* 2019).	1.0
F1 Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	Harmonic mean of precision and recall balances both metrics (Luque *et al.* 2019).	1.0
ROC	Graph of True Positive and False Positive rates (Raschka 2014).	Helpful resources for choosing categorization models according to how well they perform in terms of True Positive and False Positive rates. Random guessing is represented by the diagonal of a ROC graph and classification models that lie below the diagonal are thought to be less accurate than random guessing (Raschka 2014).	A perfect classifier would have a True Positive Rate of 1 and a False Positive Rate of 0 placing it in the upper left corner of the graph.

the underlying dynamics within our datasets.

The findings of the study present six distinct methodologies for handling our preserved dataset. In this section, the performance of various machine learning classifiers under different preprocessing techniques is presented and analyzed: Simple Implementation, Data Standardization, Data Normalization, Standardization With K-Fold, and Standardization With K-Fold and SMOTE. The study presents six distinct methodologies for handling a preserved dataset focusing on the performance of various machine learning classifiers under different preprocessing techniques. Simple implementation without standardization showed varied performance across classifiers with Support Vector Machine (SVM) exhibiting the highest accuracy.

Applying standardization to the data resulted in consistent improvements across classifiers, particularly benefiting Logistic Regression, Random Forest, and LightGBM. Normalization yielded mixed results with Logistic Regression showing the highest accuracy. Standardization with K-Fold validation provided robust estimates of model performance. Logistic Regression consistently emerged as the top performer across all preprocessing methods. Further exploration included Standardization with K-Fold and SMOTE, which notably improved model performance, particularly for Random Forest and SVM.

An alternative approach involving standardization, SMOTE oversampling, and K-Fold cross-validation showcased significant improvements in various classifiers, with CatBoost exhibiting impressive results. Comparing all methodologies revealed varying degrees of success, with the combined use of standardization, K-Fold cross-validation, and SMOTE proving effective. Random Forest and SVM consistently performed well across different methodologies. The findings offer valuable insights in addressing data preprocessing and class imbalance challenges in machine learning tasks, emphasizing the importance of careful experimentation and customization based on dataset characteristics and problem requirements.

Table 4 Evaluation metric results for each trained model across the six applied methodologies

Normalization of Data

Classifier	Acc	Prec	Rec	F1
Logistic Regression	77.27%	70.83%	57.62%	66.02%
Decision Tree	72.43%	61.21%	67.19%	64.05%
Random Forest	75.99%	66.57%	65.43%	66.00%
Gradient Boosting	75.32%	63.93%	70.91%	67.24%
SVM	72.76%	61.20%	67.45%	64.18%
KNN	71.68%	59.96%	71.24%	64.95%
Naive Bayes	74.68%	62.50%	72.73%	67.23%
XGBoost	71.43%	58.79%	77.27%	66.21%
LightGBM	72.08%	59.09%	70.91%	64.46%
CatBoost	74.03%	63.16%	65.45%	64.29%

Standardization With K Fold

Classifier	Acc	Prec	Rec	F1
Logistic Regression	72.44%	72.77%	57.62%	64.13%
Decision Tree	76.89%	69.35%	60.74%	64.80%
Random Forest	76.09%	68.91%	60.74%	64.13%
Gradient Boosting	76.30%	69.05%	61.16%	64.86%
SVM	72.21%	61.96%	68.60%	65.05%
KNN	72.16%	63.64%	71.24%	67.22%
Naive Bayes	75.13%	64.95%	63.64%	63.81%
XGBoost	73.18%	61.14%	68.75%	63.95%
LightGBM	74.38%	65.34%	65.91%	63.59%
CatBoost	76.43%	69.08%	59.87%	63.80%

Standardization, K Fold and SMOTE Implementation

Classifier	Acc	Prec	Rec	F1
Logistic Regression	75.66%	61.57%	70.91%	67.03%
Decision Tree	62.82%	54.29%	61.22%	57.03%
Random Forest	77.60%	66.44%	66.45%	65.69%
Gradient Boosting	74.87%	62.28%	72.16%	66.76%
SVM	76.69%	68.94%	71.43%	69.77%
KNN	71.95%	63.76%	71.24%	67.27%
Naive Bayes	74.38%	61.60%	75.00%	67.39%
XGBoost	74.50%	61.89%	68.45%	64.69%
LightGBM	75.49%	61.43%	68.30%	65.70%
CatBoost	76.43%	64.53%	73.30%	68.45%

Standardization, K FOLD and SMOTE Implementation (Variation)

Classifier	Acc	Prec	Rec	F1
Logistic Regression	77.34%	63.46%	74.63%	69.69%
Decision Tree	78.39%	63.95%	74.30%	74.30%
Random Forest	80.10%	100.00%	100.00%	100.00%
Gradient Boosting	89.19%	70.73%	90.67%	85.41%
SVM	82.53%	70.94%	74.70%	77.21%
KNN	100.00%	100.00%	100.00%	100.00%
Naive Bayes	74.74%	69.09%	70.91%	66.00%
XGBoost	100.00%	100.00%	100.00%	100.00%
LightGBM	100.00%	100.00%	100.00%	100.00%
CatBoost	95.18%	91.10%	95.52%	93.26%

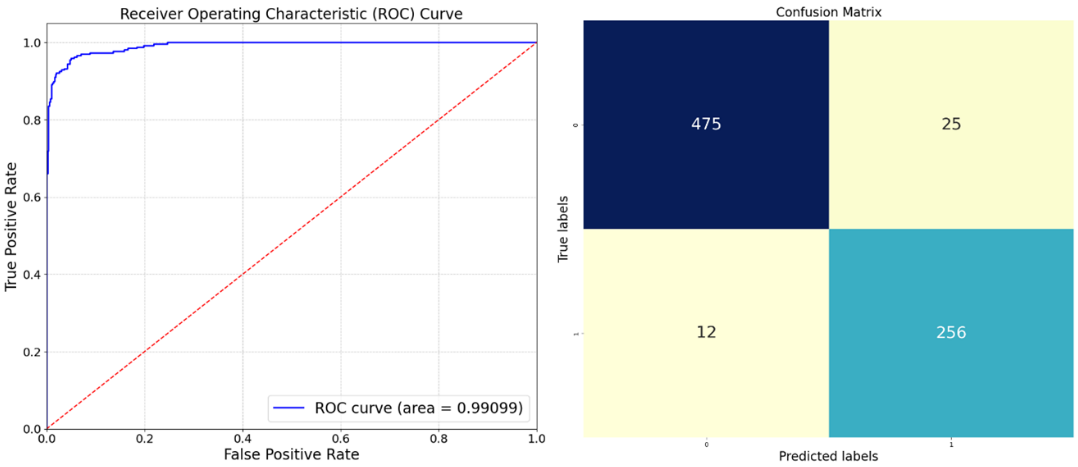

Figure 4 The confusion matrix and ROC curve illustrate the performance of the top-performing model (CatBoost) following the implementation of Standardization with K-Fold and SMOTE techniques.

Table 5 Classification performance of other classifiers in the literature

Reference	Models	Best Performance
(Chang et al., 2022) (Chang *et al.* 2023)	NB, RF, and J48 DT	F1-score: 85.17%
(Naz & Ahuja, 2020) (Naz and Ahuja 2020)	ANN, NB, DT, and DL	Accuracy: 98.07%
(Rajni & Amandeep, 2019) (Rajni and Amandeep 2019)	SVM, NB, KNN, and RB-Bayes framework	Accuracy: 72.9%
Bhoi et al. (2021) (Bhoi *et al.* 2021)	CT, SVM, k-NN, NB, RF, NN, AdaBoost (AB), LR	F1-score: 76%
Patra & Khuntia (2021) (Patra and Khuntia 2021)	Standard Deviation K Nearest Neighbor (SDKNN) classifier	Accuracy: 83.2%
(Miao, 2021) (Miao 2021)	SVM	Accuracy: 87.01%
(Mousa et al., 2023) (Mousa *et al.* 2023)	LSTM, RF, CNN	Accuracy: 85%
This study	LR, DT, RF, GB, SVM, KNN, NB, XGBoost, LightGBM, CatBoost	Accuracy: 94.27%, Precision: 89.16%, Recall: 95.15%, and F1 score: 92.06%

CONCLUSION

In conclusion, this article addressed the common challenge of imbalanced datasets, particularly in the context of classifying medical conditions such as diabetes. It investigated six distinct methodologies aimed at addressing the challenges posed by imbalanced datasets with a specific focus on classifying imbalanced diabetes datasets. The primary objective is to mitigate these challenges through customized preprocessing techniques. Through comprehensive evaluation using various classifiers and performance metrics such as accuracy, precision, recall, and F1 scores, it is evident that standardization, particularly when integrated with K-Fold cross-validation, consistently enhances model performance across classifiers. Moreover, the integration of the SMOTE oversampling technique significantly boosts model performance, particularly noted in Gradient Boosting and SVM classifiers.

Notably, CatBoost emerges as a proficient tool in handling imbalanced datasets, demonstrating impressive accuracy, precision, recall, and F1 scores adapted to the applied preprocessing techniques. These findings underscore the importance of customized preprocessing techniques in effectively addressing the challenges posed by imbalanced datasets, particularly in the context of diabetes classification, delving into the complexities of handling such datasets and highlighting the significance of employing appropriate preprocessing strategies to improve the classification of imbalanced medical datasets, thereby augmenting predictive accuracy in critical healthcare applications. Finally, among the models tested, CatBoost demonstrated exceptional performance in handling imbalanced datasets, achieving an accuracy of 95.18%, precision of 91.10%, recall of 95.52%, and an F1 score of 93.26%.

Availability of data and material

Not applicable.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

LITERATURE CITED

Abdulhadi, N. and A. Al-Mousa, 2021 Diabetes Detection Using Machine Learning Classification Methods. In *2021 International Conference on Information Technology (ICIT)*, pp. 350–354, IEEE.

Anderson, J. P., J. R. Parikh, D. K. Shenfeld, V. Ivanov, C. Marks, *et al.*, 2016 Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records. Journal of Diabetes Science and Technology **10**: 6–18.

Arias-Duart, A., E. Mariotti, D. Garcia-Gasulla, and J. M. Alonso-Moral, 2023 A Confusion Matrix for Evaluating Feature Attribution Methods. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3709–3714, IEEE.

Association, A. D., 2009 Diagnosis and Classification of Diabetes Mellitus. Diabetes Care **32**: S62–S67.

Bhavsar, H. and M. H. Panchal, 2024 A Review on Support Vector Machine for Data Classification Unpublished.

Bhoi, S. K., S. K. Panda, K. K. Jena, P. A. Abhisekh, S. Sahoo, *et al.*, 2021 Prediction of Diabetes in Females of Pima Indian Heritage: A Complete Supervised Learning Approach.

Breiman, L., 2001 Random Forests. Machine Learning **45**: 5–32.

Chang, V., J. Bailey, Q. A. Xu, and Z. Sun, 2023 Pima Indians Diabetes Mellitus Classification Based on Machine Learning (ML) Algorithms. Neural Computing & Applications **35**: 16157–16173.

Chen, T. and C. Guestrin, 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.

de Amorim, L. B. V., G. D. C. Cavalcanti, and R. M. O. Cruz, 2023 The choice of scaling technique matters for classification performance. Applied Soft Computing **133**: 109924.

D.K., T., P. B.G, and F. Xiong, 2019 Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. Pattern Recognition Letters **128**: 544–550.

Duvva, P., 2024 Did the Confusion Matrix Ever Confuse You? https://medium.com/wicds/did-the-confusion-matrix-ever-confuse-you-5fe869c10739, Accessed: March 9, 2024.

Gosain, A. and S. Sardana, 2017 Handling Class Imbalance Problem Using Oversampling Techniques: A Review. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 79–85, IEEE.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, *et al.*, 2016 LightGBM: A Highly Efficient Gradient Boosting Decision Tree Unpublished.

Kumari, S., D. Kumar, and M. Mittal, 2021 An Ensemble Approach for Classification and Prediction of Diabetes Mellitus Using Soft Voting Classifier. International Journal of Cognitive Computing in Engineering **2**: 40–46.

LaValley, M. P., 2008 Logistic Regression. Circulation **117**: 2395–2399.

Luque, A., A. Carrasco, A. Martín, and A. de Las Heras, 2019 The Impact of Class Imbalance in Classification Performance Metrics Based on the Binary Confusion Matrix. Pattern Recognition **91**: 216–231.

Miao, Y., 2021 Using Machine Learning Algorithms to Predict Diabetes Mellitus Based on PIMA Indians Diabetes Dataset. In *2021 the 5th International Conference on Virtual and Augmented Reality Simulations*, pp. 47–53, ACM.

Milo, T. and A. Somech, 2020 Automating Exploratory Data Analysis via Machine Learning: An Overview. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pp. 2617–2622, ACM.

Mousa, A., W. Mustafa, R. B. Marqas, and S. H. M. Mohammed, 2023 A Comparative Study of Diabetes Detection Using The Pima Indian Diabetes Database. University of Duhok Journal **26**: 277–288.

Murugan, S., P. K. Sivakumar, C. Kavitha, A. Harichandran, and W.-C. Lai, 2023 An Electro-Oculogram (EOG) Sensor's Ability to Detect Driver Hypovigilance Using Machine Learning. Sensors **23**.

Naz, H. and S. Ahuja, 2020 Deep Learning Approach for Diabetes Prediction Using PIMA Indian Dataset. Journal of Diabetes and Metabolic Disorders **19**: 391–403.

Patra, R. and B. Khuntia, 2021 Analysis and Prediction of Pima Indian Diabetes Dataset Using SDKNN Classifier Technique. IOP Conference Series: Materials Science and Engineering **1070**: 012059.

Powers, D. M. W., 2020 Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation .

Pradipta, G. A., R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, 2021 SMOTE for Handling Imbalanced Data Problem: A Review. In *2021 Sixth International Conference on Informatics and Computing (ICIC)*, pp. 1–8, IEEE.

Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, 2019 CatBoost: Unbiased Boosting with Categorical Features Unpublished.

Rajni and A. Amandeep, 2019 RB-Bayes Algorithm for the Prediction of Diabetic in Pima Indian Dataset. International Journal of Electrical and Computer Engineering **9**: 4866–4872.

Ramyachitra, D. D. and P. Manikandan, 2014 Imbalanced Dataset Classification and Solutions: A Review. International Journal of Computing and Business Research **5**.

Raschka, S., 2014 An Overview of General Performance Metrics of Binary Classifier Systems Unpublished.

Rigatti, S. J., 2017 Random Forest. Journal of Insurance Medicine **47**: 31–39.

Salih, A. A. and A. M. Abdulazeez, 2021 Evaluation of Classification Algorithms for Intrusion Detection System: A Review. Journal of Soft Computing and Data Mining **2**.

Shivahare, B. D., J. Singh, V. Ravi, R. R. Chandan, T. J. Alahmadi, *et al.*, 2024 Delving into Machine Learning's Influence on Disease Diagnosis and Prediction. The Open Public Health Journal **17**: e18749445297804.

Sidey-Gibbons, J. A. M. and C. J. Sidey-Gibbons, 2019 Machine Learning in Medicine: A Practical Introduction. BMC Medical Research Methodology **19**: 64.

Sohil, F., M. U. Sohali, and J. Shabbir, 2013 *An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)*, volume 6. Springer, 7th edition.

Su, J., 2024 A Fast Decision Tree Learning Algorithm Unpublished.

Taunk, K., S. De, S. Verma, and A. Swetapadma, 2019 A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pp. 1255–1260, IEEE.

How to cite this article: Abushahla, K. H., and Pala, M. A. Optimizing Diabetes Prediction: Addressing Data Imbalance with Machine Learning Algorithms. *ADBA Computer Science*, 1(1), 26-35, 2024.