Fırat Üniversitesi Müh. Bil. Dergisi
37(1), 303-320, 2025

Araştırma Makalesi
https://doi.org/10.35234/fumbd.1556260

The Impact of Balancing Techniques and Feature Selection on Machine Learning Models for Diabetes Detection

Vahid SİNAP[1*]
[1] Department of Management Information Systems, Faculty of Economics and Administrative Science, Ufuk University, Ankara, Türkiye
[*1] vahidsinap@gmail.com

(Geliş/Received: 25/09/2024; Kabul/Accepted: 24/01/2025)

Abstract: The detection of diabetes is crucial for effective management and prevention of the disease, which poses significant health risks globally. This study introduces a novel approach to diabetes detection by combining advanced data balancing techniques and feature selection methods, including Lasso (L1) regularization, to enhance the performance of predictive models in imbalanced datasets. Techniques such as Random Under Sampling (RUS), Adaptive Synthetic Sampling (ADASYN), and Synthetic Minority Over-sampling Technique (SMOTE) were employed alongside models including Random Forest (RF), CatBoost (CB), Extreme Gradient Boosting (XGB), K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), Logistic Regression (LR), and Gradient Boosting (GB) to assess their impact on model accuracy and generalization capabilities. The findings reveal that the RF model achieved the highest accuracy of 93.25% when utilizing the SMOTE technique, underscoring the importance of appropriate data handling strategies in improving predictive outcomes. Furthermore, when all features were utilized without selection, the RF model attained an accuracy of 95.31%, indicating the model's capacity to capture complex patterns when feature richness is maximized. The comprehensive methodology used in the study achieved a higher accuracy in diabetes detection than research in the literature and provided important outputs for developing reliable prediction models in healthcare.

Key words: Diabetes detection, data balancing techniques, imbalanced datasets, predictive modeling, health informatics.

Dengeleme Tekniklerinin ve Özellik Seçiminin Diyabet Tespitinde Makine Öğrenmesi Modelleri Üzerindeki Etkisi

Öz: Diyabet, küresel ölçekte önemli sağlık riskleri oluşturmaktadır. Diyabetin tespiti, hastalığın etkili yönetimi ve önlenmesi için büyük önem taşımaktadır. Bu çalışma, dengesiz veri setlerinde diyabet tespiti için çeşitli dengeleme tekniklerini ve Lasso (L1) düzenlemesi de dahil olmak üzere özellik seçim yöntemlerini birleştirerek diyabet tespitine yeni bir yaklaşım getirmektedir. Çalışmada, Random Under Sampling (RUS), Adaptive Synthetic Sampling (ADASYN) ve Synthetic Minority Over-sampling Technique (SMOTE) gibi teknikler, Random Forest (RF), CatBoost (CB), Extreme Gradient Boosting (XGB), K-En Yakın Komşu (KNN), Gaussian Naive Bayes (GNB), Lojistik Regresyon (LR) ve Gradient Boosting (GB) modelleri ile kullanılarak bu tekniklerin model doğruluğu ve genelleme yetenekleri üzerindeki etkileri değerlendirilmiştir. Bulgular, SMOTE tekniği kullanıldığında RF modelinin %93,25 ile en yüksek doğruluğa ulaştığını göstermektedir, bu da uygun veri işleme stratejilerinin tahmin sonuçlarını iyileştirmede önemini vurgulamaktadır. Ayrıca, özellik seçimi yapılmaksızın tüm özellikler kullanıldığında, RF modeli %95,31 doğruluk elde etmiş ve bu da özellik zenginliği maksimize edildiğinde modelin karmaşık desenleri yakalama kapasitesini ortaya koymaktadır. Araştırmada kullanılan kapsamlı metodoloji, diyabet tespitinde literatürdeki araştırmalardan yüksek bir doğruluğa ulaşmış ve sağlık hizmetlerinde güvenilir tahmin modelleri geliştirmek için önemli çıktılar sağlamıştır.

Anahtar kelimeler: Diyabet tespiti, veri dengeleme teknikleri, dengesiz veri setleri, tahmine dayalı modelleme, sağlık bilişimi.

1. Introduction

Diabetes mellitus (DM) is a chronic disease with an increasing prevalence, affecting millions of people worldwide. According to the World Health Organization, diabetes causes approximately 1.6 million deaths each year, reflecting the direct and indirect effects of the disease [1]. If diabetes is not controlled, serious complications such as heart disease, kidney failure, stroke and nerve damage can occur [2]. Diabetes is caused by insufficient insulin production (Type 1 diabetes) or the body's inability to use available insulin (Type 2 diabetes), leading to higher-than-normal blood glucose levels. Although Type 2 diabetes is a preventable disease, it continues to increase in prevalence, making early diagnosis and management even more important.

* Sorumlu yazar: vahidsinap@gmail.com. Yazarların ORCID Numarası: [1] 0000-0002-8734-9509

Despite advancements in diagnostic methods, gaps remain in accurately identifying diabetes at early stages, as traditional diagnostic methods often rely solely on static blood glucose measurements and clinical observations. This approach may not account for the dynamic nature of diabetes progression or provide insights for asymptomatic individuals, especially those in pre-diabetic stages. Addressing these gaps is crucial for enabling timely interventions and reducing diabetes-related complications.

Early diagnosis of diabetes enables individuals to be guided towards appropriate lifestyle changes necessary to slow the progression of the disease and prevent complications. Traditional methods of diagnosing diabetes today usually involve clinical examinations, laboratory tests and assessments based on the knowledge and experience of specialized physicians [3]. These methods include various laboratory tests that measure blood glucose levels, such as fasting blood glucose testing, oral glucose tolerance testing and hemoglobin A1c testing. However, such methods may not adequately reflect the dynamic and multidimensional nature of the disease, as they are based only on blood glucose levels measured at a specific moment in time. One of the biggest problems with traditional diagnostic methods is that they rely on limited data sources and the experience of physicians. The diagnostic process depends on the physician's assessment of the patient's symptoms, medical history and familial risk factors [4]. However, this method may ignore individual variations or nuances and is not accurate in the early stages of the disease or in asymptomatic individuals. For example, in patients who are pre-diabetic or in the early stages of diabetes, blood glucose levels are often borderline, which may be missed by conventional testing. In addition, traditional methods are largely time-consuming and costly. The collection of blood samples, laboratory analysis and interpretation of the results can prolong the diagnostic process for patients and increase the risk of diabetes progression [5]. Furthermore, in rural areas or where access to health services is limited, the lack of infrastructure and expertise to conduct laboratory tests is also a significant barrier.

Another important limitation of traditional methods is the increasing density and complexity of data. In healthcare today, it is possible to access large amounts of data from many different sources; however, processing and analyzing this data is very difficult and laborious with traditional methods. Analyzing the data can only scratch the surface, which can lead to overlooking complex underlying relationships or hidden patterns. Furthermore, results may not always be consistent as there may be subjective differences in assessment between experts [6]. These limitations highlight the need for more innovative and data-driven approaches to early diagnosis of diabetes. Advanced technologies such as machine learning and data mining can detect hidden patterns and meaningful relationships within large datasets, enabling higher accuracy and consistency in diagnostic processes. These approaches not only enable more effective use of existing health data, but also accelerate the diagnostic process, enabling individuals to access earlier intervention and treatment.

Machine learning algorithms are recognized as powerful tools for creating meaningful relationships and predictive models from large datasets. By learning relationships between variables from various data sources, these algorithms can analyze complex health data and identify previously undiscovered patterns [7]. Machine learning models developed for diabetes prediction make risk assessments using a wide variety of data types such as patients' genetic information, laboratory test results, lifestyle data and other health indicators. However, the lack of consistent results of these models in clinical applications is due to several important reasons. The success of machine learning models depends on the quality and representativeness of the datasets used [8]. Missing data, erroneous or misleading information can negatively affect the accuracy and generalization ability of the model. The heterogeneity and variability of health data can make it difficult for models to produce generalizable results [9]. In addition, existing models lack a standardized structure due to different health institutions and data collection methods [10]. This may cause the same model to perform differently on different datasets, leading to reliability issues in clinical applications. Furthermore, algorithms used for the comprehensive analysis and interpretation of health data often face performance issues due to challenges such as data imbalance. In datasets used to predict diseases like diabetes, there is frequently an imbalance between the classes; typically, there are more individuals who are healthy compared to those who are not. This imbalance can cause machine learning models to prioritize the majority class, making it difficult to accurately predict the minority class [11]. Thus, addressing data imbalance and ensuring model reliability across diverse datasets are critical for developing practical applications in healthcare. To address this, various methods and balancing strategies have been developed to enhance the accuracy of prediction models built from existing electronic health records, health screening results, and other large datasets. These approaches can lead to significant advances in the early detection and management of diabetes by enabling models to produce more generalizable and stability results. Additionally, using data from larger and more diverse patient populations allows for the development of more comprehensive and effective healthcare strategies and policies. As machine learning techniques evolve and data quality improves, these advancements hold the potential to revolutionize the early diagnosis and management of chronic diseases such as diabetes.

The main objective of this study is to address identified gaps in diabetes detection by combining data-balancing methods and advanced machine learning algorithms to create reliable predictive models, thereby improving the effectiveness of machine learning techniques in identifying the most appropriate methods for early diagnosis. Although existing research has demonstrated the effectiveness of machine learning techniques in diabetes prediction, data imbalance and feature selection have not been adequately addressed. This limits model accuracy and generalizability across different populations. In response, this study aims to reduce the effects of data imbalance -a key challenge in medical datasets that often leads to biased models- by applying advanced balancing techniques, namely SMOTE, RUS, and ADASYN. These methods enhance model reliability by either increasing the instances of the minority class or reducing the majority class instances, effectively countering the skewed distributions that commonly compromise the performance of traditional predictive models.

Furthermore, this study provides a detailed comparison using a comprehensive set of machine learning algorithms that utilize different statistical techniques such as RF, CB, XGB, KNN, GNB, LR and GB to evaluate and improve the prediction accuracy. This multi-algorithm approach not only considers the complexity and diversity of diabetes-related data but also tests various machine learning models systematically against balanced data, providing a nuanced understanding of each model's strengths and limitations. This study's diverse algorithmic framework enables a more holistic comparison, ultimately guiding practitioners towards the most suitable and reliable models for early diabetes diagnosis.

This research distinguishes itself by testing the integrated impact of data balancing and feature selection, specifically through Lasso regularization, to create high-performance models with the most relevant predictors. By addressing both data imbalance and feature relevance simultaneously, this study contributes a methodology that improves upon prior efforts and highlights the importance of balanced and optimized models in clinical applications. Comprehensive analysis of performance metrics, including accuracy and AUC, substantiates the study's effectiveness and provides essential insights for developing scalable, high-accuracy models that can be reliably applied across diverse clinical settings.

2. Related Works

In the field of diabetes detection, several significant studies have been conducted in the literature, highlighting various approaches and methodologies for improving predictive accuracy. Shin et al. [12] conducted a study aimed at creating DM prediction models using easily accessible health screening parameters. Two variable sets were employed to develop eight models, utilizing XGB and RF algorithms, with internal validation through stratified 10-fold cross-validation. The study found that the model based on 62 variables achieved the highest ROC-AUC of 0.928, while a simplified model with 27 variables still demonstrated acceptable performance, with ROC-AUCs ranging from 0.842 to 0.880. The inclusion of fasting glucose notably improved accuracy by up to 11.5%. Mir and Dhage [13] conducted a study aiming to enhance diabetes prediction using machine learning techniques. The researchers employed various classifiers, including Naive Bayes (NB), Support Vector Machine (SVM), RF, and Simple CART, utilizing the WEKA tool for analysis. The results indicated that SVM achieved the highest accuracy in predicting diabetes, suggesting it as the most effective algorithm among those tested. Sisodia and Sisodia [14] aimed to develop a model for predicting diabetes with high accuracy, addressing the challenges associated with traditional identification methods. The study utilized three machine learning classification algorithms: Decision Tree (DT), SVM, and NB, applying them to the Pima Indians Diabetes Database sourced from the UCI Machine Learning Repository. The results revealed that NB achieved the highest accuracy at 76.30%, demonstrating its effectiveness for early diabetes detection. Yahyaoui et al. [15] developed a Decision Support System (DSS) for diabetes prediction, addressing the growing prevalence of diabetes and its severe health implications. The study compared conventional machine learning techniques, SVM and RF, with a deep learning approach using a Convolutional Neural Network (CNN). Evaluating the proposed system on the Pima Indians Diabetes Database, which included 768 samples, the results demonstrated that RF achieved the highest accuracy at 83.67%, outperforming both CNN and SVM, which yielded accuracies of 76.81% and 65.38%, respectively. Khanam and Foo [16] used the Pima Indian Diabetes dataset from the UCI Machine Learning Repository to test seven machine learning algorithms for diabetes prediction. LR and SVM demonstrated strong performance, while a Neural Network model with two hidden layers achieved an accuracy of 88.6%. Sivaranjani et al. [17] employed SVM and RF algorithms to predict the likelihood of diabetes-related diseases. After data preprocessing and feature selection using step forward and backward methods, Principal Component Analysis (PCA) was applied. The RF model achieved an accuracy of 83%, outperforming the SVM, which had an accuracy of 81.4%. Hasan et al. [18] addressed outlier rejection, missing value handling, and feature selection using various machine learning classifiers, including RF and XGB. An ensemble model was created, weighting classifiers by their Area Under the

ROC Curve (AUC). The proposed model, tested on the Pima Indian Diabetes Dataset, achieved an AUC of 0.950, outperforming existing methods by 2%. Maniruzzaman et al. [19] focused on predicting diabetes using a machine learning-based system. Logistic regression was utilized to identify key risk factors, revealing seven significant variables. Four classifiers -NB, DT, Adaboost, and RF- were employed, with performance assessed using accuracy and AUC. The dataset, derived from the National Health and Nutrition Examination Survey, included 6,561 respondents. The study found that the combination of logistic regression for feature selection and RF classification achieved an accuracy of 94.25% and an AUC of 0.95, indicating strong predictive capability for diabetes.

Literature reveals a variety of approaches in diabetes detection, showcasing the effectiveness of machine learning techniques across multiple studies. Most existing studies, however, are limited in scope, often focusing on single algorithms or using imbalanced datasets without adequately addressing data skewness, which can compromise predictive reliability in clinical applications. These studies employed diverse algorithms such as RF, SVM, and NB to enhance predictive accuracy. Methods ranged from creating predictive models based on easily accessible health parameters to developing decision support systems that integrate both conventional and deep learning approaches. Many studies emphasized the importance of feature selection and data preprocessing, utilizing techniques like PCA and ensemble methods to improve performance metrics, including accuracy and AUC. However, issues such as data imbalance, lack of feature selection methods, and inconsistency in results across diverse datasets remain significant challenges in the field.

This study distinguishes itself from existing literature by integrating multiple advanced machine learning techniques and balancing methods to optimize diabetes prediction accuracy. Unlike previous works, this study adopts a holistic approach by systematically addressing data imbalance and irrelevant features, both of which are critical for creating more reliable and generalizable models. While previous works primarily focused on individual algorithms or specific datasets, this research employs a comprehensive approach that combines RF with sophisticated balancing techniques such as RUS, ADASYN, and SMOTE. By incorporating these balancing methods, the study effectively mitigates the risk of biased predictions toward the majority class, a limitation seen in prior research. Additionally, the utilization of Lasso regularization for feature selection ensures that only the most relevant variables are included, enhancing model strength. The empirical analysis demonstrates significant improvements in predictive performance, illustrating the importance of a multifaceted methodology in tackling the complexities of diabetes detection. This comprehensive strategy not only enhances predictive accuracy but also contributes to the development of more effective, scalable models that can be applied across diverse healthcare settings, marking a substantial advancement in the field.

3. Methodology

In this study, the objective is to improve the accuracy of diabetes detection through the evaluation of various data preprocessing techniques. These techniques will be assessed for their effectiveness in enhancing the quality and suitability of the input data for further analysis. Additionally, different oversampling methods will be examined to identify the most effective approach for mitigating class imbalance in the dataset. Class imbalance can significantly impact model performance, making it critical to select the appropriate oversampling method to enhance accuracy. Ultimately, the best combination of data preprocessing and oversampling techniques will be used to develop a model for precise diabetes detection. By integrating the most effective methods, this study aims to create a model that improves the reliability of diabetes diagnosis.

In the subsequent sections of the study, detailed information will be provided on several key aspects. The dataset utilized will be described, including its source and features. The classification algorithms employed, including GB, RF, CB, XGB, KNN, GNB, and LR, will be outlined, along with the data preprocessing techniques applied. Methods for addressing data imbalance, including oversampling techniques, will be discussed. Hyperparameter tuning processes will be explained, and Z-score normalization methods will be detailed. Cross-validation procedures and performance metrics used to evaluate the models will also be covered.

3.1. Dataset

The dataset used in this project is derived from the Behavioral Risk Factor Surveillance System (BRFSS) data spanning the years 1984 to 2015 [20]. Collected annually by the CDC, the BRFSS survey gathers data from over 400,000 Americans on health-related behaviors, chronic conditions, and preventive services. The rationale for selecting data only up to 2015 stems from several key considerations. Firstly, each year's dataset encompasses a substantial number of variables, ranging between 220 and 350, contingent upon the health topics prioritized during that specific survey year. The complexity of managing such a vast amount of data over several decades,

while striving to maintain a standardized format, presents considerable challenges. Notably, certain variables crucial for accurately predicting diabetes outcomes were not consistently available across all years. The inclusion of these variables, which number among the 22 key features employed in this study, is essential for ensuring the reliability of the predictive models. Additionally, the years following 2015, particularly those during the COVID-19 pandemic, introduced significant irregularities in health behaviors and access to care, which could skew analyses and predictions. The pandemic has been associated with alterations in chronic disease management that deviate from typical patterns, thereby necessitating careful consideration of any data integrated from this period. Thus, a focus was placed on a stable, pre-pandemic timeframe to ensure consistency and enhance the robustness of the findings.This study uses two distinct datasets to analyze and predict diabetes outcomes. The first dataset is a balanced collection of patient data, containing 70,692 samples with 22 features per sample. This dataset includes an equal number of individuals with and without diabetes, ensuring class balance and providing a fair basis for model evaluation. The second dataset is an imbalanced collection with 253,680 samples and the same 22 features per sample. In this dataset, a significant class imbalance is present, where one class is overrepresented compared to the other, which may impact model accuracy. The analysis will focus on exploring the relationships between features and the target variable in the imbalanced dataset to understand its structure and characteristics. Insights gained from this analysis will be used to enhance predictions in the balanced dataset, leveraging the extensive information from the larger, imbalanced dataset to improve the accuracy and reliability of diabetes predictions. Table 1 provides a detailed description of the features and their corresponding explanations.

Table 1. Description of features and their definitions.

Feature	**Description**
Diabetes_binary	Indicates whether the individual has diabetes (1: diabetic, 0: non-diabetic).
HighBP	Indicates the presence of high blood pressure (0: no high BP, 1: high BP).
HighChol	Indicates the presence of high cholesterol (0: no high cholesterol, 1: high cholesterol).
CholCheck	Indicates whether cholesterol levels have been checked in the past 5 years (0: not checked, 1: checked).
BMI	Body Mass Index, calculated as weight divided by the square of height.
Smoker	Indicates smoking status (0: never smoked, 1: smoked at least 100 cigarettes in entire life).
Stroke	Indicates whether the individual has had a stroke (1: had a stroke, 0: no stroke).
HeartDiseaseorAttack	Indicates a history of coronary heart disease or myocardial infarction (0: no, 1: yes).
PhysActivity	Indicates physical activity in the past 30 days, excluding job-related activity (0: no, 1: yes).
Fruits	Indicates whether the individual consumes fruit one or more times per day (0: no, 1: yes).
Veggies	Indicates whether the individual consumes vegetables one or more times per day (0: no, 1: yes).
HvyAlcoholConsump	Indicates heavy alcohol consumption (0: no, 1: yes), defined as more than 14 drinks per week for men and 7 for women.
AnyHealthcare	Indicates whether the individual has any form of health care coverage (0: no, 1: yes).
NoDocbcCost	Indicates if the individual could not see a doctor due to cost in the past 12 months (0: no, 1: yes).
GenHlth	Self-reported general health status on a scale from 1 (excellent) to 5 (poor).
MentHlth	Number of days in the past 30 days when the individual's mental health was not good (1-30 days).
PhysHlth	Number of days in the past 30 days when the individual's physical health was not good (1-30 days).
DiffWalk	Indicates difficulty walking (1: difficulty, 0: no difficulty).
Sex	Gender (1: male, 0: female).
Age	Age category, represented by 13 groups (e.g., 1: 18-24, 9: 60-64, 13: 80 or older).
Education	Education level on a scale from 1 to 6 (1: never attended or only kindergarten, 6: college graduate).
Income	Income level on a scale from 1 to 8 (1: less than $10,000, 5: less than $35,000, 8: $75,000 or more).

Figure 1 illustrates the correlations between various variables and diabetes, categorizing them into positive and negative correlations. Several variables exhibit a strong positive correlation with diabetes, including General Health (GenHlth), which shows the highest positive correlation, and High Blood Pressure (HighBP), which demonstrates a significant association with increased diabetes risk. Other variables, such as High Cholesterol (HighChol), Body Mass Index (BMI), and Heart Disease or Attack (HeartDiseaseorAttack), also show notable positive correlations. These strong correlations suggest that individuals with poorer general health, elevated blood pressure, high cholesterol, higher BMI, or a history of heart disease are at a higher risk of developing diabetes.

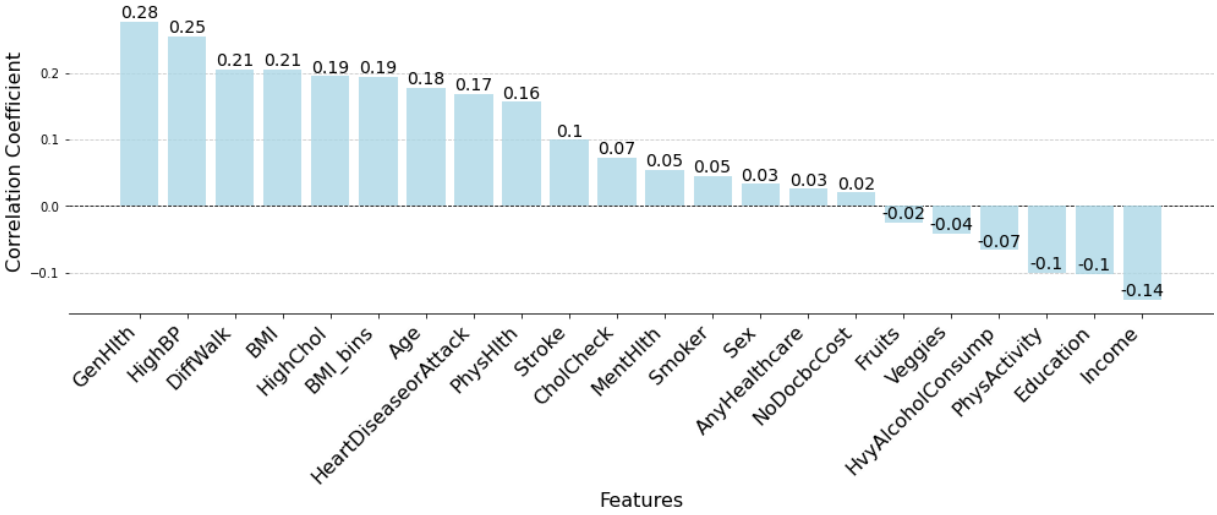

Figure 1. Correlation analysis of key variables with diabetes.

3.2. Data preprocessing

In the data preparation phase, the dataset underwent several preprocessing steps to ensure the data quality and suitability for subsequent analysis. Initially, the dataset was examined for missing values, and it was confirmed that there were no missing values. The next step involved checking duplicate entries, as duplicates can introduce biases and affect the model's performance. A total of 24,206 duplicate rows were detected and removed, resulting in a cleaned dataset with 229,474 entries. Following the cleanup, variables were grouped based on their characteristics. The target variable, Diabetes binary, indicates the presence or absence of diabetes. The remaining variables were classified into two groups: boolean variables and numerical variables. Boolean variables were identified as those with only two unique values, excluding the target variable, while numerical variables included all other features that did not fall into the boolean category. This categorization facilitated tailored preprocessing and analysis for each type of variable, optimizing the overall data preparation process.

3.3. Classification algorithms

In this study, various classification algorithms have been employed to predict diabetes outcomes effectively. Each algorithm was selected based on its suitability for handling the characteristics of the dataset and its performance in similar studies. The first algorithm utilized is GB, a powerful ensemble learning technique that builds models sequentially, where each new model attempts to correct the errors of the previous one. GB is particularly effective for datasets with imbalanced classes, as it focuses on learning from misclassified instances, thus enhancing predictive accuracy. The method combines weak learners, typically decision trees, into a strong learner, making it versatile for various classification tasks [21].

Another algorithm used in the research is RF. This algorithm operates by constructing multiple decision trees during training and outputting the mode of the classes for classification tasks. The inherent feature of RF to perform well with high-dimensional data and its robustness against overfitting makes it a valuable choice, especially in scenarios with substantial feature sets [22]. Furthermore, RF provides insights into feature importance, allowing for better understanding of the factors influencing diabetes prediction.

CB, which stands out for its ability to handle categorical features without extensive preprocessing, is another algorithm used in the research. CB implements an innovative approach to gradient boosting, reducing the risk of overfitting and improving generalization [23]. Its efficiency in managing large datasets while delivering competitive results in prediction accuracy is advantageous for this study's objectives.

XGB is recognized for its scalability and performance. XGB enhances the traditional gradient boosting framework by introducing parallel processing and regularization techniques, making it effective for handling complex datasets. Its capacity to optimize both training time and model performance through fine-tuning hyperparameters has proven beneficial in previous studies [24].

The KNN is a straightforward yet effective instance-based learning method. KNN classifies instances based on the majority class of their nearest neighbors in the feature space [25]. This algorithm is particularly useful for its simplicity and ease of interpretation, although it may require careful consideration of distance metrics and the number of neighbors to ensure optimal performance.

GNB has been included for its probabilistic approach to classification. Based on Bayes' theorem, GNB assumes independence among predictors, which simplifies the computation of posterior probabilities. Despite its simplifying assumptions, GNB has shown effective performance, especially in scenarios with high-dimensional data, making it a reliable choice for diabetes prediction tasks [26].

LR has been applied, serving as a fundamental statistical method for binary classification. LR models the probability of class membership using a logistic function, providing interpretability and efficiency [27]. Its application in healthcare-related predictive modeling is well-established, enabling insights into the relationships between features and the likelihood of diabetes occurrence.

3.4. Addressing data imbalance

Data imbalance is a common challenge in classification tasks, particularly when one class significantly outnumbers the other, as seen in the diabetes dataset used in this study, as shown in Figure 2. This imbalance can lead to biased models that perform well on the majority class but poorly on the minority class, ultimately compromising the model's predictive accuracy and generalizability. To address this issue, it is essential to apply techniques that balance the dataset, ensuring that both classes are adequately represented during training. Three methods, namely SMOTE, RUS and ADASYN, were used in this study.

SMOTE generates synthetic samples for the minority class by interpolating between existing minority instances, thus enhancing the representation of the minority class without duplicating data [28]. This approach helps to create a more balanced training set, improving the model's ability to learn the characteristics of both classes effectively [29]. RUS addresses imbalance by reducing the size of the majority class through random selection, creating a more balanced dataset by removing excess majority samples. While this method effectively balances the classes, it can potentially discard valuable information from the majority class, making it essential to carefully assess its impact on the model's performance [30]. ADASYN, an advanced variant of SMOTE, focuses on generating synthetic samples in areas where the minority class is underrepresented. This adaptive approach prioritizes instances that are harder to classify, thereby improving the model's capacity to differentiate between classes in complex decision boundaries [31]. By applying these techniques, the study aims to mitigate the negative effects of class imbalance, enhance the predictive power of the models, and ensure strong and reliable performance across both diabetic and non-diabetic classifications.

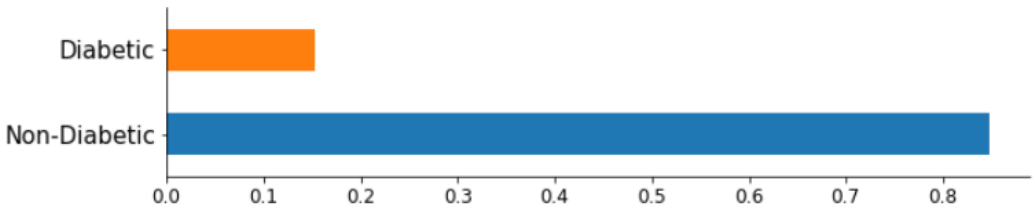

Figure 2. Distribution of diabetes status in the dataset (%).

3.5. Hyperparameter tuning

Hyperparameters are configuration settings used to control the training process of a model, unlike model parameters which are learned from the data. Optimal tuning of these hyperparameters can significantly enhance a model's ability to generalize well on unseen data, thereby improving its predictive accuracy. One of the most used methods for hyperparameter tuning is Grid Search. This technique involves exhaustively searching through a manually specified subset of the hyperparameter space to identify the best combination of parameters.

During Grid Search, the model is trained and evaluated for each combination of hyperparameters specified in the grid, usually using cross-validation, to ensure vigorous performance across different data splits. Although computationally intensive, Grid Search provides a thorough exploration of the hyperparameter space, making it a reliable method for finding optimal settings [32]. For this study, we employed Grid Search to determine the best hyperparameters for RF, CB, XGB, KNN, GNB, LR, and GB models. The optimal hyperparameter settings identified for each model, which were found to maximize performance metrics such as accuracy, and AUC, are shown in Table 2.

Table 2. Optimal hyperparameters for selected models determined via Grid Search.

Model	Hyperparameter	Optimal Value
RF	Number of Trees (n_estimators)	96
	Maximum Depth (max_depth)	11
	Minimum Samples Split (min_samples_split)	3
	Minimum Samples Leaf (min_samples_leaf)	2
	Max Features (max_features)	'sqrt'
CB	Depth (depth)	7
	Learning Rate (learning_rate)	0.09
	Number of Estimators (iterations)	483
XGB	Number of Estimators (n_estimators)	94
	Learning Rate (learning_rate)	0.08
	Maximum Depth (max_depth)	5
	Subsample (subsample)	0.76
	Colsample by Tree (colsample_bytree)	0.84
GB	Number of Estimators (n_estimators)	94
	Learning Rate (learning_rate)	0.08
	Maximum Depth (max_depth)	5
	Subsample (subsample)	0.76
	Colsample by Tree (colsample_bytree)	0.84
KNN	Number of Neighbors (n_neighbors)	5
	Weight Function (weights)	'uniform'
	Algorithm	'auto'
GNB	Smoothing Parameter (var_smoothing)	1e-09
LR	Regularization Strength (C)	1.0
	Solver	'lbfgs'
	Maximum Iterations (max_iter)	100

3.6. Feature selection

By identifying and retaining only the most relevant features, researchers can improve the accuracy of their predictive models while reducing computational complexity. Effective feature selection helps mitigate the risk of overfitting, enhances model generalization, and clarifies the relationships between the input features and the target variable [33]. In this study, Lasso was employed as a feature selection method to identify the most significant predictors of diabetes. Lasso regression applies a penalty to the absolute size of the coefficients, effectively shrinking some of them to zero, which aids in selecting a subset of features that contribute the most to the predictive capability of the model [34].

The strength of the relationship between each feature and the target variable, diabetes status, was assessed using correlation coefficients, categorized as follows:

- 0.5 and above: Very strong relationship – This feature has a significant impact on the target variable.
- 0.3 to 0.5: Strong relationship – This feature has a considerable effect on the target variable.
- to 0.3: Moderate relationship – This feature has a moderate effect on the target variable.
- 0.02 to 0.1: Weak relationship – This feature has a slight effect on the target variable.
- 0 to 0.02: Very weak or negligible relationship – This feature has very little or no effect on the target variable.

The relationship strengths of the features in relation to the target variable are categorized as follows: GenHlth (General Health) shows a very strong relationship with a score of 0.648. HighBP (High Blood Pressure) and BMI (Body Mass Index) demonstrate strong relationships, with scores of 0.532 and 0.406, respectively, along with Age at 0.395. HighChol (High Cholesterol) and DiffWalk exhibit moderate relationships, with scores of 0.298 and 0.267. HeartDiseaseorAttack (Heart Disease or Attack), PhysHlth (Physical Health), Income, and PhysActivity (Physical Activity) reflect weak relationships, with scores of 0.188, 0.172, 0.152, and 0.142, respectively. Finally, Education, HvyAlcoholConsump (Heavy Alcohol Consumption), Stroke, MentHlth (Mental Health), and Smoker (Smoker Status) are categorized as having very weak relationships, with scores of 0.109, 0.068, 0.078, 0.056, and 0.026, respectively.

3.7. Z-score normalization

Z-score normalization, also known as standardization, is a widely used technique in data preprocessing that transforms features to have a mean of zero and a standard deviation of one [35]. By normalizing the data, we enhance the performance of various machine learning algorithms that are sensitive to the scale of input features. The Z-score normalization formula is given in Equation 1.

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

Where, Z is the standardized value, x is the original value, μ is the mean of the feature, σ is the standard deviation of the feature. This transformation allows for comparison of scores across different features by placing them on a common scale. The resulting Z-scores indicate how many standard deviations a data point is from the mean, providing insights into the relative standing of each observation within its distribution.

3.8. Cross validation

In this study, 5-fold cross-validation was utilized as a method for evaluation of the predictive models. This technique systematically partitions the dataset into five equal-sized folds. For each iteration, one-fold is designated as the validation set while the remaining four folds are combined to form the training set [36]. This process is repeated five times, ensuring that each fold serves as the validation set exactly once. The use of 5-fold cross-validation helps to mitigate overfitting by providing a more generalized assessment of model performance across different subsets of the data. By averaging the evaluation metrics obtained from each fold, a comprehensive estimate of the model's predictive capability is achieved. The formula for calculating the performance metrics in 5-fold cross-validation is given in Equation 2.

$$\text{Performance Metric} = \frac{1}{k}\sum_{i=1}^{k} \text{Metric}_i \tag{2}$$

where k represents the number of folds, and Metric_i is the performance metric calculated for the i-th fold.

3.9. Performance metrics

In this study, the effectiveness of the predictive models was assessed using two primary performance metrics, which include Accuracy and AUC. Accuracy is a fundamental metric that indicates the proportion of correctly predicted instances among the total instances in the dataset [37]. It is shown in Equation 3.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

In this context, TP (True Positives) refers to the number of correctly predicted positive instances, while TN (True Negatives) indicates the number of correctly predicted negative instances. FP (False Positives) represents the number of incorrectly predicted positive instances, and FN (False Negatives) denotes the number of incorrectly predicted negative instances.

AUC, referring to the Receiver Operating Characteristic (ROC) curve, measures the model's ability to distinguish between classes [38]. AUC is calculated as shown in Equation 4.

$$\text{AUC} = \int_0^1 \text{TPR}\, d(\text{FPR}) \tag{4}$$

TPR (True Positive Rate) represents sensitivity, which is calculated as shown in Equation 5.

$$\text{TPR} = \frac{TP}{TP + FN} \tag{5}$$

FPR (False Positive Rate) is calculated using the formula provided in Equation 6.

$$\text{FPR} = \frac{FP}{FP + TN} \tag{6}$$

4. Experimental Study and Findings

In this section of the study, the focus is on addressing the challenges posed by imbalanced datasets in predictive modeling by evaluating the effectiveness of various balancing techniques and their influence on model performance. Imbalanced data poses significant challenges in predictive modeling, often leading to biased results favoring the majority class. To address this, the study explores several balancing methods, comparing their effectiveness in enhancing model performance. By systematically examining these techniques, the research aims to provide insights into the best practices for handling imbalanced data in predictive modeling contexts. The research was conducted in a structured manner, progressing through the following steps: First, analyses were performed on the original imbalanced dataset to establish baseline results. Subsequently, the dataset was balanced using RUS to assess the effect of reducing the majority class. The study then employed the ADASYN sampling technique to generate synthetic samples of the minority class, followed by the SMOTE to further balance the data. Finally, the impact of feature selection was evaluated on the dataset balanced specifically using the SMOTE technique, providing a comprehensive comparison of the methods.

All model development and analysis were conducted using Python, taking advantage of its extensive libraries for machine learning and data processing. The primary libraries included scikit-learn for implementing machine learning algorithms, imbalanced-learn for handling imbalanced data through techniques like SMOTE, Pandas and NumPy for data manipulation and preprocessing, and Matplotlib and Seaborn for visualizations. Code was developed and executed in the Jupyter Notebook environment, allowing for iterative testing and refinement of the model. The models were trained on a PC with the following specifications: a Ryzen 7800x3D processor running at 4.2 GHz, an NVIDIA 4070 Ti GPU for accelerated computations, 32 GB of DDR5 RAM at 6000 MHz, and the Windows 11 operating system. Figure 3 presents the workflow of the study for diabetes prediction.

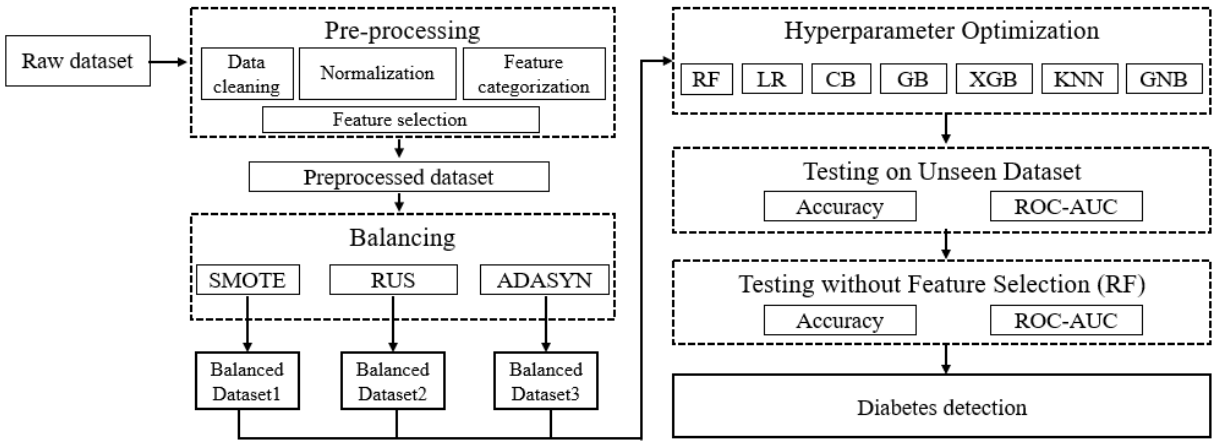

Figure 3. Workflow of the study for diabetes prediction model development.

Figure 3 presents the workflow of the study for diabetes prediction model development, illustrating the sequential steps including data preprocessing, application of classification algorithms, implementation of data balancing techniques such as SMOTE, RUS, and ADASYN, and evaluation of the RF model's performance without feature selection. The analyses conducted on the imbalanced dataset are presented in Table 3.

Table 3. Analysis results on the imbalanced dataset.

Model	Training Accuracy	Training AUC	Test Accuracy	Test AUC	Unseen Accuracy	Unseen AUC
RF	0.983416	0.998015	0.842514	0.764099	0.914520	0.984570
CB	0.864584	0.837797	0.859334	0.816036	0.587544	0.844531
XGB	0.861199	0.835693	0.858288	0.812885	0.581085	0.842366
KNN	0.874766	0.894319	0.835323	0.711257	0.653924	0.881901
GNB	0.774726	0.768106	0.777802	0.768734	0.667854	0.782006
LR	0.849264	0.803782	0.856502	0.807526	0.556091	0.816166
GB	0.853205	0.814605	0.859421	0.816742	0.563201	0.825097

Table 3 presents the analysis results for different models applied to the imbalanced dataset, focusing on key performance metrics, including testing and unseen data accuracy and AUC scores. The RF model stood out with the highest test accuracy of 0.8425 and a notable test AUC of 0.7641. Additionally, RF showed strong generalizability on unseen data, achieving an accuracy of 0.9145 and an AUC of 0.9846, which are the highest among all evaluated models.

Table 4. Analysis results on the RUS-balanced dataset.

Model	Training Accuracy	Training AUC	Test Accuracy	Test AUC	Unseen Accuracy	Unseen AUC
RF	0.950887	0.992859	0.366350	0.603088	0.631913	0.690164
CB	0.883607	0.939894	0.348658	0.547894	0.561971	0.575401
XGB	0.880251	0.936391	0.356850	0.559054	0.565707	0.586397
KNN	0.857530	0.944120	0.533075	0.633808	0.650868	0.689271
GNB	0.818123	0.905807	0.535515	0.623527	0.609337	0.646294
LR	0.845981	0.909715	0.490936	0.648997	0.626888	0.668883
GB	0.862950	0.924309	0.411801	0.624394	0.593857	0.645043

Table 4 displays the performance metrics of various models applied to the dataset balanced using the RUS technique. Among the models evaluated, the KNN model demonstrates relatively better performance, achieving a test accuracy of 0.5331 and a test AUC of 0.6338. KNN also shows competitive results on unseen data, with an

accuracy of 0.6509 and an AUC of 0.6893, which are among the highest in this configuration. Comparatively, the performance of models under RUS is generally lower than on the imbalanced dataset presented in Table 3. The RF, which previously excelled, shows a significant reduction in test accuracy (0.3664) and AUC (0.6031). This highlights the impact of balancing techniques like RUS, which, while mitigating the class imbalance, may also lead to information loss and reduced overall model performance.

Table 5. Analysis results on the ADASYN-balanced dataset.

Model	Training Accuracy	Training AUC	Test Accuracy	Test AUC	Unseen Accuracy	Unseen AUC
RF	0.984322	0.999205	0.687860	0.779964	0.851847	0.942934
CB	0.761683	0.846848	0.706728	0.813942	0.758895	0.836873
XGB	0.769956	0.854639	0.704593	0.810164	0.760039	0.838074
KNN	0.787682	0.875128	0.676225	0.749124	0.751321	0.820569
GNB	0.695081	0.769712	0.731001	0.767885	0.701464	0.781333
LR	0.732424	0.805529	0.720019	0.807786	0.741243	0.816781
GB	0.740081	0.817603	0.710868	0.816711	0.747513	0.824921

Table 5 presents the analysis results on the ADASYN-balanced dataset. Among the evaluated models, the RF model demonstrates superior performance, particularly evident in the test and unseen data. RF achieved a test accuracy of 0.6879 and an AUC of 0.7800. Furthermore, RF excelled on the unseen data with an accuracy of 0.8518 and an AUC of 0.9429, highlighting its strong generalization capability and effectiveness in predicting outcomes in previously unseen instances. In comparison to other models, the RF model consistently showed higher AUC values across different data splits, reinforcing its reliability in handling the ADASYN-balanced dataset. Other models, such as CB and XGB, also performed well with relatively high AUC scores (above 0.81 on the test set and around 0.84 on unseen data), but RF's superior unseen AUC of 0.9429 emphasizes its distinctive advantage in predictive performance.

Table 6. Analysis results on the SMOTE-balanced dataset.

Model	Training Accuracy	Training AUC	Test Accuracy	Test AUC	Unseen Accuracy	Unseen AUC
RF	0.985237	0.998741	0.827915	0.758419	0.932534	0.980670
CB	0.870121	0.947975	0.841119	0.800728	0.640775	0.829975
XGB	0.862297	0.940380	0.843211	0.803321	0.624817	0.813572
KNN	0.833034	0.869892	0.853364	0.661417	0.491912	0.675342
GNB	0.840268	0.920058	0.835062	0.793648	0.619213	0.800814
LR	0.856084	0.934584	0.836238	0.806860	0.637444	0.812978
GB	0.552343	0.610201	0.530542	0.609531	0.557897	0.615662

Table 6 shows the performance results on the SMOTE-balanced dataset. Among the evaluated models, RF stands out with consistently high performance across all metrics, particularly on unseen data. RF achieved a test accuracy of 0.8279 and a test AUC of 0.7584. Notably, RF excelled on unseen data with an accuracy of 0.9325 and an AUC of 0.9807. The overall findings underscore the RF model's strong adaptability and efficiency in handling imbalanced data, especially when balanced using SMOTE.

Table 7 presents the performance metrics of the RF model on the dataset without any feature selection, illustrating how the model's accuracy evolves as additional features are incrementally included. The RF model was selected for in-depth analysis across varying feature sets due to its consistently superior performance compared to other models in previous evaluations. The table reveals a clear trend: as the number of included features increases, there is a notable enhancement in model performance across all key metrics, particularly on unseen data. For example, with only a single feature, PhysHlth, the RF model achieved an unseen accuracy of 0.5926 and an unseen AUC of 0.6100. As more features were incorporated, such as PhysHlth, BMI, MentHlth, the unseen accuracy improved to 0.6639 and the unseen AUC to 0.7295. This progression continues, with the model reaching an unseen accuracy of 0.9532 and an unseen AUC of 0.9877 when 21 features were utilized. The data

strongly suggests that the inclusion of additional relevant features enables the RF model to capture more complex patterns in the dataset, thereby significantly enhancing its predictive accuracy and generalization capability. The most prominent gains are observed in the unseen metrics, emphasizing the RF model's improved ability to generalize to new data as feature richness increases.

Table 7. Performance of the RF model without feature selection.

Model	Test Accuracy	Test AUC	Unseen Accuracy	Unseen AUC	Features	No. of Features
RF	0.690953	0.605988	0.592568	0.609960	[PhysHlth]	1.0
RF	0.666768	0.695607	0.653518	0.712813	[PhysHlth, BMI]	2.0
RF	0.648379	0.661457	0.663930	0.729464	[PhysHlth, BMI, MentHlth]	3.0
RF	0.766385	0.724543	0.841030	0.928445	[PhysHlth, BMI, MentHlth, Age, HighBP, DiffWal...	9.0
RF	0.793054	0.733316	0.884501	0.958254	[PhysHlth, BMI, MentHlth, Age, HighBP, DiffWal...	10.0
RF	0.828700	0.757375	0.932114	0.980489	[PhysHlth, BMI, MentHlth, Age, HighBP, DiffWal...	15.0
RF	0.832534	0.766317	0.940788	0.983989	[PhysHlth, BMI, MentHlth, Age, HighBP, DiffWal...	16.0
RF	0.840988	0.777247	0.949998	0.987455	[PhysHlth, BMI, MentHlth, Age, HighBP, DiffWal...	20.0
RF	0.842383	0.777327	0.953170	0.987694	[PhysHlth, BMI, MentHlth, Age, HighBP, DiffWal...	21.0

The analysis reveals significant changes in model performance across various balancing methods and feature sets. The RF model demonstrated an accuracy of 91.45% and an AUC of 98.46% on the imbalanced dataset. However, when applying the RUS technique, the accuracy dropped to 36.63%, indicating a reduction of approximately 54.82%. In contrast, on the ADASYN-balanced dataset, RF's accuracy improved to 68.79%, reflecting a gain of about 32.16%. Further, the SMOTE balancing technique yielded an accuracy of 82.79%, resulting in a notable increase of around 14.00%. Finally, the analysis without feature selection showed that including more features consistently enhanced the model's accuracy, reaching 94.98% with 21 features, illustrating the importance of both balancing techniques and feature richness in predictive modeling.

Figure 4 illustrates the comparison of unseen accuracy values for various models across different datasets, including the imbalanced dataset, RUS-balanced dataset, ADASYN-balanced dataset, and SMOTE-balanced dataset. The bar chart presents the performance of various machine learning models, highlighting their unseen accuracy values across the different datasets. This comparison highlights the impact of different balancing techniques on model efficacy, with a clear indication of how unseen accuracy varies across models and datasets, thereby underscoring the importance of appropriate data handling strategies in enhancing predictive performance in diabetes detection.

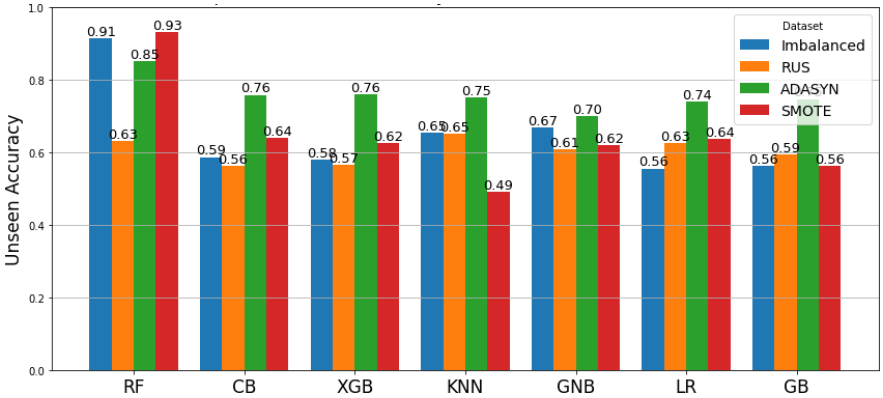

Figure 4. Comparison of unseen accuracy across different machine learning models and datasets.

5. Discussion

The present study investigates the impact of various balancing techniques and feature selection on predictive model performance in the context of imbalanced datasets, focusing on diabetes detection. The results highlight the challenges associated with class imbalance and the importance of employing appropriate methods to enhance model accuracy and generalization capabilities in predicting diabetes outcomes.

The RF model consistently emerged as the top performer across various datasets, demonstrating a test accuracy of 91.45% and an AUC of 98.46% on the original imbalanced dataset. This strong performance can be attributed to RF's ensemble learning approach, which effectively combines the predictions of multiple decision trees [39], thereby enhancing its ability to identify complex patterns in data relevant to diabetes detection. However, applying the RUS technique resulted in a drastic decrease in accuracy to 36.63%, underscoring the potential drawbacks of this method, particularly regarding information loss during data reduction [40]. RUS works by randomly removing instances from the majority class to balance the dataset, which can lead to the exclusion of valuable information that may be important for accurate predictions. This loss of data may reduce the model's ability to learn important features associated with diabetes, such as specific physiological indicators or lifestyle factors that are often present in the majority class. Moreover, the imbalance created by removing majority class instances can lead to a distorted representation of the underlying data distribution, making it challenging for the model to generalize effectively to unseen data. This aligns with findings from recent studies, which have highlighted that naive approaches to balancing can compromise model performance and accuracy, especially in high-stakes domains like healthcare where accurate predictions are critical [41].

The ADASYN and SMOTE techniques significantly enhanced the performance of the RF model, resulting in test accuracies of 68.79% and 82.79%, respectively. These improvements underscore the effectiveness of these advanced sampling methods in addressing class imbalance while maintaining critical data characteristics. The ADASYN technique, which focuses on generating synthetic minority instances based on the density of existing data points, allows for a more nuanced approach to balancing [42]. By adapting the number of generated instances to the local distribution of minority class samples, ADASYN effectively preserves the underlying structure of the data, thereby enhancing the model's ability to capture complex relationships relevant to diabetes detection. This adaptive nature is particularly beneficial in healthcare applications, where nuanced variations in data can carry significant predictive power. Similarly, the SMOTE technique employs a synthetic oversampling strategy, creating new minority class instances by interpolating between existing samples. This method not only increases the number of minority class instances but also enriches the dataset by introducing variability, which can lead to improved model strength [43]. By ensuring that the minority class is adequately represented, SMOTE enhances the RF model's capacity to generalize to unseen data, thus leading to improved accuracy and AUC metrics. These findings align with existing literature that emphasizes the efficacy of adaptive sampling methods in boosting predictive outcomes in imbalanced settings [44]. Such techniques facilitate a more balanced representation of data, enabling models to learn from a comprehensive view of the feature space associated with diabetes risk factors. This is fundamental, as accurate predictions are paramount in clinical settings, where misclassifications can have significant health implications.

The analysis conducted without feature selection demonstrated a noteworthy positive correlation between the number of features included in the model and its performance metrics. Specifically, the RF model achieved an accuracy of 95.31% when utilizing 21 features. As the number of relevant features increases, the RF model benefits from a richer representation of the underlying data landscape. Each additional feature contributes unique information that can help the model discern subtle variations in risk factors associated with diabetes. For instance, including features such as BMI, age, and blood pressure allows the model to leverage known risk factors that significantly impact diabetes prevalence [45]. This is critical in medical domains, where the complexity of patient data often requires multifactorial approaches to achieve reliable predictions. The findings align with existing literature that emphasizes the importance of feature richness in improving predictive accuracy and generalization capabilities [46]. Studies have shown that models with a higher number of relevant features are better equipped to identify complex interactions among variables, thus enhancing their overall interpretability and predictive power [47]. In the context of healthcare, this becomes vital, as accurately capturing the multifaceted nature of health-related data can lead to better risk assessments and more effective intervention strategies.

As demonstrated in Table 8, numerous studies have explored diverse machine learning models and datasets for diabetes detection, yielding varied levels of accuracy. For instance, the generalized regression neural network (GRNN) model by Zhang et al. [48] achieved a 75.49% accuracy on the CDC BRFSS2015 dataset, while Al-Absi et al. [49] attained a high 92% accuracy using the DiaNet v2 model on the Qatar Biobank dataset. Moreover, ensemble approaches, such as those used by Hasan et al. [18], and model combinations, like the LR and RF

ensemble applied by Maniruzzaman et al. [19], also showcased strong performance, reaching accuracies of 94.3% and 94.25%, respectively. However, the proposed method in this study, which utilized RF along with SMOTE on the CDC BRFSS2015 dataset, outperformed previous approaches with an accuracy of 95.31%. This result underscores the efficacy of RF in conjunction with SMOTE for managing imbalanced datasets and improving prediction accuracy in diabetes detection. The superior performance of our model may be attributed not only to the dataset choice and preprocessing steps but also to the optimized parameter tuning process applied. Specifically, Grid Search was employed to identify the best hyperparameters for the RF model, ensuring that it was finely tuned for optimal performance. Furthermore, cross-validation was implemented to validate the stability of the model, providing a reliable accuracy assessment across different data subsets.

The extensive data preprocessing steps in this study are among the other factors contributing to the higher accuracy compared to other studies in the field. The dataset's integrity was carefully examined, with missing values and duplicate entries eliminated to prevent the introduction of noise or bias. Additionally, variables were categorized based on their characteristics allowing for tailored preprocessing methods that optimized data handling and analysis for each type. Through this meticulous preparation, the model's ability to detect nuanced patterns was enhanced, contributing to the stability and reliability of the predictive framework and ultimately yielding superior accuracy in diabetes detection.

Table 8. Performance results of studies on diabetes detection.

Reference	Year	Method	Dataset Used	Accuracy
Zhang et al. [48]	2024	GRNN	CDC BRFSS2015 Dataset	GRNN: 75.49%
Al-Absi et al. [49]	2024	DiaNet v2	Qatar Biobank (QBB) Dataset	DiaNet v2: 92%
Maulana et al. [50]	2023	XGB	Pima Indians Diabetes Dataset	XGB: 82.68%
Khaleel and Al-Bakry [51]	2023	LR	Pima Indians Diabetes Dataset	LR: 94%
Shin et al. [12]	2022	XGB, RF	Health Promotion Center of Seoul St. Mary's Hospital	XGB: 85.8% (62 variables), 80.7% (27 variables)
Khanam and Foo [16]	2021	LR, SVM, CNN	Pima Indians Diabetes Dataset	CNN: 88.6%
Sivaranjani et al. [17]	2021	SVM, RF	Pima Indians Diabetes Dataset	RF: 83%
Hasan et al. [18]	2020	RF, XGB, Ensemble	Pima Indians Diabetes Dataset	XGB: 94.3%
Maniruzzaman et al. [19]	2020	LR, NB, DT, Adaboost, RF	National Health and Nutrition Examination Survey	LR + RF: 94.25%
Yahyaoui et al. [15]	2019	SVM, RF, CNN	Pima Indians Diabetes Dataset	RF: 83.67%
Mir and Dhage [13]	2018	NB, SVM, RF, Simple CART	Pima Indians Diabetes Dataset	SVM: 79.13%
Sisodia and Sisodia [14]	2018	DT, SVM, NB	Pima Indians Diabetes Dataset	NB: 76.30%
This study	-	RF + SMOTE	CDC BRFSS2015 Dataset	RF: 95.31%

6. Conclusion

This study underscores the critical role of advanced data-balancing techniques and feature selection in enhancing predictive model accuracy for diabetes detection, particularly in imbalanced datasets. Unlike prior research that often overlooks the impact of data imbalance on model accuracy, this work uniquely integrates multiple balancing methods with a comprehensive machine learning framework to address this issue systematically. The findings reveal that the RF model, combined with the SMOTE technique, achieved the highest predictive performance, highlighting SMOTE's effectiveness in mitigating class imbalance while maintaining essential data characteristics. In contrast, methods like RUS showed a significant drop in accuracy due to

information loss, emphasizing the potential drawbacks of certain balancing techniques and the need for careful method selection to avoid compromising data quality.

Another novel aspect of this study lies in its evaluation of feature selection through Lasso regularization, which demonstrated a clear positive correlation between the number of relevant features included and model performance. The findings reveal that as more relevant features are integrated, predictive accuracy improves, underscoring the importance of comprehensive feature inclusion in diabetes detection. Prioritizing relevant variables through this approach strengthens predictive capability and provides an efficient method for selecting features in medical diagnostics. This comprehensive integration of data balancing and feature selection represents a methodological advancement in developing high-performing and reliable models for healthcare applications. The insights from this study contribute valuable knowledge to the field by identifying optimal data handling strategies, particularly emphasizing the utility of SMOTE, to enhance predictive accuracy and support early diabetes diagnosis for improved health outcomes.

Looking ahead, future research could explore several avenues to build on the findings of this study. First, investigating the effectiveness of additional advanced balancing techniques, such as ensemble methods or generative adversarial networks, could provide further insights into improving model performance. Second, applying the developed methodologies to diverse populations and various healthcare settings would help validate the generalizability of the proposed models. Additionally, examining the interplay between feature selection methods and different algorithms could reveal more optimal combinations tailored for specific datasets. Finally, longitudinal studies could assess the real-world impact of these predictive models on early diabetes diagnosis and patient outcomes, thereby contributing to evidence-based practices in healthcare.

Acknowledgments

A preliminary version of this study was presented as a short abstract at the 1st International Data Analytics Congress, in 2024. The current manuscript is an extended version of that work, incorporating additional data, comprehensive analysis, and further discussions to provide a more in-depth exploration of the topic.

References

[1] World Health Organization. Diabetes. Available at: https://www.who.int/en/health-topics/noncommunicable-diseases/diabetes/#tab=tab_1 [Accessed 03 September 2024].
[2] Soumya D, Srilatha B. Late stage complications of diabetes and insulin resistance. J Diabetes Metab 2011; 2(9): 1000167.
[3] Sacks DB, Bruns DE, Goldstein DE, Maclaren NK, McDonald JM, Parrott M. Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. Clin Chem 2002; 48(3): 436-472.
[4] American Diabetes Association. Standards of medical care in diabetes—2019 abridged for primary care providers. Clin Diabetes 2019; 37(1): 11.
[5] Harris MI, Eastman RC. Early detection of undiagnosed diabetes mellitus: a US perspective. Diabetes Metab Res Rev 2000; 16(4): 230-236.
[6] Crow H, Gage H, Hampson S, Hart J, Kimber A, Storey L, Thomas H. Measurement of satisfaction with health care: implications for practice from a systematic review of the literature. Health Technol Assess 2002; 6(32): 1-10.
[7] Sinap V. A comparative study of loan approval prediction using machine learning methods. Gazi Univ J Sci Part C: Design Technol 2024; 12(2): 644-663.
[8] Gong Y, Liu G, Xue Y, Li R, Meng L. A survey on dataset quality in machine learning. Inform Software Technol 2023; 162: 107268.
[9] Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. J Big Data 2021; 8: 1-37.
[10] Neelima S, Govindaraj M, Subramani DK, ALkhayyat A, Mohan DC. Factors influencing data utilization and performance of health management information systems: a case study. Indian J Inform Sources Serv 2024; 14(2): 146-152.
[11] Thabtah F, Hammoud S, Kamalov F, Gonsalves A. Data imbalance in classification: Experimental evaluation. Inform Sci 2020; 513: 429-441.
[12] Shin J, Kim J, Lee C, Yoon JY, Kim S, Song S, Kim HS. Development of various diabetes prediction models using machine learning techniques. Diabetes Metab J 2022; 46(4): 650-657.
[13] Mir A, Dhage SN. Diabetes disease prediction using machine learning on big data of healthcare. In: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA); 16-18 August 2018; Pune, India. New York, NY, USA: IEEE; 2018. pp. 1-6.
[14] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Comput Sci 2018; 132: 1578-1585.
[15] Yahyaoui A, Jamil A, Rasheed J, Yesiltepe M. A decision support system for diabetes prediction using machine learning and deep learning techniques. In: 2019 1st International Informatics and Software Engineering Conference (UBMYK); 6-7 November 2019; Ankara, Turkey. New York, NY, USA: IEEE; 2019. pp. 1-4.

[16] Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. ICT Express 2021; 7(4): 432-439.
[17] Sivaranjani S, Ananya S, Aravinth J, Karthika R. Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS); 19-20 March 2021; Coimbatore, India. New York, NY, USA: IEEE; 2021. pp. 141-146.
[18] Hasan MK, Alam MA, Das D, Hossain E, Hasan M. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 2020; 8: 76516-76531.
[19] Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM. Classification and prediction of diabetes disease using machine learning paradigm. Health Inform Sci Syst 2020; 8: 1-14.
[20] Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System: Annual survey data. Available at: https://www.cdc.gov/brfss/annual_data/annual_data.htm [Accessed 07 September 2024].
[21] Tanha J, Abdi Y, Samadi N, Razzaghi N, Asadpour M. Boosting methods for multi-class imbalanced data classification: an experimental review. J Big Data 2020; 7: 1-47.
[22] Sinap V. Comparative analysis of machine learning techniques for detecting potability of water. J Sci Rep A 2024; 058: 135-161.
[23] Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data 2020; 7(1): 94.
[24] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug. p. 785-794.
[25] Wilson DR, Martinez TR. Reduction techniques for instance-based learning algorithms. Mach Learn 2000; 38: 257-286.
[26] Ontivero-Ortega M, Lage-Castellanos A, Valente G, Goebel R, Valdes-Sosa M. Fast Gaussian Naïve Bayes for searchlight classification analysis. NeuroImage 2017; 163: 471-479.
[27] Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002; 35(5-6): 352-359.
[28] Elreedy D, Atiya AF, Kamalov F. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Mach Learn 2024; 113(7): 4903-4923.
[29] Abacı İ, Yıldız K. SMOTE vs. KNNOR: An evaluation of oversampling techniques in machine learning. Gümüşhane University Journal of Science 2023; 13(3): 767-779.
[30] Susan S, Kumar A. The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent state of the art. Eng Rep 2021; 3(4): e12298.
[31] Mukherjee M, Khushi M. SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features. Appl Syst Innov 2021; 4(1): 18.
[32] Belete DM, Huchaiah MD. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. Int J Comput Appl 2022; 44(9): 875-886.
[33] Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. A review of feature selection methods on synthetic data. Knowl Inf Syst 2013; 34: 483-519.
[34] Li Z, Sillanpää MJ. Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theor Appl Genet 2012; 125: 419-435.
[35] Raju VG, Lakshmi KP, Jain VM, Kalidindi A, Padma V. Study the influence of normalization/transformation process on the accuracy of supervised classification. In: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT); 20-22 August 2020; Tirunelveli, India. New York, NY, USA: IEEE; 2020. pp. 729-735.
[36] Rodríguez-Pérez R, Fernández L, Marco S. Overoptimism in cross-validation when using partial least squares-discriminant analysis for omics data: A systematic study. Anal Bioanal Chem 2018; 410(23): 5981-5992.
[37] Dinga R, Penninx BW, Veltman DJ, Schmaal L, Marquand AF. Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines. BioRxiv. 2019; 743138.
[38] Obuchowski NA, Bullen JA. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. Physics in Medicine & Biology. 2018; 63(7): 07TR01.
[39] Belgiu M, Drăguţ L. Random forest in remote sensing: A review of applications and future directions. ISPRS J Photogramm Remote Sens 2016; 114: 24-31.
[40] Elhassan T, Aljurf M. Classification of imbalance data using Tomek link (T-link) combined with random under-sampling (RUS) as a data reduction method. Glob J Technol Optim 2016; S1:111.
[41] Zheng K, Cai S, Chua HR, Wang W, Ngiam KY, Ooi BC. Tracer: A framework for facilitating accurate and interpretable analytics for high stakes applications. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data; 14-19 June 2020; Portland, OR, USA. New York, NY, USA: ACM; 2020. pp. 1747-1763.
[42] Öter E, Doğan Y. A comparative study on data balancing methods for Alzheimer's disease classification. Cukurova Univ J Fac Eng 2024; 39(2): 489-501.
[43] Talukder MA, Islam MM, Uddin MA, Hasan KF, Sharmin S, Alyami SA, Moni MA. Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. J Big Data 2024; 11(1): 33.
[44] Khushi M, Shaukat K, Alam TM, Hameed IA, Uddin S, Luo S, Reyes MC. A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access 2021; 9: 109960-109975.

[45] Boutilier JJ, Chan TC, Ranjan M, Deo S. Risk stratification for early detection of diabetes and hypertension in resource-limited settings: Machine learning analysis. J Med Internet Res 2021; 23(1): e20123.
[46] Friedrich T, Schlauderer S, Overhage S. Some things are just better rich: How social commerce feature richness affects consumers' buying intention via social factors. Electron Mark 2021; 31: 159-180.
[47] Tredennick AT, Hooker G, Ellner SP, Adler PB. A practical guide to selecting models for exploration, inference, and prediction in ecology. Ecology 2021; 102(6): e03336.
[48] Zhang Z, Ahmed KA, Hasan MR, Gedeon T, Hossain MZ. A deep learning approach to diabetes diagnosis. In: Asian Conference on Intelligent Information and Database Systems; 10-12 April 2024; Singapore. Singapore: Springer Nature Singapore; pp. 87-99.
[49] Al-Absi HR, Pai A, Naeem U, Mohamed FK, Arya S, Sbeit RA, et al. DiaNet v2 deep learning-based method for diabetes diagnosis using retinal images. Sci Rep. 2024; 14(1): 1595.
[50] Maulana A, Faisal FR, Noviandy TR, Rizkia T, Idroes GM, Tallei TE, Idroes R. Machine learning approach for diabetes detection using fine-tuned XGBoost algorithm. Infol J Data Sci. 2023; 1(1): 1-7.
[51] Khaleel FA, Al-Bakry AM. Diagnosis of diabetes using machine learning algorithms. Mater Today Proc. 2023; 80: 3200-3203.