




SMOTE-ENN Resampling to Optimize Diabetes Prediction in Imbalanced Data

Agustinus Eko Setiawan^{1,2}, Supriadi Rustad^{2,3*}, Abdul Syukur², Moch Arief Soeleman²,
Muhamad Akrom^{2,3}, Andreas Wilson Setiawan⁴

¹ Faculty of Technology and Informatics, Aisyah University, Lampung 35372, Indonesia

² Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia

³ Research Center for Quantum Computing and Materials Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia

⁴ Faculty of Medicine, Universitas Dian Nuswantoro, Semarang 50131, Indonesia

Corresponding Author Email: srustad@dsn.dinus.ac.id

Copyright: ©2025 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/isi.300505>

ABSTRACT

Received: 23 February 2025

Revised: 18 May 2025

Accepted: 25 May 2025

Available online: 31 May 2025

Keywords:

SMOTE-ENN, diabetes mellitus, prediction, imbalanced data

This study reveals the role of resampling techniques in enhancing the performance of various ML models to detect Diabetes Mellitus (DM) in imbalanced datasets, namely BRFSS 2015 and PIMA Indian. For this purpose, five ML algorithms, KNN, RF, SVM, GB, and XGB, were implemented on both datasets before and after resampling using SMOTE, SMOTE-TOMEK, and SMOTE-ENN. The results uncovered that all resampling techniques improve the performance of all ML models. The best improvement was achieved by combining the KNN model and SMOTE-ENN technique, with an accuracy of 0.97 and other evaluation metrics of 0.96-0.99 for the BRFSS 2015. For the PIMA Indian, the combination performs perfectly with all evaluation metrics, with a value of 1.0. This study observed that resampling improves the correlation between each feature and the target, making it easier for the model to recognize data patterns. It was also found that in unbalanced datasets, the role of resampling is more worthy of attention than the choice of the algorithm. With the proper technique, namely SMOTE-ENN, the difference in performance between ML models was only a maximum of 2 to 4%. It makes the ML DM prediction more flexible when applied to the health sector.

1. INTRODUCTION

The increasing prevalence of chronic diseases poses a significant threat to global public health, primarily driven by rapid urbanization and lifestyle changes. Among these, Diabetes Mellitus (DM) has become one of the most widespread and challenging conditions affecting individuals across all age groups and socioeconomic backgrounds [1]. DM is generally classified into four main types: type 1 diabetes (T1D), also known as insulin-dependent or juvenile diabetes; type 2 diabetes (T2D), which is the most common form; gestational diabetes; and diabetes caused by specific medical conditions or genetic factors [2]. Both T1D and T2D can lead to serious and life-threatening complications if not appropriately managed, including diabetic foot syndrome, stroke, heart attack, liver cirrhosis, and chronic kidney failure [3]. Early and accurate detection of diabetes is therefore critical to prevent these complications and to reduce the burden on healthcare systems worldwide.

DM is a metabolic disorder that causes increased blood sugar levels [4]. The increased blood sugar levels can damage nerves, eyes, blood vessels, and other organs [5]. High blood sugar levels cause metabolic disorders in the body since it does not produce enough insulin, or the ineffective use of insulin that occurs in the body [6]. DM disease can be considered one

of the main challenges in the world health community because this disease is growing very rapidly [7]. According to statistical research, an estimated 463 million individuals globally had diabetes in 2019, and that number is expected to rise to 578 million in 2030 and 700 million in 2045. The number of diabetes patients is expected to increase by 25% in 2030 and by 51% in 2045 [8].

DM has been regarded as the seventh leading cause of premature death, and as such, about 1.6 million people die each year [9]. DM is dangerous for pregnant women and babies in the womb because there is a high potential for the mother to pass the disease on to her baby [6]. Given the ever-increasing risks associated with DM disease, early detection is a crucial issue, so accurate DM calcification and prediction are urgently needed for treatment and prevention efforts [10]. In computer science, diabetes classification and prediction are challenging tasks because the class distribution for all attributes is not linearly separable [11]. ML techniques are widely used in disease classification and prediction due to continuous technological advancements [12, 13]. With accurate ML algorithms for diabetes classification and prediction, individuals at risk and implementing early preventive interventions can be identified [14]. Some ML algorithms that have been widely employed for DM classification and prediction are KNN [15], RF [16], SVM

[17], and GB [18].

A common problem faced by ML in classification tasks is class imbalance [19]. This is known as imbalanced data when the number of samples in one class is much higher or lower than in other classes [20]. This can cause the ML model to be too biased towards the majority class, thus performing poorly in predicting the minority class [21]. Several resampling techniques have been developed to address the problem of class imbalance, including oversampling, undersampling, or a combination of both. Oversampling may result in overfitting of the model [22], while undersampling can eliminate essential parts of the majority class so that the decision boundary between classes is more challenging to learn, and it affects classification accuracy [23, 24]. By combining the advantages of oversampling and undersampling, hybrid sampling approaches enable researchers to concurrently decrease the number of majority classes and raise the number of minority classes [25].

Previous research [26] applied the SMOTE-ENN technique to overcome the class imbalance problem in the BRFSS Diabetes 2015 dataset in predicting DM. The SMOTE-ENN method combines SMOTE oversampling and ENN undersampling to reduce the number of samples in the majority class as a hybrid technique to achieve a balanced dataset [27]. SMOTE-ENN has proven more effective than the individual SMOTE method in building an accurate prediction model. The study employed several ML algorithms, namely KNN, RF, XG Boost, Bagging, and AdaBoost, and found that KNN was the best model with an accuracy of 0.98. The study produces a reliable prediction model for diagnosing diabetes mellitus based on accuracy, precision, recall, F1-score, and ROC/AUC values. Although the KNN model has the best accuracy, other models, RF, XGB, Bagging, and AdaBoost, have good performance, with accuracies of 0.96, 0.95, 0.93, and 0.94, respectively. Unfortunately, there is no explanation of the performance of the ML models on the dataset before resampling.

The study begins by revealing the difference in ML model performance on the BRFSS dataset before and after resampling using three techniques, namely SMOTE, SMOTE-TOMEK, and SMOTE-ENN, for various algorithms, namely KNN, RF, SVM, GB, and XGB [26]. The aim is to find out whether the resampling technique improves the performance of each algorithm and explain the cause of the increase in model performance due to resampling. The study is then applied to another imbalance PIMA Indian dataset from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) in the United States, and available at the UCI Respiratory ML [28]. In this dataset, 35% of patients are positive for diabetes, while 65% of patients are negative [29]. The study aimed to find the best resampling technique and ML model for the imbalanced PIMA Indian dataset. Following this introduction, section 2 discusses related previous research, section 3 presents details of the research methodology and dataset, and section 4 discusses the research results before being closed by the conclusion in section 5.

2. LITERATURE REVIEW

This section discusses previous studies relevant to using ML methods to predict DM. The application of ML methods for DM prediction has received increasing attention in recent years. Various studies have explored artificial intelligence

approaches to enhance early detection of DM by leveraging different datasets and model architectures [9, 30]. A key challenge frequently addressed is class imbalance in medical datasets, often mitigated through techniques such as SMOTE and its variants. For example, one study applied SMOTE to the BRFSS diabetes dataset to improve classification performance and compared several ML models, including Decision Tree (DT), LR, RF, KNN, and Gaussian Naive Bayes (GNB) [31]. Among these, the RF model achieved the highest accuracy (0.82%), closely followed by KNN (0.80%).

The effectiveness of enhanced resampling strategies was further demonstrated in another study, which utilized the SMOTE-ENN technique with the BRFSS diabetes dataset. This method provided better performance than SMOTE alone in developing an accurate prediction model. The study tested multiple algorithms KNN, RF, XGB, Bagging, and AdaBoost with KNN producing the highest accuracy at 0.98% and RF at 0.95% [26]. These results emphasize the potential of combining resampling and ensemble methods to build reliable tools for identifying risk factors and supporting early medical diagnosis.

Other research has focused on evaluating specific algorithm pairs. For instance, a comparative analysis between KNN and Naive Bayes (NB) revealed that NB outperformed KNN with average accuracies of 76.07% and 73.33%, respectively [15]. In another study, the performance of multiple ML algorithms was examined using the PIMA Indian Diabetes (PID) and German diabetes datasets. Algorithms such as NB, KNN, SVM, DT, RF, and LR were assessed using the WEKA 3.8.6 tool. On the PID dataset, LR recorded the highest accuracy at 74.0%, while KNN had the lowest at 66.1% [32]. Further findings from studies utilizing the SVM and GB algorithms reported accuracies of 84% and 76%, respectively [17, 18], reinforcing the role of model selection and dataset suitability in influencing predictive outcomes.

More comprehensive approaches have combined multiple datasets and advanced architectures. For example, one study constructed ML models using both the PIMA Indian and Iraqi LMCH diabetes datasets. It employed RF, SVM, and a two-growth deep neural network (2GDNN), achieving high performance metrics with precision, sensitivity, and F1-scores around 97%. On the PIMA dataset, accuracy reached 97.25% in testing and 99.01% in training, while the LMCH dataset yielded a comparable 97.33% accuracy [33].

Finally, the integration of both supervised and unsupervised learning algorithms has also been explored. In one study, classification models RF, SVM, NB, DT, and the k-means clustering method were applied to data from Frankfurt Hospital and the UCI PIMA Indian Diabetes Database. The RF algorithm delivered the highest accuracy 97.6% on the Frankfurt dataset, whereas SVM performed best on the PIDD dataset with an accuracy of 83.1% [34]. Validation was performed using training and testing splits, confirming the models' generalization performance.

Based on the literature review presented above, this study developed a reliable DM prediction model by utilizing multiple datasets and overcoming the class imbalance problem in the dataset. The problem of data imbalance can interfere with the learning process and reduce prediction accuracy, so it needs to be overcome through special techniques [35]. In addition, DM is a chronic disease that can cause various serious complications if not detected and treated early [36]. Therefore, an accurate prediction model is required to help diagnose diabetes more effectively. This study uses multiple

datasets to evaluate the accuracy of ML algorithms in predicting DM in various data conditions. This is crucial to test the robustness of the model [37] so that it can be appropriately implemented in various clinical contexts. The results of this study are also expected to provide valuable insights for the development of DM diagnosis and treatment applications in the future.

Adjusting the hyperparameter tuning on each algorithm, which can control the training process, can improve the accuracy of results [38, 39]. Aside from the use of hyperparameter tuning, the cross-validation method using K-Fold was utilized to obtain an optimal ML model [40].

3. METHODOLOGY

This study employed two diabetes datasets, the BRFSS and PIMA Diabetes datasets, with the research flow presented in Figure 1. Data preprocessing was carried out for both datasets, including data cleaning and imputation, outlier handling using the Interquartile Range (IQR) method, and data normalization to ensure all features were in the same range. Furthermore, class balancing was conducted using SMOTE, SMOTE-TOMEK, and SMOTE-ENN. The data was divided into training data (80%) and testing data (20%) for modeling purposes. The training was performed with internal validation using K-Fold cross-validation with 10 folds to ensure the resulting model was not overfitting. Several ML models were trained using this data: K-NN, RF, SVM, GB, and XGB. Grid Search Cross-Validation (Grid CV) was used for hyperparameter tweaking to determine the best parameters for every model. Using test data, the models were assessed in the

last step, and the best model was applied to forecast the consequences of diabetes.

3.1 Dataset descriptions

The first dataset used in this analysis is a subset of the 14,268 samples from the BRFSS Diabetes behavioural risk factor data, each including one output attribute and 20 input attributes. Blood pressure, cholesterol, smoking, diabetes, obesity, age, sex, race, food, exercise, alcohol usage, BMI, household income, marital status, sleep, time since last checkup, education, health insurance, and mental health were among these characteristics [41]. The detailed attribute information and descriptions are presented in Table 1.

The second dataset is PIMA Indian, containing 768 samples; each sample had 8 input attributes and 1 output attribute. These attributes encompassed the number of Pregnancies, plasma glucose levels, blood pressure, skin thickness, insulin levels, body mass index, the ability to analyze diabetes risk, age, and the classification results of whether someone is a diabetic.

Each attribute had a different range of values, such as the number of pregnancies, 0-17 times, glucose levels, 0-199 mg/dl, and blood pressure, 0-122 mm/Hg. These attributes provide a comprehensive understanding of the data used for diabetes prediction research so that they can help identify important factors that influence the diagnosis and prevention of diabetes. Specifically, all patients here were women aged at least 21 years with PIMA Indian heritage [42]. A detailed attribute information table for the PIMA Indian dataset is presented in <https://github.com/agustinus58/Diabetes-Mellitus-Prediction-on-Imbalance-Datasets>.

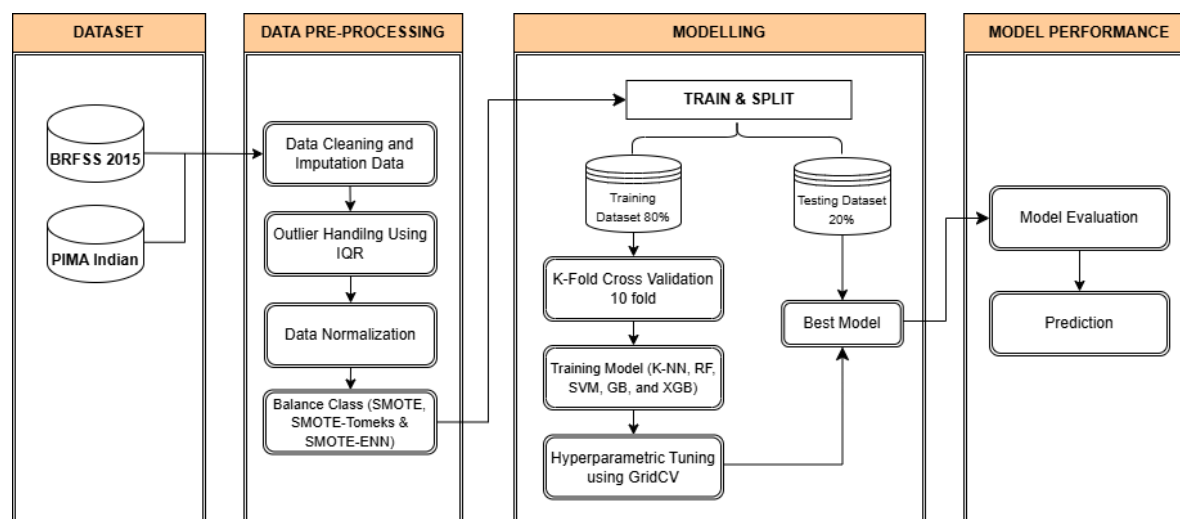


Figure 1. Proposed methodology for DM prediction assessment

Table 1. Attribute information of BRFSS diabetes

No	Variable Name	Description
0	Diabetes binary	0 = no diabetes 1 = prediabetes or diabetes
1	High BP	0 = no high BP 1 = high BP
2	High Chol	0 = no high cholesterol 1 = high cholesterol
3	Chol Check	0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
4	BMI	Body Mass Index
5	Smoker	In your lifetime, have you smoked at least 100 cigarettes? [Note: 100 smokes are in 5 packs.] 1 = yes, 0 = no
6	Stroke	You had a stroke, did you know that? 1 = yes, 0 = no
7	Heart Disease or Attack	Myocardial infarction (MI) or coronary heart disease (CHD) 1 = yes, 0 = no
8	Phys Activity	30 days' worth of physical exercise, excluding work 1 = yes, 0 = no

9	Fruits	Eat fruit at least once a day. 1 = yes, 0 = no
10	Veggies	Eat Vegetables at least once a day 1 = yes, 0 = no
11	Hvy Alcohol Consump	Heavy drinkers (adult women who consume more than seven drinks per week, and adult men who consume more than fourteen drinks per week) 1 = yes; 0 = no
12	Any Healthcare	possess health care coverage, such as health insurance, prepaid plans like HMOs, etc. 1 = yes; 0 = no.
13	No Docbc Cost	Have you ever needed to see a doctor in the last 12 months but been unable to do so due to financial constraints? 1 = yes, 0 = no
14	Gen Hlth	If you had to rate your overall health on a scale of 1 to 5, where 1 represents excellent, 2 very good, 3 good, 4 fair, and 5 poor, how would you rank it?
15	Ment Hlth	In light of your mental health, which includes emotional problems, stress, and depression, how many days throughout the past 30 days did you have poor mental health? Range from 1 to 30 days
16	Phys Hlth	Considering any illnesses or injuries, how many days during the past 30 days did you experience poor physical health? Please answer with a number between 1 and 30 days.
17	Diff Walk	Do you find it difficult to climb stairs or walk? 1 = yes, 0 = no
18	Sex	0 = female 1 = male
19	Age	13-level age group (codebook: _AGEG5YR) 1 = 18–24 9 = 60–64. 13 = 80 years of age or older
20	Education	The scale of education level (EDUCA see codebook) 1–6 1 = Only completed kindergarten or never went to school 2 = Elementary Grades 1 through 8 Grades 9–11 (some high school) = 3. 4 = GED (high school graduation) or Grade 12 5 = One to three years of college (technical school or college) 6 = 4 years or more of college (college graduate)
21	Income	Income scale (refer to the codebook for INVOE2): On a scale of 1–8, 1 is less than \$10,000, 5 is less than \$35,000, and 8 is at least \$75,000.

3.2 Data preprocessing

Data preprocessing involves various steps, such as handling missing values, removing duplicates, and managing outliers [43]. In data cleaning, imputation was performed to fill in missing data using the mean, median, or other appropriate approaches so that the dataset remains intact [44]. At the same time, outlier handling was carried out to reduce the impact of extreme values that can interfere with the analysis results, such as by using IQR (Interquartile Range) to remove inappropriate outliers [45]. Data normalization is the next step, converting data values to a uniform scale so that differences in scale between variables do not affect the results of the ML model. Finally, class balancing is a technique for balancing the distribution of classes in an unbalanced dataset, such as by oversampling or undersampling methods, so that the model is not biased towards the majority class and can more accurately predict the minority class. These three steps are handy for ensuring the quality of the data to build effective and reliable models.

After data cleaning and IQR application, the next step was data normalization using min-max scaling with the aim of standardizing the scale of features in the datasets [46]. This was done by calculating each feature's minimum and maximum values and then applying a linear transformation to map the original values into the new range [0, 1]. Eq. (1) governs the min-max scaling, where x'_i is the result of normalizing data x_i , and x_i is the i -th data to be normalized, while x_{min} and x_{max} are the maximum and minimum values of the dataset features.

$$x'_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Data resampling was done using SMOTE oversampling and its combination with Tomek Links (SMOTE-TOMEK) and ENN undersampling (SMOTE-ENN) methods to balance datasets. The goal was to increase the predictive power of the model by balancing the classes before the machine-learning model was built [47]. Following class balancing with SMOTE, SMOTE-TOMEK, and SMOTE-ENN, Table 2 shows the percentage of samples between the regular and diabetes classes.

Table 2. Number of samples after class balancing

Dataset	Imbalance Class	SMOTE	SMOTE-TOMEK	SMOTE-ENN
Brfss	11788:2321	11788:	11764:	11257:
Diabetes		11788	11764	6347
Pima	398:192	398:398	393:393	339:307
Indian				

3.3 Modelling

As depicted in Figure 1, ML modeling starts with splitting the dataset into training and test sets in a ratio of 80:20. The popular Pareto principle is the basis for the 80:20 split's rationale [48]. However, that is simply the thumb rule that practitioners employ. There don't appear to be any explicit guidelines on the ideal ratio for the supplied dataset [49]. While 20% of the data is used to test the model, the remaining 80% is used to train five ML models: KNN, RF, SVM, GB, and XGB.

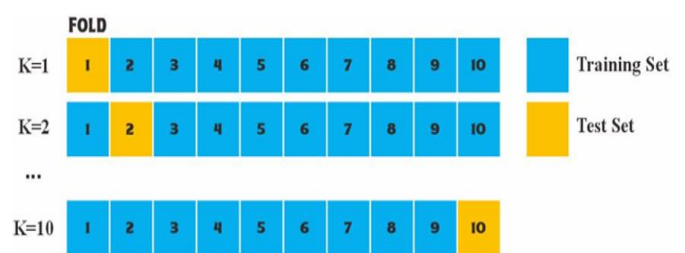


Figure 2. Illustration of K-fold cross-validation

The training was conducted by applying K-fold cross-validation with a value of $K = 10$, randomly dividing the training data into 10 folds, where nine folds were used to train the model. At the same time, the remaining 1-fold was employed to validate internally. The model was trained using the remaining folds in the first iteration, while the initial fold served as validation. The model was trained using the remaining folds in the second iteration, whereas the second fold was used for testing. As shown in Figure 2, this procedure was repeated until all 10 folds had been employed to validate the model.

The performance of a classifier algorithm is highly

dependent on the parameter values that define its model architecture, known as hyperparameters. To obtain the best results from a classifier algorithm on a dataset, hyperparameter tuning needs to be done by trying various values until the optimal value is found that produces the best performance [50]. This study performed the hyperparameter tuning process using the Grid search CV algorithm using 10-fold cross-

validation by testing various ranges of hyperparameter values shown in Table 3 using nested loops while building the classifier model. Table 4 presents the details of the hyperparameters with the best values on the PIMA Indian dataset, and other BRFS Diabetes 2015 datasets can be found in <https://github.com/agustinus58/Diabetes-Mellitus-Prediction-on-Imbalance-Datasets>.

Table 3. Classifiers and their hyperparameters of the prediction model

Algorithm	Parameter Name	Value
KNN	N neighbors	3,5,7,9
	weight	Uniform, distance
	Algorithm	Auto, ball tree, kd tree, brute
	Leaf size	10, 20, 30
SVM	p	1,2
	C	0.01, 0.1, 1, 10, 100
	Kernel	Linear, rbf
	Gamma	Scale, auto
Random Forest	Class weight	None, balanced
	N estimators	50, 100, 200
	Max depth	None, 10, 20, 30
	Min samples split	2, 5, 10
Gradient Boosting	Min samples leaf	1, 2, 4
	Max features	Sqrt, log2, None
	bootstrap	True, False
	N estimators	50, 100, 200
XG Boost	Learning rate	0.01, 0.1, 0.2
	Max depth	3, 4, 5, 6
	Min samples split	2, 5, 10
	Min samples leaf	1, 2, 4
	subsample	0.6, 0.8, 1.0
	Max features	Auto, sqrt, log2
	N estimators	100, 200, 300
	Learning rate	0.01, 0.05, 0.1
	Max depth	3, 4, 5
	subsample	0.8, 0.9, 1.0
	Colsample bytree	0.8, 0.9, 1.0

Table 4. Hyperparameter tuning training parameter

Dataset	Model	Parameter Name	Best Value			
			Imbalance Data	SMOTE	SMOTE-TOMEK	SMOTE-ENN
Pima Indian	KNN	N neighbors	3	7	3	3
		Weights	Distance	Distance	Distance	Distance
		Algorithm	Auto	Auto	Auto	Auto
		Leaf size	10	10	10	10
	Random Forest	p	2	2	1	1
		N estimators	100	100	20	50
		Max depth	None	None	None	None
		Min samples split	5	2	2	5
	SVM	Min samples leaf	2	1	1	1
		Max features	sqrt	Log2	Log2	Sqrt
		Bootstrap	False	False	False	True
		C	0.01	100	100	10
	Gradient Boosting	Kernel	Linear	rbf	Rbf	Linear
		Gamma	Scale	Scale	Scale	Scale
		Class weight	Balanced	None	None	None
		N estimators	50	100	50	50
	XG Boosting	Learning rate	0.2	0.2	0.2	0.2
		Max depth	6	6	5	5
		Min samples split	5	10	2	2
		Min samples leaf	1	1	1	2
		Subsample	0.8	1	1	0.6
		Max features	Log2	Sqrt	Log2	Sqrt
		N estimators	100	200	300	200
		Learning rate	0.1	0.1	0.05	0.05
		Max depth	3	5	3	4
		Subsample	0.8	1	1	1
		Colsample bytree	0.8	1	1	0.8

3.4 Evaluation metrics

In this study, five evaluation metrics were employed to assess the performance of the classification model [51]: accuracy, precision, recall, F1-score, and AUC. The selection of the most appropriate evaluation metric for a classification task depends on the specific objectives and context of the problem. Utilizing multiple metrics can provide a more comprehensive understanding of the model's performance. The formulas for these evaluation metrics are presented in Eqs. (2)–(6). In these equations: TP (True Positive) denotes the number of correctly identified positive instances; TN (True Negative) represents the number of correctly identified negative instances; FN (False Negative) indicates the number of positive instances incorrectly classified as negative; and FP (False Positive) refers to negative instances incorrectly classified as positive. Additionally, TPR and FPR correspond to the true positive rate and false positive rate, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

$$AUC = \int_0^1 TPR(FPR) d(FPR) \quad (6)$$

4. RESULTS AND DISCUSSION

This section discusses the implementation results of various ML algorithms, such as KNN, RF, SVM, GB, and XGB, to predict DM using two datasets: BRFS and PIMA. Each algorithm underwent a hyperparameter tuning process with Grid CV to obtain the best performance. Performance evaluation was performed using five metrics: accuracy, precision, recall, F1-score, and ROC/AUC. The performance results of each model were then compared to determine the most appropriate algorithm for each dataset and the most consistent algorithm between metrics in both datasets. In addition, the effect of resampling methods such as SMOTE, SMOTE-TOMEK, and SMOTE-ENN on the model

performance was also analyzed.

4.1 Before resampling

As an initial step, five ML learning algorithms were applied to both datasets without any resampling to evaluate their performance on the original data. Table 5 displays the evaluation metrics—accuracy, precision, recall, F1-score, and ROC/AUC—for each algorithm on each dataset. For the BRFS dataset, the performance metrics ranged as follows: accuracy (82–83%), precision (51–75%), recall (7–16%), F1-score (13–26%), and ROC/AUC (73–82%). For the PIMA Indian dataset, the ranges were: accuracy (81–87%), precision (74–89%), recall (76–82%), F1-score (77–82%), and ROC/AUC (85–92%). These results suggest that the accuracy values achieved by the five algorithms on both datasets are relatively modest and not particularly impressive. In addition to the accuracy value, what is more important to notice is its consistency with the values of other metrics, such as precision, recall, F1-score, and ROC/AUC, which also need to be considered to see the quality of the model.

Before resampling, the consistency of values between metrics for both datasets could be low. The most severe inconsistency occurred with the KNN model on the BRFS Diabetes 2015 dataset, where it has an accuracy of 0.82, while the precision, recall, F1-score, and ROC/AUC show lower values, namely 0.51, 0.15, 0.23, and 0.74, respectively. Such inconsistencies are also observed in the Pima Indian dataset, although they are not as severe. Inconsistent metrics can provide a misleading picture of model performance; for example, high accuracy may be due to the dominance of the majority class, while the performance of the minority class can be abysmal [52].

Table 5 displays the initial analysis results on the BRFS Diabetes 2015 and PIMA Indian datasets, which show that specific models, such as KNN, produced inconsistent metric performance. This finding suggests that dataset characteristics have a significant influence on algorithm performance [53]. Balancing plays a vital role in improving the consistency of performance metrics [54]. This technique improves the representation of minority classes in the dataset, supporting the algorithm to produce more accurate results [55, 56].

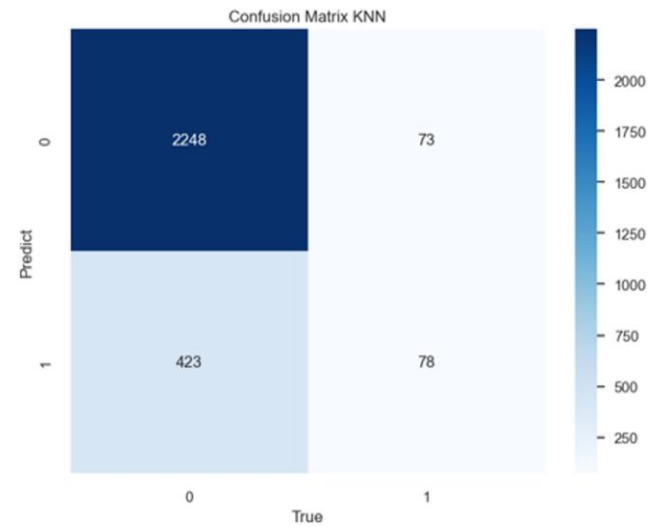
Table 5 shows the values of evaluation metrics provided by different ML models before resampling for the two datasets. It was observed that the evaluation metrics values were inconsistent, where other metrics do not always follow higher accuracy values. The most striking difference is between the accuracy and recall values. As previously reported, imbalanced data may produce inconsistent performance between evaluation metrics [57].

Table 5. Evaluation metrics of five ml models for each dataset before resampling

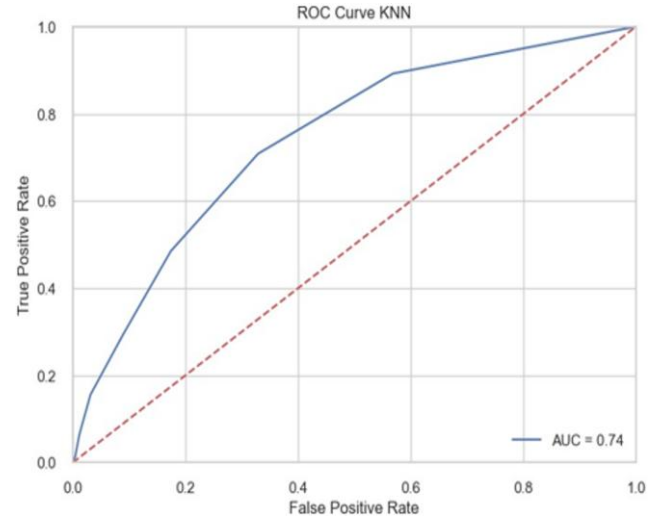
Dataset	Algorithm	Accuracy	Precision	Recall	F1 Score	Roc/Auc
Brfss Diabetes 2015	SVM	0.83	0.75	0.07	0.13	0.73
	KNN	0.82	0.51	0.15	0.23	0.74
	RF	0.83	0.63	0.11	0.19	0.81
	GB	0.83	0.65	0.17	0.27	0.82
	XGB	0.83	0.66	0.16	0.26	0.82
Pima Indian	SVM	0.81	0.74	0.80	0.77	0.85
	KNN	0.86	0.82	0.82	0.82	0.91
	RF	0.83	0.86	0.69	0.77	0.93
	GB	0.87	0.89	0.76	0.82	0.92
	XGB	0.87	0.87	0.78	0.82	0.92

Accuracy is not the only metric that determines model quality [58]. It measures the number of correct predictions without considering the distribution or balance of the data, making it less relevant for imbalanced data [59]. Using accuracy as the only metric in unbalanced data can be misleading. For example, suppose the majority of the data contains non-diabetic patients. In that case, a model that always predicts "non-diabetic" will still achieve high accuracy despite failing to detect actual cases. Therefore, additional metrics like recall (which measures the model's ability to identify positive cases) and F1-score (which represents the balance between precision and recall) are essential for a more comprehensive evaluation of the model's performance.

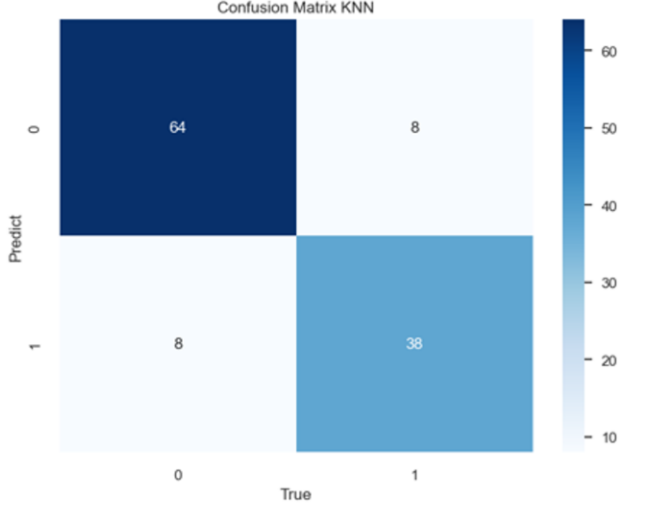
Figure 3(a) presents the confusion matrix of the KNN model applied to the BRFSS Diabetes 2015 dataset before resampling. The model correctly predicted 2,248 positive cases (class 0) and 78 negative cases (class 1), but misclassified 423 positive cases as negative and 73 negative cases as positive. The higher risk lies in misclassifying diabetic patients as non-diabetic, which underscores the challenges posed by using an imbalanced dataset [60]. Figure 3(b) displays the ROC curve of the KNN model, with an AUC value of 0.74, indicating moderate classification performance. An AUC between 0.70 and 0.80 is considered acceptable but warrants caution, whereas values above 0.90 are regarded as excellent.



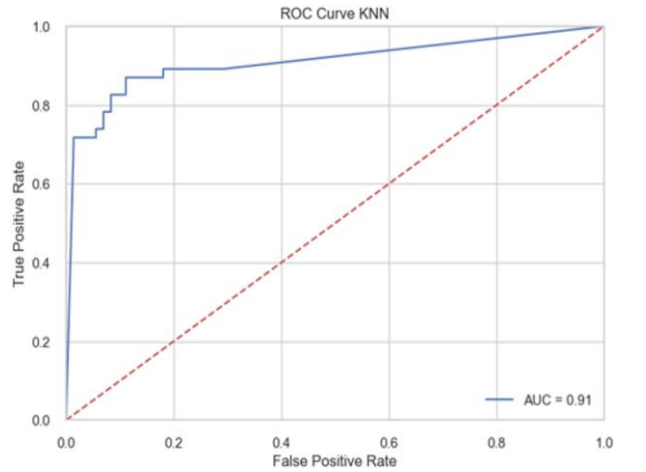
(a) Confusion matrix value of the KNN BRFSS 2015



(b) ROC/AUC value of the KNN BRFSS 2015



(c) Confusion matrix value of the KNN PIMA Indian



(d) ROC/AUC value of the KNN PIMA Indian

Figure 3. (a), (c) Confusion matrix value of the KNN model. (b), (d) ROC/AUC value of the KNN model before resampling the BRFSS 2015 and PIMA Indian

Figure 3(c) shows the confusion matrix for the PIMA Indian dataset, where the model correctly classified 64 instances of class 0 and 38 instances of class 1, with eight misclassifications in each class. The model achieved an accuracy of 86.44%, and for class 1, it yielded a precision, recall, and F1-score of 82.61%.

Finally, Figure 3(d) illustrates the ROC curve of the KNN model on the same dataset, with an AUC of 0.91, indicating excellent classification ability in distinguishing between the two classes. A higher AUC value reflects better model performance [61]. However, it is worth noting that the KNN model's AUC value of 0.91 before resampling on the PIMA Indian dataset, while high, still falls within the moderate range and suggests room for further improvement.

Figures 3, 4, and 5 illustrate the clinical implications of the model's performance. In a clinical context, an FN means that a diabetic patient is not detected by the model, which may delay diagnosis and treatment, thus worsening the patient's condition. Conversely, FP occurs when individuals who do not have diabetes are misclassified as diabetics, which may cause excessive anxiety and lead to unnecessary follow-up examinations. Therefore, an effective screening model should prioritize reducing FN while keeping FP rates within acceptable limits, to ensure early and accurate detection without harming patients.

4.2 After resampling

Three methods, SMOTE, SMOTE-TOMEK, and SMOTE-ENN, were applied to the BRFSS and PIMA datasets to evaluate the impact of resampling techniques on enhancing the performance of the five machine learning algorithms. These

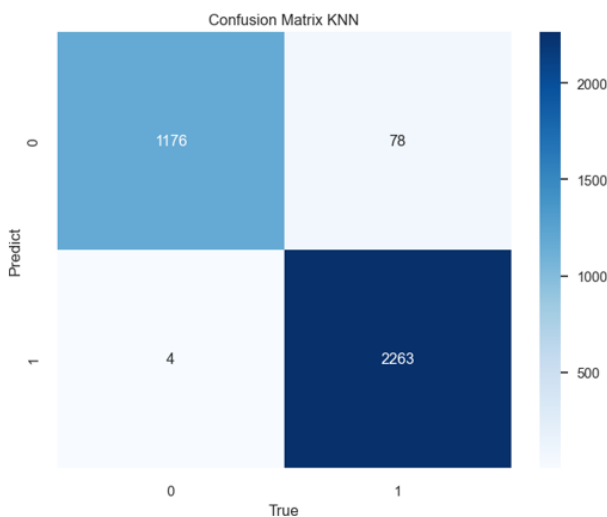
techniques aimed to identify the most effective approach for addressing class imbalance in the BRFSS Diabetes 2015 and PIMA Indian datasets. The following section outlines the performance of the five algorithms after resampling, based on their accuracy, precision, recall, F1-score, and ROC/AUC scores for each dataset.

Table 6. Evaluation metrics of five ML models for the BRFSS 2015 diabetes after resampling

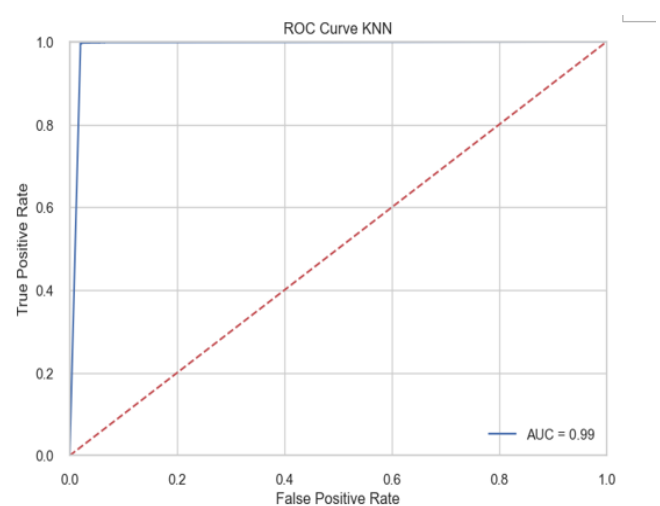
Resampling	Algorithm	Accuracy	Precision	Recall	F1-Score	ROC/AUC
SMOTE	KNN	0.90	0.84	0.98	0.91	0.96
	RF	0.91	0.96	0.85	0.90	0.97
	SVM	0.91	0.87	0.96	0.92	0.96
	GB	0.90	0.95	0.84	0.89	0.96
	XGB	0.90	0.96	0.84	0.89	0.96
SMOTE-TOMEK	KNN	0.90	0.84	0.99	0.91	0.95
	RF	0.90	0.95	0.85	0.91	0.97
	SVM	0.91	0.87	0.96	0.91	0.96
	GB	0.89	0.94	0.85	0.89	0.96
	XGB	0.90	0.95	0.84	0.89	0.96
SMOTE-ENN	KNN	0.97	0.96	0.99	0.98	0.99
	RF	0.94	0.97	0.94	0.96	0.99
	SVM	0.96	0.95	0.99	0.97	0.98
	GB	0.94	0.97	0.94	0.95	0.99
	XGB	0.94	0.97	0.94	0.95	0.99

Table 7. Evaluation metrics of five ML models for PIMA Indian dataset after resampling

Resampling	Algorithm	Accuracy	Precision	Recall	F1-Score	ROC/AUC
SMOTE	KNN	0.90	0.89	0.92	0.91	0.97
	RF	0.91	0.92	0.91	0.92	0.97
	SVM	0.89	0.91	0.88	0.89	0.93
	GB	0.91	0.91	0.91	0.91	0.98
	XGB	0.92	0.91	0.94	0.92	0.97
SMOTE-TOMEK	KNN	0.94	0.92	0.98	0.95	0.98
	RF	0.94	0.97	0.92	0.95	0.98
	SVM	0.89	0.93	0.86	0.90	0.95
	GB	0.94	0.97	0.91	0.94	0.98
	XGB	0.94	0.96	0.94	0.95	0.98
SMOTE-ENN	KNN	1.0	1.0	1.0	1.0	1.0
	RF	0.99	0.98	1.0	0.99	1.0
	SVM	0.97	0.96	0.98	0.97	0.97
	GB	0.99	0.98	1.0	0.99	1.0
	XGB	0.99	0.98	1.0	0.99	1.0

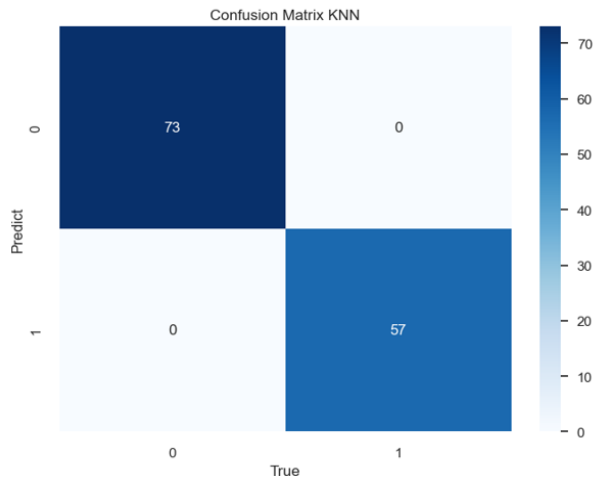


(a) Confusion matrix value of the KNN BRFSS 2015

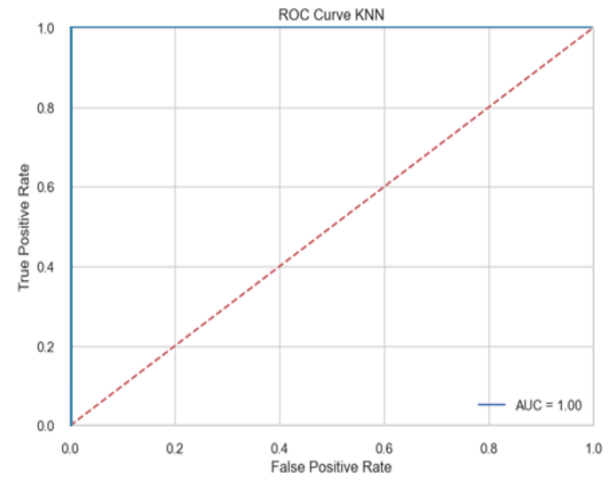


(b) ROC/AUC value of the KNN BRFSS 2015

Figure 4. (a) Confusion matrix value of the KNN model, (b) ROC/AUC value of the KNN model after resampling the BRFSS 2015



(a) Confusion matrix value of the KNN PIMA Indian



(b) ROC/AUC value of the KNN PIMA Indian

Figure 5. (a) Confusion matrix value of the KNN model, (b) ROC/AUC value of the KNN model after resampling the PIMA Indian dataset

Tables 6 and 7 present the performance of the five models after resampling using SMOTE, SMOTE-TOMEK, and SMOTE-ENN. Compared to metric values before resampling in Table 5, any resampling technique could improve model performance for all algorithms and all evaluation metric values. The SMOTE-ENN is the best technique for the highest performance for all five models and both datasets [26]. For the BRFS Diabetes 2015 dataset, the KNN model was the best, with an accuracy of 0.97, approaching the results of previous research of 0.98 [26]. The values of the other four metrics were in the range of 0.96-0.99, which is also close to previous results [26]. This was not only true for the BRFS dataset; combining the SMOTE-ENN and KNN models was also the best resampling technique and model for the PIMA Indian dataset, with perfect values (1.0) for all evaluation metrics.

As shown in Tables 6 and 7, resampling techniques enhance model performance and reduce the variation in results across different algorithms, leading to minimal differences in accuracy, precision, recall, F1-score, and ROC/AUC. For instance, in the PIMA Indian dataset, the maximum performance gap between models was limited to just 2–4% across all evaluation metrics.

From this perspective, for imbalanced diabetes data, the issue of finding a resampling technique is more worthy of attention than finding the best algorithm. A good resampling technique can improve the correlation between features and targets, facilitating any algorithm's recognition of data patterns.

The role of the resampling technique on model performance can be explained through Figures 4(a) and (b), which present the confusion matrix and ROC/AUC of the KNN model for the BRFS Diabetes 2015 dataset. Out of 1,180 positive cases, the model only made four errors, indicating that the model's ability to detect positive cases is very high. In addition, out of 2341 negative cases, the model made 78 errors, denoting a relatively small error rate compared to the total number of negative cases.

This overview was also observed in the other four models. Compared to Figure 4 before resampling, the SMOTE-ENN method could help the five ML models detect positive cases accurately and almost not miss truly positive cases. The ROC/AUC curves demonstrate that the ML models perform very well in distinguishing between the positive class (class 0)

and the negative class (class 1). This curve was also observed in other ML models for both datasets. The model could achieve a very high (TPR), which means it correctly detects almost all positive cases, while maintaining a very low (FPR), which means only a few negative cases are incorrectly classified as positive. With an AUC value of 0.99, the model exhibited almost perfect ability to separate the two classes. An AUC value approaching 1 indicates that the model effectively makes accurate predictions.

Testing with an AUC value above 0.90 denotes that the model has a very superior and reliable classification performance. A high AUC suggests that the model can distinguish between positive and negative classes with very good accuracy [61].

The role of the SMOTE-ENN technique in the KNN algorithm for the PIMA Indian dataset is also very clearly observed in the confusion matrix and ROC/AUC presented in Figures 5 (a) and (b). The model successfully predicted all data correctly, namely 73 data as class 0 (TN) and 57 data as class 1 (TP), without any errors in the prediction (False Positive and False Negative are zero).

This implies that the model accuracy reached 1.0. Furthermore, the ROC Curve strengthens these results with an AUC value of 1.0, indicating that the KNN model performs perfectly in distinguishing between the two classes. Overall, the KNN model showed optimal performance on the PIMA Indian dataset.

4.3 Discussion

ML algorithms (KNN, RF, SVM, GB, and XGB) were applied to two datasets without resampling to evaluate the initial performance of the models. Without resampling, the performance of all five ML models on the BRFS and PIMA datasets is less than satisfactory, with varying and inconsistent values of accuracy, precision, recall, F1-score, and ROC/AUC. For example, the most extreme inconsistencies were observed for the KNN algorithm implemented on the BRFS dataset, where its accuracy was 0.82, but the recall was only 0.15.

Three resampling methods, namely SMOTE, SMOTE-TOMEK, and SMOTE-ENN, were employed to address the class imbalance in the BRFS Diabetes 2015 and PIMA Indian

datasets by combining oversampling and undersampling techniques. The results displayed significant improvements in accuracy, precision, recall, F1-score, and ROC/AUC. SMOTE-ENN, a combination of SMOTE and ENN, gave the best results, significantly increasing metric values. Resampling techniques can minimize performance differences between algorithms so that they do not differ significantly in accuracy, precision, recall, F1-score, and ROC/AUC. For example, for the PIMA Indian dataset, the differences in performance between models were relatively small, ranging only from 2% to 4% across metrics such as accuracy, precision, recall, F1-score, and ROC/AUC. Among the resampling techniques used, SMOTE-ENN consistently performed the best in handling class imbalance, establishing it as the most effective method in this analysis.

Figure 6 illustrates the correlation values between features and targets for the BRFSS Diabetes 2015 dataset before and after resampling using the SMOTE-ENN technique. In general, resampling increases the correlation values between features and the target, either positive or negative. Features with positive correlations, such as high BP, BMI, and stroke, experienced significant increases after resampling, indicating that data redistribution helped strengthen the positive

relationship between these features and diabetes. Meanwhile, several features with negative correlations, such as PhysActivity (physical activity) and Veggies (vegetable consumption), also increase by resampling, emphasizing the inverse relationship between these features and diabetes risk. Overall, resampling provides a more representative dataset and allows for a more in-depth analysis of the relationship between variables. Virtual samples improve distribution uniformity and strengthen the relationship between features and targets [62].

Figure 7 presents the correlation values between features and targets for the PIMA Indian dataset before and after resampling. It appears that the resampling keeps the pattern of the dataset, namely that the glucose and insulin features have the highest correlation values among the other features, meaning that these two features are dominant. Resampling on the PIMA Indian dataset increased the correlation value between features and targets, and this increase was observed in all features, including the most dominant features, namely glucose and insulin. Resampling improves the correlation between features and targets, making it easier for the ML model to recognize patterns, thereby increasing its accuracy, including the values of other evaluation metrics.

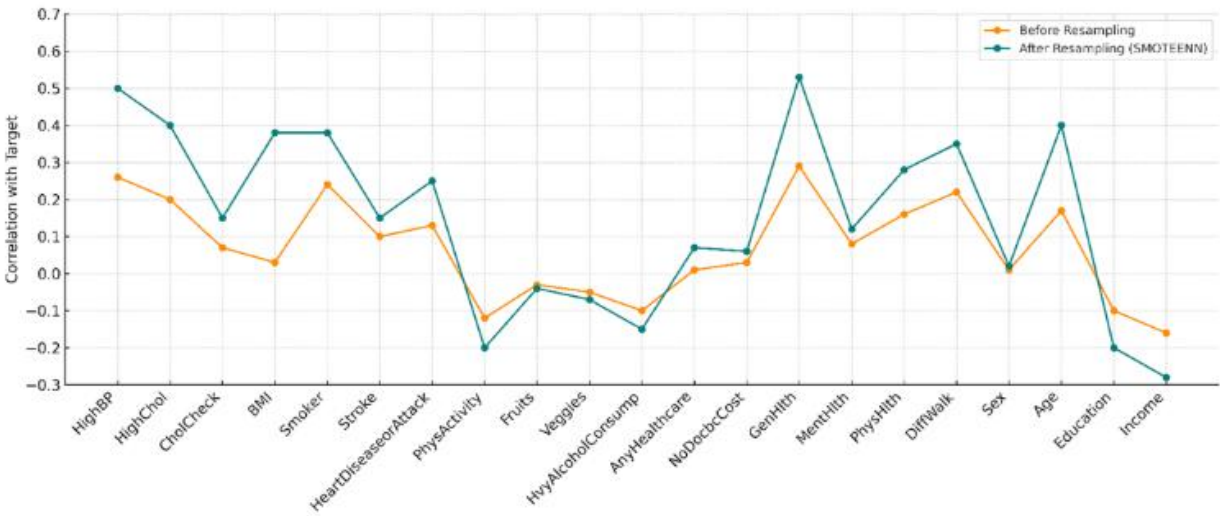


Figure 6. Correlation between features and targets before and after resampling

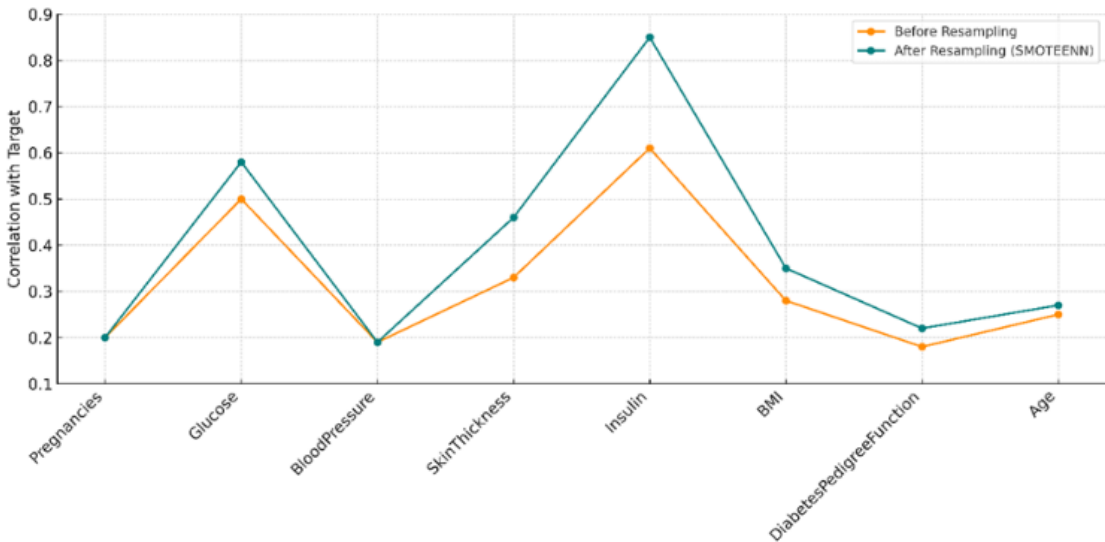


Figure 7. Correlation between features and targets before and after resampling

Table 8. Comparison of the proposed method with existing studies using the BRFSS 2015 diabetes

Study	Dataset	Highest Accuracy	Accuracy	Precision	Recall	F1-Score	ROC/AUC
[31]	BRFSS 2015	RF	0.82	0.83	0.80	0.82	0.82
[26]	BRFSS 2015	KNN	0.98	0.98	0.98	0.98	0.98
[51]	BRFSS 2015	RF	0.96	0.80	0.80	-	0.99
Our Proposed Method	BRFSS 2015	KNN (SMOTE-ENN)	0.97	0.96	0.99	0.98	0.99

Table 9. Comparison of the proposed method with existing studies using the PIMA Indian dataset

Study	Dataset	Highest Accuracy	Accuracy	Precision	Recall	F1-Score	ROC/AUC
[16]	PIMA Indian	RF	0.75	0.84	0.78	0.81	1.0
[17]	PIMA Indian	SVM	0.84	0.71	0.66	0.67	0.80
[18]	PIMA Indian	GB	0.76	0.68	0.59	0.63	0.82
[34]	PIMA Indian	SVM	0.83	-	0.53	0.64	-
[32]	PIMA Indian	SVM	0.74	0.74	0.74	0.74	0.74
Our Proposed Method	PIMA Indian	KNN (SMOTE-ENN)	1.0	1.0	1.0	1.0	1.0

The proposed method in this study has produced good results in several assessment metrics for predicting DM disease. The proposed method exhibited high accuracy, precision, recall, F1-score, and ROC/AUC performance. Tables 8 and 9 show the performance of the proposed framework, evaluated based on methodology and accuracy, compared with several relevant studies. Table 7 compares the performance of ML models on the BRFSS Diabetes 2015 dataset. The proposed method, KNN, obtained an accuracy value of 0.97, close to the previous study result of 0.98 [26], indicating excellent performance.

This method excels in detecting positive cases and maintaining performance balance while effectively handling imbalanced data, and the results are comparable to previous methods. Table 8 also compares the ML models on the PIMA Indian dataset for diabetes prediction. The previous study achieved the highest accuracy of 0.84 using SVM, while other models, such as RF and GB, had lower accuracy. The proposed method, KNN with SMOTE-ENN, produced a performance of all metrics reaching a perfect value of 1.0. This method effectively detects all positive cases (recall = 1.0) and produces accurate predictions without errors (precision = 1.0) thanks to resampling using the SMOTE-ENN technique.

5. CONCLUSIONS

This study demonstrated that applying the SMOTE-ENN resampling technique significantly enhanced the performance of machine learning algorithms in predicting diabetes mellitus (DM) on the imbalanced BRFSS Diabetes 2015 and PIMA Indian datasets. SMOTE-ENN improved accuracy, precision, recall, F1-score, and ROC/AUC, resulting in more consistent models in identifying both positive and negative cases. On the BRFSS Diabetes 2015 dataset, the KNN algorithm achieved an accuracy of 0.97, with other metrics also showing strong performance, closely aligning with findings from previous research. In the PIMA Indian dataset, the KNN model combined with SMOTE-ENN delivered optimal results, achieving perfect scores (1.0) across all evaluation metrics.

The confusion matrix analysis uncovered that the SMOTE-ENN technique significantly improved the model's ability to detect minority classes (positive cases of diabetes), which were previously often overlooked in imbalanced datasets. For example, in the BRFSS Diabetes 2015 dataset, the KNN

model, after resampling, only made a few errors in predicting positive and negative classes, resulting in a very low (FPR). An even better result was observed in the PIMA Indian dataset, where the KNN model successfully predicted all data without error. The model achieved a perfect AUC value (1.0) for the latter dataset, indicating an excellent ability to distinguish between positive and negative classes.

The SMOTE-ENN resampling technique improves all five ML models performance because it strengthens the relationship between features and target. The positive correlation of key features, such as high BP, BMI, and Stroke, with the target significantly increased after resampling. In addition, features with negative correlations, such as Phys Activity and Veggies, also strengthened the inverse relationship, emphasizing the role of these features in supporting diabetes prediction.

This proves that the data redistribution could improve dataset representation and reduce distortion due to data imbalance. After resampling, the increase in correlation between features and targets indicates a more representative data redistribution. The resampling technique improves the overall model performance and minimizes the performance differences between algorithms. In imbalanced datasets, the issue of resampling techniques is more worthy of attention than that of searching for the best algorithm. The practical implication of this study is that SMOTE-ENN can be used as part of an automated screening system for early detection of diabetes in primary healthcare, including clinics and health centers. This approach can also be integrated into web-based or mobile health applications. In the future, the research will be extended to test the effectiveness of this method on multiclass data as well as real-time scenarios, such as complication prediction or DM severity classification.

ACKNOWLEDGMENT

All computations were done using the Computing Facility at the Center for Quantum Computing and Materials Informatics Research, Faculty of Computer Science, Dian Nuswantoro University. The authors gratefully acknowledge the support provided by the Doctoral Program in Computer Science at Dian Nuswantoro University, Indonesia, and the financial assistance from the Faculty of Technology and Informatics, Aisyah University, Indonesia.

REFERENCES

- [1] Misra, A., Gopalan, H., Jayawardena, R., Hills, A.P., Soares, M., Reza-Albarrán, A.A., Ramaiya, K.L. (2019). Diabetes in developing countries. *Journal of Diabetes*, 11(7): 522-539. <https://doi.org/10.1111/1753-0407.12913>
- [2] American Diabetes Association. (2018). 2. Classification and diagnosis of diabetes: Standards of medical care in diabetes—2018. *Diabetes Care*, 41(Supplement_1): S13-S27. <https://doi.org/10.2337/dc18-S002>
- [3] Zhou, H., Myrzashova, R., Zheng, R. (2020). Diabetes prediction model based on an enhanced deep neural network. *EURASIP Journal on Wireless Communications and Networking*, 2020: 1-13. <https://doi.org/10.1186/s13638-020-01765-7>
- [4] Prakash, B.B.N.S., Naveen, B., Akhiluzzama, M.D., Rajarajeswari, P. (2023). Comparative performance analysis of quantum algorithm with machine learning algorithms on diabetes mellitus. In *2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, Bengaluru, India, pp. 1178-1183. <https://doi.org/10.1109/IITCEE57236.2023.10090957>
- [5] Ghosh, P., Azam, S., Karim, A., Hassan, M., Roy, K., Jonkman, M. (2021). A comparative study of different machine learning tools in detecting diabetes. *Procedia Computer Science*, 192: 467-477. <https://doi.org/10.1016/j.procs.2021.08.048>
- [6] Gupta, H., Varshney, H., Sharma, T.K., Pachauri, N., Verma, O.P. (2022). Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. *Complex & Intelligent Systems*, 8(4): 3073-3087. <https://doi.org/10.1007/s40747-021-00398-7>
- [7] Hameed, N., Shabut, A.M., Ghosh, M.K., Hossain, M.A. (2020). Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques. *Expert Systems with Applications*, 141: 112961. <https://doi.org/10.1016/j.eswa.2019.112961>
- [8] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., IDF Diabetes Atlas Committee. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas. *Diabetes Research and Clinical Practice*, 157: 107843. <https://doi.org/10.1016/j.diabres.2019.107843>
- [9] Naz, H., Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders*, 19: 391-403. <https://doi.org/10.1007/s40200-020-00520-5>
- [10] Zohair, M., Chandra, R., Tiwari, S., Agarwal, S. (2024). A model fusion approach for severity prediction of diabetes with respect to binary and multiclass classification. *International Journal of Information Technology*, 16(3): 1955-1965. <https://doi.org/10.1007/s41870-023-01463-9>
- [11] Hasan, M.K., Alam, M.A., Das, D., Hossain, E., Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, 8: 76516-76531. <https://doi.org/10.1109/ACCESS.2020.2989857>
- [12] Patil, S., Patil, K.R., Patil, C.R., Patil, S.S. (2020). Performance overview of an artificial intelligence in biomedics: A systematic approach. *International Journal of Information Technology*, 12(3): 963-973. <https://doi.org/10.1007/s41870-018-0243-8>
- [13] Mohapatra, D., Bhoi, S.K., Mallick, C., Jena, K.K., Mishra, S. (2022). Distribution preserving train-test split directed ensemble classifier for heart disease prediction. *International Journal of Information Technology*, 14(4): 1763-1769. <https://doi.org/10.1007/s41870-022-00868-2>
- [14] Bhat, S.S., Banu, M., Ansari, G.A., Selvam, V. (2023). A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms. *Healthcare Analytics*, 4: 100273. <https://doi.org/10.1016/j.health.2023.100273>
- [15] Febrian, M.E., Ferdinan, F.X., Sendani, G.P., Suryanigrum, K.M., Yunanda, R. (2023). Diabetes prediction using supervised machine learning. *Procedia Computer Science*, 216: 21-30. <https://doi.org/10.1016/j.procs.2022.12.107>
- [16] Tigga, N.P., Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, 167: 706-716. <https://doi.org/10.1016/j.procs.2020.03.336>
- [17] Zulkifli, Z., Makkiyah, F.A., Antoni, D., Fitriana, F., Jamaan, T., Taufik, A. (2024). Multi-algorithm to measure the accuracy level of diabetes status prediction.
- [18] Nusrat, F., Uzbaş, B., Baykan, Ö.K. (2020). Gradient boosting classification kullanak diabetes mellitus tahmini. *European Journal of Science and Technology*. <https://doi.org/10.31590/ejosat.803504>
- [19] Doan, Q.H., Keshtegar, B., Kim, S.E., Thai, D.K. (2024). Generative adversarial networks for overlapped and imbalanced problems in impact damage classification. *Information Sciences*, 675: 120752. <https://doi.org/10.1016/j.ins.2024.120752>
- [20] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73: 220-239. <https://doi.org/10.1016/j.eswa.2016.12.035>
- [21] García, V., Sánchez, J.S., Mollineda, R.A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1): 13-21. <https://doi.org/10.1016/j.knosys.2011.06.013>
- [22] Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Cham: Springer. <https://doi.org/10.1007/978-3-319-98074-4>
- [23] Prayogo, R.D., Karimah, S.A. (2021). Feature selection and adaptive synthetic sampling approach for optimizing online shopper purchase intent prediction. In *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, Bandung, Indonesia, pp. 1-5. <https://doi.org/10.1109/ICAICTA53211.2021.9640270>
- [24] Tang, B., He, H. (2015). KernelADASYN: Kernel based adaptive synthetic data generation for imbalanced learning. In *2015 IEEE Congress on Evolutionary Computation (CEC)*, Sendai, Japan, pp. 664-671. <https://doi.org/10.1109/CEC.2015.7256954>
- [25] Hassan, H., Ahmad, N.B., Anuar, S. (2020). Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble

- approach in educational data mining. *Journal of Physics: Conference Series*, 1529(5): 052041. IOP Publishing. <https://doi.org/10.1088/1742-6596/1529/5/052041>
- [26] Ullah, Z., Saleem, F., Jamjoom, M., Fakieh, B., Kateb, F., Ali, A.M., Shah, B. (2022). Detecting high - risk factors and early diagnosis of diabetes using machine learning methods. *Computational Intelligence and Neuroscience*, 2022(1): 2557795. <https://doi.org/10.1155/2022/2557795>
- [27] Chowdhury, M.M., Ayon, R.S., Hossain, M.S. (2024). An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFS dataset. *Healthcare Analytics*, 5: 100297. <https://doi.org/10.1016/j.health.2023.100297>
- [28] Kahn, M. Diabetes. UCI Machine Learning Repository. Available: <https://archive.ics.uci.edu/dataset/34/diabetes>.
- [29] Butt, U.M., Letchmunan, S., Ali, M., Hassan, F.H., Baqir, A., Sherazi, H.H.R. (2021). Machine learning based diabetes classification and prediction for healthcare applications. *Journal of Healthcare Engineering*, 2021(1): 9930985. <https://doi.org/10.1155/2021/9930985>
- [30] Maghsoodi, A.I., Torkayesh, A.E., Wood, L.C., Herrera-Viedma, E., Govindan, K. (2023). A machine learning driven multiple criteria decision analysis using LS-SVM feature elimination: Sustainability performance assessment with incomplete data. *Engineering Applications of Artificial Intelligence*, 119: 105785. <https://doi.org/10.1016/j.engappai.2022.105785>
- [31] Chang, V., Ganatra, M.A., Hall, K., Golightly, L., Xu, Q.A. (2022). An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics*, 2: 100118. <https://doi.org/10.1016/j.health.2022.100118>
- [32] Kangra, K., Singh, J. (2023). Comparative analysis of predictive machine learning algorithms for diabetes mellitus. *Bulletin of Electrical Engineering and Informatics*, 12(3): 1728-1737. <https://doi.org/10.11591/eei.v12i3.4412>
- [33] Olisah, C.C., Smith, L., Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220: 106773. <https://doi.org/10.1016/j.cmpb.2022.106773>
- [34] Edeh, M.O., Khalaf, O.I., Tavera, C.A., Tayeb, S., Ghoulali, S., Abdulsahib, G.M., Louni, A. (2022). A classification algorithm-based hybrid diabetes prediction model. *Frontiers in Public Health*, 10: 829519. <https://doi.org/10.3389/fpubh.2022.829519>
- [35] Rezvani, S., Wang, X. (2023). A broad review on class imbalance learning techniques. *Applied Soft Computing*, 143: 110415. <https://doi.org/10.1016/j.asoc.2023.110415>
- [36] Antar, S.A., Ashour, N.A., Sharaky, M., Khattab, M., Ashour, N.A., Zaid, R.T., Al-Karmalawy, A.A. (2023). Diabetes mellitus: Classification, mediators, and complications; A gate to identify potential targets for the development of new effective treatments. *Biomedicine & Pharmacotherapy*, 168: 115734. <https://doi.org/10.1016/j.biopha.2023.115734>
- [37] Freiesleben, T., Grote, T. (2023). Beyond generalization: A theory of robustness in machine learning. *Synthese*, 202(4): 109. <https://doi.org/10.1007/s11229-023-04334-9>
- [38] Nematzadeh, S., Kiani, F., Torkamanian-Afshar, M., Aydin, N. (2022). Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases. *Computational Biology and Chemistry*, 97: 107619. <https://doi.org/10.1016/j.compbiolchem.2021.107619>
- [39] Ashraf, N.M., Mostafa, R.R., Sakr, R.H., Rashad, M.Z. (2021). Optimizing hyperparameters of deep reinforcement learning for autonomous driving based on whale optimization algorithm. *PloA ONE*, 16(6): e0252754. <https://doi.org/10.1371/journal.pone.0252754>
- [40] Charilaou, P., Battat, R. (2022). Machine learning models and over-fitting considerations. *World Journal of Gastroenterology*, 28(5): 605. <https://doi.org/10.3748/wjg.v28.i5.605>
- [41] Teboul, A. (2022). Diabetes Health Indicators Dataset. Kaggle. <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/>.
- [42] Ayon, S.I., Islam, M.M. (2019). Diabetes prediction: A deep learning approach. *International Journal of Information Engineering and Electronic Business*, 13(2): 21. <https://doi.org/10.5815/ijieeb.2019.02.03>
- [43] Kwak, S.K., Kim, J.H. (2017). Statistical data preparation: Management of missing values and outliers. *Korean Journal of Anesthesiology*, 70(4): 407-411. <https://doi.org/10.4097/kjae.2017.70.4.407>
- [44] Trier, A.M., Mack, M.R., Kim, B.S. (2019). The neuroimmune axis in skin sensation, inflammation, and immunity. *The Journal of Immunology*, 202(10), 2829-2835. <https://doi.org/10.4049/jimmunol.1801473>
- [45] Dash, C.S.K., Behera, A.K., Dehuri, S., Ghosh, A. (2023). An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal*, 6: 100164. <https://doi.org/10.1016/j.dajour.2023.100164>
- [46] Kim, Y.S., Kim, M.K., Fu, N., Liu, J., Wang, J., Srebric, J. (2025). Investigating the impact of data normalization methods on predicting electricity consumption in a building using different artificial neural network models. *Sustainable Cities and Society*, 118: 105570. <https://doi.org/10.1016/j.scs.2024.105570>
- [47] Chachoui, Y., Azizi, N., Hotte, R., Bensebaa, T. (2024). Enhancing algorithmic assessment in education: Equifused-data-based SMOTE for balanced learning. *computers and education: Artificial Intelligence*, 6: 100222. <https://doi.org/10.1016/j.caeai.2024.100222>
- [48] Newman, M.E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5): 323-351.
- [49] Joseph, V.R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(4): 531-538. <https://doi.org/10.1002/sam.11583>
- [50] Akrom, M., Rustad, S., Dipojono, H.K. (2024). Development of quantum machine learning to evaluate the corrosion inhibition capability of pyrimidine compounds. *Materials Today Communications*, 39: 108758. <https://doi.org/10.1016/j.mtcomm.2024.108758>
- [51] Pias, T.S., Su, Y., Tang, X., Wang, H., Faghani, S., Yao, D. (2024). Enhancing fairness and accuracy in type 2 diabetes prediction through data resampling. In *Proceedings of Machine Learning Research LEAVE*

- UNSET, pp. 1-17. <https://doi.org/10.1101/2023.05.02.23289405>
- [52] Luque, A., Carrasco, A., Martín, A., de Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91: 216-231. <https://doi.org/10.1016/j.patcog.2019.02.023>
- [53] Sahid, M.A., Babar, M.U.H., Uddin, M.P. (2024). Predictive modeling of multi-class diabetes mellitus using machine learning and filtering Iraqi diabetes data dynamics. *PloS ONE*, 19(5): e0300785. <https://doi.org/10.1371/journal.pone.0300785>
- [54] Gurcan, F., Soylu, A. (2024). Learning from imbalanced data: Integration of advanced resampling techniques and machine learning models for enhanced cancer diagnosis and prognosis. *Cancers*, 16(19): 3417. <https://doi.org/10.3390/cancers16193417>
- [55] Huang, S., Yang, J., Fong, S., Zhao, Q. (2020). Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters*, 471: 61-71. <https://doi.org/10.1016/j.canlet.2019.12.007>
- [56] Kraiem, M.S., Sánchez-Hernández, F., Moreno-García, M.N. (2021). Selecting the suitable resampling strategy for imbalanced data classification regarding dataset properties. An approach based on association models. *Applied Sciences*, 11(18): 8546. <https://doi.org/10.3390/app11188546>
- [57] Gaudreault, J.G., Branco, P. (2024). Empirical analysis of performance assessment for imbalanced classification. *Machine Learning*, 113(8): 5533-5575. <https://doi.org/10.1007/s10994-023-06497-5>
- [58] Uddin, M.A., Islam, M.M., Talukder, M.A., Hossain, M.A.A., Akhter, A., Aryal, S., Muntaha, M. (2024). Machine learning based diabetes detection model for false negative reduction. *Biomedical Materials & Devices*, 2(1): 427-443. <https://doi.org/10.1007/s44174-023-00104-w>
- [59] Opitz, J. (2024). A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics*, 12: 820-836. https://doi.org/10.1162/tacl_a_00675
- [60] Canbek, G., Taskaya Temizel, T., Sagioglu, S. (2022). PToPI: A comprehensive review, analysis, and knowledge representation of binary classification performance measures/metrics. *SN Computer Science*, 4(1): 13. <https://doi.org/10.1007/s42979-022-01409-1>
- [61] Çorbacioğlu, Ş.K., Aksel, G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish Journal of Emergency Medicine*, 23(4): 195-198. https://doi.org/10.4103/tjem.tjem_182_23
- [62] Rustad, S., Akrom, M., Sutojo, T., Dipojono, H.K. (2024). A feature restoration for machine learning on anti-corrosion materials. *Case Studies in Chemical and Environmental Engineering*, 10: 100902. <https://doi.org/10.1016/j.cscee.2024.100902>