Celal Bayar University Journal of Science
Volume 22, Issue 2, 2026, p 225-235
Doi: 10.18466/cbayarfbe.1766229 Research Article

Celal Bayar University Journal of Science

A Genetic Algorithm-Enhanced Method for Missing Value Imputation in Healthcare Datasets

Hatice Nizam-Ozogur[1]*, Zeynep Orman[2]

[1] Istanbul University, Department of Artificial Intelligence and Data Engineering, Istanbul, Türkiye.
[2] Istanbul University-Cerrahpasa, Department of Computer Engineering, Istanbul, Türkiye.
* hatice.nizamozogur@istanbul.edu.tr
* Orcid No: 0000-0002-9722-4355

Received: August 16, 2025
Accepted: January 26, 2026
DOI: 10.18466/cbayarfbe.1766229

Abstract

In healthcare datasets, imbalanced class distributions and missing data pose significant challenges to the performance and stability of machine learning models, thereby hindering accurate analysis and disease diagnosis. Addressing these challenges is crucial for improving both the precision and reliability of healthcare data analysis. This paper proposes a novel preprocessing framework specifically designed for healthcare datasets to mitigate issues related to incomplete data and class imbalance. The framework introduces a new imputation method, GA-MICE, which enhances the Multiple Imputation by Chained Equations (MICE) technique using a Genetic Algorithm (GA) to improve the accuracy of handling missing data. Additionally, the framework incorporates the GASMOTEPSO_ENN method, which combines the Synthetic Minority Oversampling Technique (SMOTE) and Edited Nearest Neighbor (ENN) algorithms with GA and Particle Swarm Optimization (PSO) heuristics to effectively address class imbalance. After preprocessing, six machine learning classifiers are employed to categorize individuals as either patients or healthy subjects. The model's performance is evaluated using multiple metrics, including accuracy, precision, recall, F1-score, Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUC). Experimental results demonstrate the effectiveness of the proposed approach in managing missing data and addressing class imbalance, achieving performance close to or exceeding existing methodologies reported in the literature.

Keywords: Disease classification, healthcare dataset, heuristic approach, imbalanced data, missing value.

1. Introduction

Real-world datasets frequently contain missing values in certain attributes, a problem that is particularly common in healthcare-related data. Such missing values typically arise from incomplete patient records, data entry errors by healthcare professionals, or malfunctions in medical equipment. Since many machine learning algorithms are unable to process incomplete datasets directly, their applicability in real-world healthcare environments is often limited [1]. Moreover, the presence of missing values can significantly degrade the performance of classification and prediction models by introducing bias into the learning process. In healthcare applications, this bias may lead to inaccurate diagnoses, inappropriate treatment decisions, and, in extreme cases, serious or life-threatening consequences for patients.

In addition to missing values, class imbalance represents another major challenge commonly observed in real-world datasets. Class imbalance occurs when one class is represented by a substantially larger number of instances than another. This issue is particularly pronounced in healthcare datasets, where rare diseases or minority classes are often associated with a limited number of samples. As a result, machine learning models tend to be biased toward the majority class, leading to misleading predictions and unreliable clinical decision-making. Therefore, addressing class imbalance is a critical prerequisite for achieving accurate and trustworthy healthcare data analysis.

To mitigate these challenges, numerous studies have proposed various techniques for handling missing values and imbalanced datasets in healthcare applications. One of the most straightforward approaches for dealing with missing data is the deletion method, which removes instances containing missing values from the dataset. This method is generally applicable when the proportion of missing data is relatively low, typically below 10%–15% of the dataset [2]. However, deletion-based approaches are often unsuitable for healthcare datasets, as they may result in the loss of valuable and rare patient records that could otherwise enhance predictive performance.

In practical healthcare datasets, missing values frequently occur at higher rates, and the affected instances are often too important to discard. In such scenarios, missing value imputation methods provide a more effective alternative [3]. These methods employ statistical or machine learning models to estimate and replace missing values [4]. Missing value handling techniques are broadly categorized into single imputation and multiple imputation approaches. Single imputation assigns a single estimated value to each missing entry, whereas multiple imputation performs several iterations to better capture the uncertainty and variability associated with missing data. Given the critical nature of healthcare decision-making, robust and reliable imputation techniques are essential to minimize the adverse effects of missing values.

Various strategies have also been proposed to address class imbalance in healthcare datasets, including resampling techniques, ensemble learning approaches, and cost-sensitive algorithms. Despite these efforts, achieving stable and robust performance on imbalanced healthcare data remains a challenging task. Furthermore, most existing studies focus on handling either missing values or class imbalance in isolation. Consequently, there is a clear need for integrated preprocessing frameworks that simultaneously address both challenges to improve the accuracy and reliability of healthcare data analysis.

To address this gap, this study proposes a comprehensive preprocessing framework specifically designed for healthcare datasets characterized by both missing values and imbalanced class distributions. The proposed system consists of three main stages: missing value imputation, class imbalance correction, and classification. A novel imputation approach, termed GA-MICE, is introduced by integrating the Multiple Imputation by Chained Equations (MICE) technique with a Genetic Algorithm (GA). To mitigate the effects of class imbalance, the GASMOTEPSO_ENN sampling method—previously introduced in our earlier work [5]—is incorporated into the framework.

The effectiveness of the proposed preprocessing system is evaluated using three healthcare datasets that exhibit both missing values and imbalanced class distributions. Multiple machine learning classifiers are employed to distinguish between healthy individuals and patients. The primary objective of this study is to develop an efficient and reliable imputation method tailored to healthcare datasets with missing values. A secondary objective is to construct an integrated preprocessing framework that jointly addresses missing value imputation and class imbalance by combining the proposed GA-MICE method with the GASMOTEPSO_ENN sampling strategy.

The remainder of this paper is organized as follows. Section 2 reviews related work on missing value imputation and class imbalance handling in healthcare datasets. Section 3 describes the proposed methodology. Section 4 presents the experimental framework, including the datasets, experimental setup, and evaluation metrics. Section 5 reports and analyzes the experimental results. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Related Work

Missing values and class imbalance are two pervasive challenges in data analysis and mining, particularly in healthcare applications, where they can significantly degrade predictive accuracy and introduce bias into decision-making processes [6]. Missing values frequently arise in healthcare datasets due to complex data acquisition procedures, patient non-disclosure, and technical or human-related recording errors [7]. These deficiencies can severely limit the reliability of data-driven models, especially in critical applications such as early disease diagnosis and treatment planning. In parallel, class imbalance—where one class substantially outnumbers others—further complicates learning, as models tend to favor majority classes, leading to suboptimal performance on clinically important minority cases. Consequently, addressing both missing values and class imbalance is essential to ensure robust and reliable healthcare data analysis.

A substantial body of research has focused on developing effective missing value imputation techniques for healthcare datasets. Khan and Hoque [8] proposed a hybrid imputation strategy that integrates single and multiple imputation approaches by extending the MICE framework. Their method demonstrated superior performance on both public UCI datasets and a large-scale real medical dataset, achieving notable improvements in F-score and error reduction. Raja and Thangavel [9] introduced a rough K-means centroid-based imputation approach and evaluated its effectiveness against several clustering-based imputation techniques using RMSE and MAE metrics across multiple benchmark datasets. Similarly, Madhu et al. [10] proposed an XGBoost-based imputation method capable of handling both continuous and categorical attributes, showing strong performance across datasets with varying levels of missingness. Rani et al. [11] developed a classifier-optimized hybrid imputation framework that

integrates MICE, KNN, and statistical imputation methods, yielding significant reductions in RMSE and notable improvements in classification accuracy even under high missing data rates.

Beyond standalone imputation techniques, several studies have investigated frameworks that jointly consider missing values and class imbalance. Wang et al. [12] introduced the DMP_MI approach, which combines Naive Bayes-based imputation with adaptive synthetic (ADASYN) for class balancing and Random Forest (RF) for classification, achieving superior results on the PIMA Indians Diabetes Disease Dataset (PIDD) dataset. Yadav et al. [13] addressed missing values, outliers, and class imbalance using mean imputation, ADASYN, feature selection, and MLP classifiers, reporting improved diagnostic accuracy. Cost-sensitive learning approaches have also been explored to mitigate class imbalance. Mienye and Sun [14] demonstrated that modifying objective functions to account for imbalance yields better performance than traditional classifiers across multiple medical datasets, including Chronic Kidney Disease Dataset (CKD).

Ensemble and hybrid learning strategies have further contributed to addressing imbalanced healthcare datasets. Alruily et al. [15] proposed an ensemble framework combining RF, XGBoost, and LightGBM, supported by KNN-based imputation and SMOTE sampling, achieving high predictive performance on stroke datasets. Jing [16] investigated focal loss functions integrated with SMOTE and PCA-k-means clustering to enhance minority class detection, while Rana et al. [17] combined KNN imputation with SMOTE and Tomek links to improve stroke risk prediction using Artificial Neural Networks (ANN) models. Other studies employed under-sampling [18], advanced deep learning-based imputation models such as datawig [19], and hybrid feature selection strategies alongside SMOTE to enhance CKD prediction performance [20,21].

Recent research has also emphasized advanced feature selection and deep learning architecture. Shams et al. [22] proposed an RFE-GRU-based model for diabetes diagnosis, achieving strong performance across multiple evaluation metrics. Rahman et al. [23] introduced an ensemble framework integrating MICE-based imputation, Borderline SVM-SMOTE, and optimized feature selection techniques, reporting near-perfect classification results for CKD detection. In addition, Eram et al. [24] highlighted the importance of recall-oriented models in stroke prediction, demonstrating that prioritizing sensitivity can significantly reduce false negatives in critical medical applications.

Although existing studies have reported promising results in handling missing values and class imbalance in healthcare datasets, most approaches address these challenges independently or rely on conventional imputation and sampling techniques without joint optimization. Furthermore, heuristic optimization methods such as Genetic Algorithm and Particle Swarm Optimization are rarely integrated simultaneously for both missing value imputation and class balancing. This limitation underscores the need for a unified preprocessing framework capable of jointly optimizing these two critical challenges. Motivated by this gap, the present study proposes a hybrid preprocessing system that integrates a GA-enhanced MICE-based imputation method with an optimization-driven sampling strategy (GASMOTESPO_ENN) to improve classification performance in healthcare datasets.

3. Proposed Methodology

In this section, a hybrid prediction system shown in Figure 1 is developed to address the classification challenge of medical patients resulting from predicted missing values and imbalanced distributed classes.

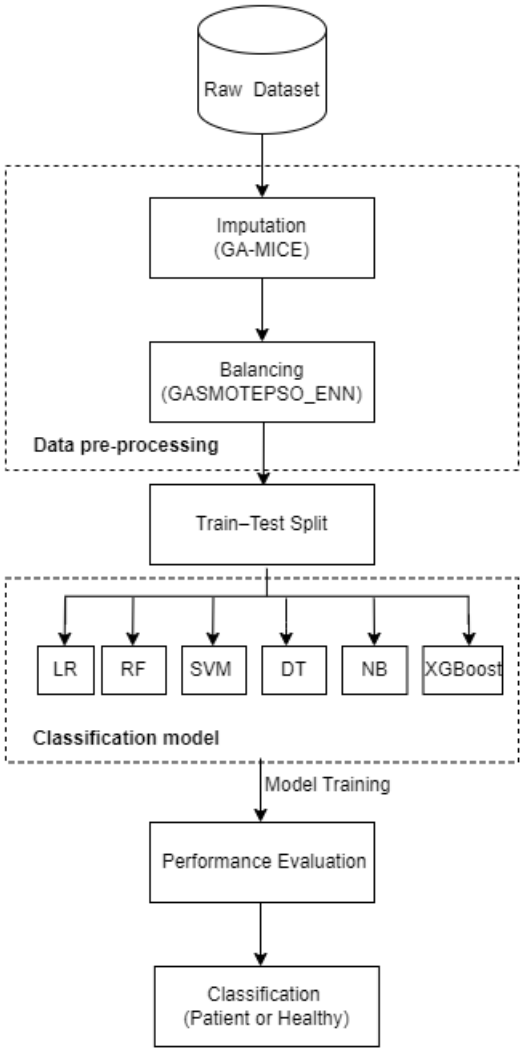


Figure 1.The proposed system's structural framework.

The model consists of three stages: imputation of missing values, balancing imbalanced class distributions, and classifying individuals as patient or healthy employing

machine learning models. The details of the proposed method are outlined in the sections that follow.

3.1. A Novel Missing Value Imputation Method

This section introduces a novel GA–MICE approach that incorporates GA heuristics to enhance the MICE technique for addressing missing data. Algorithm 1 presents the pseudo-code of the proposed GA–MICE imputation method.

Algorithm 1. GA-MICE missing value imputation method.

Input:
- Dataset $D$with variables $\{V_1, V_2, \dots, V_n\}$containing missing values
- Maximum number of iterations max_iter
- Convergence threshold ϵ

Output:
- Imputed dataset D^*

1. Initialize iteration counter: iter ← 0
2. Initialize missing values in Dusing a simple imputation method (e.g., mean or median imputation)
3. **Repeat**
   a. For each variable $V_i \in D$with missing values:
      i. Define $V_i$as the target variable
      ii. Define predictor variables $U = \{V_j \in D : j \neq i$ and V_j has no missing values$\}$
      iii. Fit a regression model $M_i$using complete cases, predicting $V_i$from U
      iv. Predict and update missing values in $V_i$using model M_i
   b. Calculate the maximum absolute difference between imputed values in the current and previous iteration:
      $\text{diff} \leftarrow \max \mid D^{(\text{iter})} - D^{(\text{iter}-1)} \mid$
   c. Increment iteration counter: iter ← iter + 1
4. **Until** diff $< \epsilon$or iter $\geq$ max_iter
5. Return imputed dataset $D^* \leftarrow D^{(\text{iter})}$

The GA–MICE algorithm is initiated by calculating the missing value rates for all variables, after which the variables are sorted in ascending order according to their proportion of missingness. Prior to the iterative MICE procedure, missing values are initialized using a genetic algorithm with fixed parameter settings. In this stage, the GA generates an initial population of 100 randomly initialized individuals, where each individual represents a candidate solution for imputing missing values. The population evolves over a maximum of 100 generations, during which each individual is evaluated using a fitness function. The fitness function assesses each candidate solution by replacing the missing values in the dataset with the values proposed by the individual and computing the sum of all elements in the resulting completed dataset. This objective function is defined as:

$$Fitness(x) = \sum_{i=1}^{n}\sum_{j=1}^{m} \widehat{D}_{i,j} \quad (1)$$

where $\widehat{D}$ denotes the imputed dataset obtained by substituting the missing values in the original dataset D with the corresponding entries from the solution vector x. The GA aims to maximize this fitness value throughout the evolutionary process. During each generation, offspring are produced using a two-point crossover operator with a crossover probability of 0.8, facilitating effective recombination of high-quality solutions and global exploration of the search space. To maintain population diversity and prevent premature convergence, mutation is applied using a Gaussian mutation operator with zero mean and unit variance, controlled by a mutation probability of 0.2 and an independent per-gene mutation probability of 0.1. Parent selection is carried out using tournament selection with a tournament size of three, ensuring a balanced trade-off between selection pressure and diversity. To ensure reproducibility of the experimental results, all GA–MICE experiments were conducted using a fixed random seed (random seed = 42). This evolutionary cycle continues until the predefined maximum number of generations is reached. At the conclusion of the GA optimization, the best-performing individual according to the fitness function is selected, and its genetic representation is used to initialize the missing values before entering the MICE framework. The imputation process then proceeds iteratively using MICE. In each iteration, a target variable V with missing values is selected, and its missing entries are imputed using regression models trained on the remaining fully observed predictor variables U, excluding V itself to avoid circular dependency. Separate regression models are constructed for each variable, allowing independent estimation of missing values based on inter-variable relationships.

Convergence of the imputation process is assessed by computing the maximum absolute difference between imputed values in successive iterations. When this difference falls below the predefined threshold ε, or when the maximum number of iterations max_{iter} is reached, the algorithm terminates, indicating convergence to a stable imputation solution. By integrating GA-based optimization into the initialization stage of the MICE procedure, the proposed GA–MICE method effectively explores the solution space and improves imputation accuracy. Consequently, the GA–MICE algorithm provides a robust and reliable approach for missing value imputation, particularly well-suited for large and complex datasets where accurate data completion is essential.

3.2. Data Sampling Method

In this section, a heuristic-based hybrid sampling method, termed GASMOTEPSO_ENN, was employed

to address class imbalance in the datasets. This approach effectively integrates SMOTE oversampling and Edited Nearest Neighbor (ENN) under sampling techniques, leveraging PSO and GA heuristics to achieve a balanced class distribution. The GASMOTE algorithm was applied with the following parameters: $k = 3$ nearest neighbors, GA population size of 50, 20 generations, crossover rate of 0.8, mutation rate of 0.1, and a maximum neighbor parameter (maxn) of 2, using a fixed random seed of 42 for reproducibility. Subsequently, the synthetic samples generated were refined using the SMOTE_PSO algorithm with a swarm size (n_{pop}) of 10, inertia weight (w) of 1.0, cognitive and social coefficients ($c_1 = c_2 = 2.0$), neighborhood size ($k = 3$), convergence threshold ($\varepsilon = 0.01$), and 15 iterations, executed on a single CPU core ($n_{jobs} = 1$). This was followed by ENN undersampling with 10 nearest neighbors and an 'all' selection criterion to eliminate noisy or ambiguous samples. This sequential preprocessing pipeline—comprising advanced oversampling, heuristic optimization, and noise reduction—yielded balanced and cleaned datasets, thereby enhancing the robustness and accuracy of subsequent classification tasks. The GASMOTEPSO_ENN method [5] can be summarized in the following steps:

Step 1: The GASMOTE algorithm, as proposed in [25], was employed to generate balanced data by combining GA heuristics with SMOTE oversampling. The GA optimizes sampling rates for each minority class instance, where each individual in the population encodes these sampling rates. The population of size *P* undergoes genetic operations including selection, crossover, and mutation, with fitness evaluations guiding the evolution toward optimal sampling strategies. Upon termination, the optimized sampling rates are used to generate synthetic minority class samples via SMOTE, producing a balanced dataset.

Step 2: The SMOTE_PSO algorithm, introduced in [26], was applied to further enhance classifier performance and dataset balance. This method integrates SMOTE with PSO by training an SVM on the processed GASMOTE data to identify support vectors and construct an optimal separating hyperplane. Synthetic samples are then generated within the margin region of the SVM, with PSO optimizing the placement of these samples to improve classifier generalization. Following balancing, outliers were detected using the Interquartile Range (IQR) technique and imputed using the GA_MICE method to maintain data integrity.

Step 3: Finally, the GASMOTEPSO_ENN hybrid approach combined oversampling strategies with ENN under sampling, which removes samples misclassified by their nearest neighbors. This step eliminates noisy and irrelevant data points, further refining the dataset and improving the overall quality of the training data.

The GASMOTEPSO_ENN approach effectively balances class distributions by synergistically combining oversampling and under sampling techniques supported by heuristic optimizations. This integrated methodology enhances the predictive performance and reliability of machine learning classifiers while preserving the structural integrity of the dataset.

3.3. Classification

In this study, various machine learning classification algorithms are employed to distinguish between healthy individuals and patients in healthcare datasets. Logistic Regression (LR) models the probability of class memberships through a linear combination of input features, offering simplicity and interpretability in both binary and multi-class settings [27, 28]. RF, an ensemble method based on multiple decision trees, enhances prediction accuracy and robustness while mitigating overfitting [29, 30]. Support Vector Machine (SVM) identifies optimal hyperplanes for classification, particularly effective in handling non-linear boundaries [31, 32]. Decision Tree (DT) methods provide a clear, interpretable structure for decision-making but may suffer from overfitting in deep trees [6, 33]. Naive Bayes (NB), grounded in Bayes' theorem, is computationally efficient and effective for high-dimensional data [34]. Extreme Gradient Boosting (XGBoost), a powerful boosting technique, improves prediction by sequentially correcting errors from previous models while incorporating regularization to prevent overfitting [35, 36]. These diverse models were selected for their complementary strengths and proven effectiveness in healthcare-related machine learning tasks.

4. Experimental Framework

The experimental framework is carefully designed to systematically analyze the impact of the proposed preprocessing system on binary classification tasks containing health datasets specified by imbalanced missing values. All experiments were performed on the same hardware configuration: an AMD64 processor with 8 physical cores, 16 logical threads, and a maximum clock speed of 1801 MHz, running Windows 10 (Version 10.0.26200). The experimental design includes thorough descriptions of health datasets, classifier parameters, and performance measures under separate subsections.

4.1. Dataset Description

In this section, the proposed preprocessing system is evaluated using three publicly available healthcare datasets obtained from the Kaggle platform: PIDD, the Cerebral Stroke Prediction (CSP) dataset, and CKD dataset [37]. A detailed description of each dataset is provided below.

The PIDD dataset [38] consists exclusively of female individuals aged 21 years and above. It contains a total of

768 instances, including 268 diabetics and 500 non-diabetic samples. The dataset comprises several clinical and physiological attributes commonly used for early diabetes diagnosis, such as number of pregnancies, body mass index (BMI), insulin concentration, age, blood pressure, skin thickness, glucose level, and the Diabetes Pedigree Function. Notably, some attributes contain zero values that do not represent valid physiological measurements and are treated as missing values in this study. The proportion of missing values varies across attributes, accounting for 0.6% in glucose, 4.5% in blood pressure, 29.5% in skin thickness, 48.6% in insulin, and 1.4% in BMI.

The CSP dataset [39] consists of 5,110 instances described by 12 attributes and exhibits a highly imbalanced class distribution. Stroke cases constitute only 4.8% (249 samples) of the dataset, while the remaining instances represent non-stroke cases. This dataset is designed to predict the likelihood of stroke occurrence based on demographic, clinical, and lifestyle-related features, including gender, age, hypertension, heart disease, marital status, occupation type, residence type, smoking status, BMI, average glucose level, and stroke occurrence. Missing values are primarily observed in the smoking status attribute (approximately 30%) and the BMI attribute (approximately 3.9%).

The CKD dataset [40] is used to predict chronic kidney disease based on clinical and laboratory measurements. The dataset contains 400 samples, including 250 CKD patients and 150 healthy individuals, described by 25 predictive attributes and one class label. These attributes include age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cells, pus cell clumps, bacteria presence, random blood glucose, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, and comorbid conditions such as hypertension, diabetes mellitus, and coronary artery disease. The dataset is characterized by a substantial proportion of missing values across multiple attributes, making it particularly suitable for evaluating the effectiveness of the proposed imputation framework.

4.2. Classifier Setting

In machine learning, the classifier setting refers to the systematic configuration of a classification model, including algorithm selection, data preprocessing, hyperparameter optimization, and performance evaluation. An appropriate classifier configuration is critical for building robust and reliable predictive models, particularly in healthcare applications where data quality and class distribution significantly affect outcomes.

All datasets were processed using the proposed preprocessing framework (Figure 1). Missing values were imputed using GA-MICE, followed by class imbalance correction via GASMOTEPSO_ENN. Noise and outliers were handled within the pipeline to improve data quality, and datasets were split into training and testing sets with a 33% test ratio.

Classification experiments were conducted on three datasets using six machine learning classifiers to distinguish patient and healthy instances. Hyperparameter tuning was performed with GridSearchCV and cross-validation, while various scaling methods were evaluated to identify the optimal configuration for each classifier. The search spaces are summarized in Table 1.

Table 1. Hyperparameter search spaces used for classifier optimization via GridSearchCV.

Classifier	*Hyperparameters*
LR	scaler: [StandardScaler, MinMaxScaler, Normalizer, MaxAbsScaler] penalty: ['l1', 'l2', 'elasticnet', 'none'] C: [0.001, 0.01, 0.1, 1.0, 10, 100] class_weight: ['dict', 'balanced', None] random_state: [42, 50, 100] solver: ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'] max_iter: [10, 50, 100]
RF	scaler: [StandardScaler, MinMaxScaler, Normalizer, MaxAbsScaler] n_estimators: [10, 30, 50, 100] criterion: ['gini', 'entropy'] max_features: ['auto', 'sqrt', 'log2'] random_state: [42, 50, 100] class_weight: ['balanced', 'balanced_subsample', 'dict', None] bootstrap: [0, 1] max_depth: [2, 3]
SVM	scaler: [StandardScaler, MinMaxScaler, Normalizer, MaxAbsScaler] C: [0.1, 1, 10, 100, 1000] gamma: ['scale', 'auto', 1, 0.1, 0.01, 0.001, 0.0001] random_state: [42, 50, 100] kernel: ['linear', 'rbf'] class_weight: ['dict', 'balanced', None] decision_function_shape: ['ovo', 'ovr'] degree: [2, 3, 5]
DT	scaler: [StandardScaler, MinMaxScaler, Normalizer, MaxAbsScaler] criterion: ['gini', 'entropy'] splitter: ['best', 'random'] max_features: ['auto', 'sqrt', 'log2', None] random_state: [42, 50, 100] class_weight: ['dict', 'balanced', None] ccp_alpha: [0.0, 0.005, 0.010, 0.015, 0.030, 0.035] max_depth: [1, 2, 3, 4, 5, 6, 7, None]
NB	scaler: [StandardScaler, MinMaxScaler, Normalizer, MaxAbsScaler]
XGBoost	scaler: [StandardScaler, MinMaxScaler, Normalizer, MaxAbsScaler] max_depth: [2, 3, 5] n_estimators: [10, 50, 100] learning_rate: [0.1, 0.01, 0.05] colsample_bytree: [0.5, 1, 2] reg_alpha: [0, 0.5, 1, 5] reg_lambda: [0, 0.5, 1, 5]

The integration of exhaustive hyperparameter tuning and cross-validation facilitated the selection of robust and generalizable classification models. Predictive performance was systematically evaluated and compared using standart performance metrics.

4.3. Performance Evaluation

Evaluation metrics are essential for assessing the performance and effectiveness of machine learning models. The metrics are calculated using the confusion matrix presented in Table 2. The confusion matrix is a thorough evaluation of a model's effectiveness that divides expected and actual outcomes into four categories: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Table 2. Confusion matrix.

	Predicted Positive	***Predicted Negative***
Actual Positive	TP	FN
Actual Negative	FP	TN

Important model evaluation metrics include accuracy, precision, recall, specificity, F1-score, and the Matthews Correlation Coefficient (MCC) with their relational expressions provided in Eqs. (2-6). Accuracy is the ratio of properly identified samples to the overall number of samples. Precision evaluates the proportion of true positive predictions among all positive predictions. Recall, commonly referred to as TPR, assesses the model's ability to accurately recognize positive cases by computing the true positive ratio compared to all true positives. F1-score, calculated as the harmonic mean of precision and recall, offers a balanced performance measure, particularly beneficial for imbalanced datasets. MCC is a composite metric that balances model performance by combining true positives, true negatives, false positives, and false negatives. Moreover, these metrics give precise evaluation and comparison of model performance across various experimental scenarios.

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

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

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

$$F1 - Score\ (F1) = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall} \quad (5)$$

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (6)$$

5. Results and Discussion

In this study, the classification performance of patients and healthy individuals was evaluated using six machine learning models on datasets processed with the proposed preprocessing framework. Classifier performance was assessed using the evaluation metrics, and the results are presented in Tables 3–5, with the highest F1-scores highlighted in bold. All results are reported as mean ± standard deviation over multiple independent runs.

The computational efficiency of the proposed MICE-GA imputation method was also evaluated on the PIDD, CKD, and CSP datasets. The recorded imputation times of 0.54 s, 0.60 s, and 0.75 s, respectively, demonstrate that the proposed approach achieves efficient and scalable performance with low computational overhead across datasets of varying sizes and complexity.

Table 3. Classifier performance analysis on the PIDD dataset (mean ± standard deviation).

Algorithm	***Acc.***	***Pre.***	***Rec.***	***F1***	***MCC***
LR	0.78 ± 0.03	0.74 ± 0.04	0.86 ± 0.02	0.79 ± 0.03	0.57 ± 0.05
RF	0.84 ± 0.02	0.77 ± 0.03	0.96 ± 0.01	0.85 ± 0.02	0.70 ± 0.04
SVM	**0.93 ± 0.01**	**0.92 ± 0.01**	**0.94 ± 0.01**	**0.93 ± 0.01**	**0.86 ± 0.02**
DT	0.85 ± 0.03	0.82 ± 0.04	0.88 ± 0.02	0.85 ± 0.03	0.70 ± 0.05
NB	0.73 ± 0.04	0.73 ± 0.03	0.73 ± 0.04	0.73 ± 0.04	0.47 ± 0.06
XGBoost	0.92 ± 0.01	0.92 ± 0.01	0.89 ± 0.02	0.96 ± 0.01	0.84 ± 0.02

Table 4. Classifier performance analysis on the CSP dataset (mean ± standard deviation).

Algorithm	***Acc.***	***Pre.***	***Rec.***	***F1***	***MCC***
LR	0.92 ± 0.02	0.92 ± 0.02	0.90 ± 0.03	0.91 ± 0.02	0.84 ± 0.03
RF	0.92 ± 0.02	0.91 ± 0.02	0.93 ± 0.02	0.92 ± 0.02	0.85 ± 0.03
SVM	0.96 ± 0.01	0.95 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.92 ± 0.02
DT	0.94 ± 0.02	0.93 ± 0.02	0.95 ± 0.02	0.94 ± 0.02	0.88 ± 0.03
NB	0.92 ± 0.03	0.91 ± 0.03	0.92 ± 0.03	0.92 ± 0.03	0.84 ± 0.04
XGBoost	**0.97 ± 0.01**	**0.97 ± 0.01**	**0.98 ± 0.01**	**0.97 ± 0.01**	**0.95 ± 0.01**

Table 5. Classifier performance analysis on the CKD dataset (mean ± standard deviation).

Algorithm	***Acc.***	***Pre.***	***Rec.***	***F1***	***MCC***
LR	0.99 ± 0.01	1.00 ± 0.002	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01
RF	**1.00 ± 0.003**	**1.00 ± 0.002**	**1.00 ± 0.003**	**1.00 ± 0.003**	**1.00 ± 0.003**

Celal Bayar University Journal of Science
Volume 22, Issue 2, 2026, p 225-235
Doi: 10.18466/cbayarfbe.1766229

SVM	**1.00 ± 0.002**	**1.00 ± 0.001**	**1.00 ± 0.002**	**1.00 ± 0.002**	**1.00 ± 0.002**
DT	0.98 ± 0.02	0.95 ± 0.03	1.00 ± 0.002	0.98 ± 0.02	0.95 ± 0.03
NB	0.98 ± 0.01	0.98 ± 0.01	0.99 ± 0.01	0.98 ± 0.01	0.96 ± 0.02
XGBoost	**1.00 ± 0.001**	**1.00 ± 0.001**	**1.00 ± 0.001**	**1.00 ± 0.001**	**1.00 ± 0.001**

The experimental results demonstrate the effectiveness of the proposed preprocessing system in improving the performance of machine learning classifiers across all datasets. On the PIDD dataset, advanced classifiers such as SVM and XGBoost achieved consistently high accuracy, precision, recall, F1-score, and MCC, indicating that the combined application of GA-MICE and GASMOTEPSO_ENN successfully mitigates the effects of missing values and class imbalance. Similar performance patterns were observed on the CSP dataset, where SVM and XGBoost again yielded superior results across multiple evaluation metrics, confirming the robustness of the preprocessing strategy and the effectiveness of the trained classifiers. For the CKD dataset, all evaluated classifiers exhibited excellent predictive performance, reflecting the strong data representation achieved after preprocessing.

Overall, these findings underline the critical role of jointly addressing missing data and class imbalance in enhancing classification performance. To further evaluate the discriminative capability of the classifiers, ROC curves and the corresponding AUC values for the PIDD, CSP, and CKD datasets are presented in Figures 2–4, respectively.

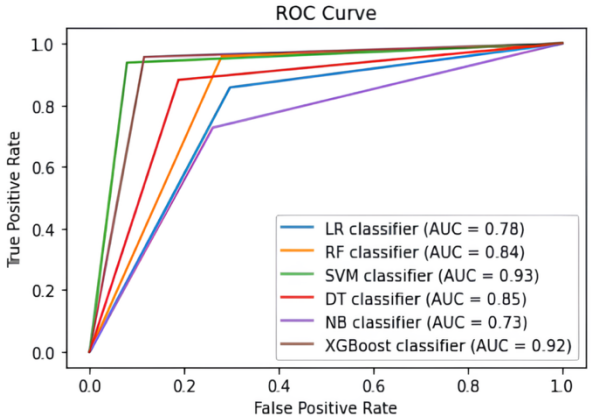


Figure 2. ROC curves for classifiers on the PIDD dataset.

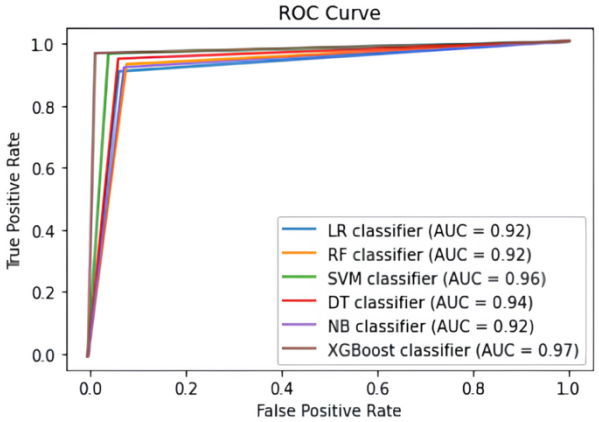


Figure 3. ROC curves for classifiers on the CSP dataset.

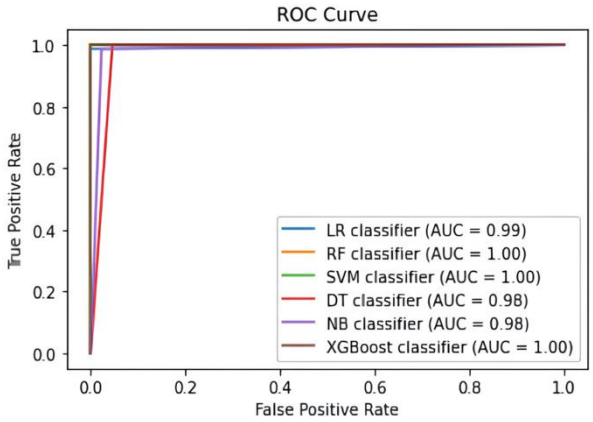


Figure 4. ROC curves for classifiers on the CKD dataset.

As illustrated in Figure 2, the results obtained on the PIDD dataset indicate that SVM and XGBoost exhibit superior discriminative capability between positive and negative classes compared to the other classifiers, whereas NB demonstrates relatively weaker performance. For the CSP dataset, Figure 3 shows that all evaluated classifiers achieve strong discrimination ability, as reflected by their AUC values. Among them, XGBoost attains the highest AUC score of 0.97, highlighting its robustness on this dataset.

The ROC analysis for the CKD dataset, presented in Figure 4, reveals exceptionally high discriminative performance across all classifiers, with AUC values exceeding 0.98. Even comparatively simpler models, such as LR, DT, and NB, demonstrate strong classification capability, while ensemble-based classifiers, including RF, SVM, and XGBoost, achieve AUC values approaching 1.00, indicating near-perfect discrimination. For several classifiers, the proposed framework yields perfect classification performance on the CKD dataset.

The performance of the proposed preprocessing system in the present research was compared to the findings of relevant literature studies dealing with missing value imputation and class imbalance issues. A comparative analysis was conducted based on the model evaluation metrics to evaluate the relative performance of each

classifier. The experimental findings are presented in Tables 6, 7, and 8.

Table 6. Comparison of the PIDD dataset with models in literature.

Reference	Method	Acc.	Pre.	Rec.	F1	AUC
Wang et al. [12]	*DMP MI*	0.87	0.80	0.85	0.83	0.92
Yadav et al. [13]	*ADASYN+ MLP*	0.84	0.91	0.85	0.88	0.83
Mienye and Sun [14]	*CS XGBoost*	0.79	0.77	0.84	0.80	0.83
Madhu et al. [10]	*MissForest +XGBoost*	0.77	-	-	-	-
Shams et al. [22]	*RFE-GRU*	0.91	0.91	0.91	0.90	0.93
This paper	***Proposed system***	**0.93**	**0.92**	**0.94**	**0.93**	**0.93**

Table 7. Comparison of the CSP dataset with models in literature.

Reference	Method	Acc.	Pre.	Rec.	F1	AUC
Alruily et al.[15]	*Ensemble RXLM*	0.96	0.96	0.96	0.96	0.99
Jing [16]	*PCA+DNN+Focal Loss*	0.92	-	-	0.77	0.90
Rana et al.[17]	*SMOTE+TOMEK+DNN*	0.86	0.99	0.87	0.92	0.84
Sailasya and Kumari [18]	*Undersampling +NB*	0.82	0.79	0.86	0.82	0.82
Eram et al. [24]	*SVM+SHAP*	-	-	0.94	-	-
This paper	***Proposed system***	**0.97**	**0.97**	**0.98**	**0.97**	**0.97**

Table 8. Comparison of the CKD dataset with models in literature.

Reference	Method	Acc.	Pre.	Rec.	F1	AUC
Mienye and Sun [14]	*CS RF*	0.99	0.99	1.00	0.99	1.00
Maisha et al. [19]	*Datawig+RF*	0.98	-	-	-	-
Wibowo and Pulupi [20]	*LR+SMOTE*	0.98	1.00	-	0.98	-
Imran et al. [21]	*SMOTE+FNN*	-	0.97	0.99	0.99	0.99
Rahman et al. [23]	LightGBM	1.00	0.99	0.99	1.00	1.00
Rabby et al. [41]		**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Senan et al. [42]	**RF**	**1.00**	**1.00**	**1.00**	**1.00**	**-**
Metun et al. [43]	**RF, Extra Boost**	**1.00**	**1.00**	**1.00**	**1.00**	**-**
This paper	***Proposed system***	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

A comparative evaluation with existing studies on the PIDD dataset, summarized in Table 6, indicates that the proposed system consistently outperforms previously reported approaches. While the DMP_MI method introduced by Wang et al. [12] and the models proposed by Yadav et al. [13] and Mienye and Sun [14] achieved acceptable performance levels, their reported metrics remained below those obtained by the proposed approach. Similarly, although the RFE-GRU model developed by Shams et al. [22] achieved competitive accuracy, it was still surpassed by the proposed system. Overall, the proposed method achieved superior accuracy, precision, recall, F1-score, and AUC values on the PIDD dataset, demonstrating its effectiveness and reliability for diabetes diagnosis.

A similar trend is observed for the CSP dataset, as shown in Table 7. The proposed system achieved higher accuracy and F1-score compared to several state-of-the-art methods, including ensemble-based approaches [15], deep learning-based models [16,17], and sampling-driven frameworks [18]. Although certain studies reported strong performance in specific metrics, such as recall [24], the proposed system provided a more balanced and consistently high performance across all evaluation measures, confirming its robustness for stroke prediction tasks.

The comparative analysis on the CKD dataset, presented in Table 8, further highlights the competitive performance of the proposed system. While several studies reported near-perfect or perfect results using advanced classifiers and preprocessing strategies [14,19–23,41–43], the proposed framework achieved equally optimal scores across all evaluation metrics, including accuracy, precision, recall, F1-score, and AUC. These results demonstrate that the proposed preprocessing and classification strategy is highly effective and performs on par with, or better than, existing state-of-the-art approaches for CKD prediction.

It should be noted that the consistently high performance observed, particularly on the CKD dataset, may be influenced by both intrinsic dataset characteristics—such as strong class separability and low noise—and the adopted preprocessing strategy. Applying certain preprocessing steps prior to the train–test split may introduce limited information leakage, potentially leading to optimistic performance estimates. Nevertheless, the reported results remain consistent with findings in the existing literature and are supported using cross-validation-based hyperparameter optimization, which helps mitigate overfitting.

Overall, the experimental results across all datasets demonstrate that the proposed system can be effectively applied to healthcare datasets characterized by missing values and class imbalance. When evaluated under appropriate and well-defined experimental protocols, the proposed approach exhibits performance that is competitive with, and in many cases superior to, existing methods reported in the literature.

6. Conclusion and Future Work

Imbalanced class distributions and missing data remain two of the most critical challenges limiting the reliability of machine learning models in healthcare applications. These issues can substantially degrade predictive performance, particularly in the diagnosis of rare diseases

and the analysis of incomplete clinical records. Addressing both challenges in an integrated manner is therefore essential for developing robust and trustworthy healthcare decision-support systems.

In this study, a unified preprocessing framework was proposed to simultaneously address missing values and class imbalance in healthcare datasets. Missing values were handled using the GA-MICE imputation method, which enhances the conventional MICE approach through Genetic Algorithm–based optimization, while class imbalance was mitigated using the GASMOTEPSO_ENN hybrid sampling strategy. The effectiveness of the proposed framework was evaluated using six machine learning classifiers across three publicly available healthcare datasets. Performance was assessed using multiple evaluation metrics, including accuracy, precision, recall, F1-score, MCC, and AUC.

The experimental results demonstrate that the proposed preprocessing system significantly improves classification performance by enhancing data quality and mitigating bias introduced by missing and imbalanced data. Across all datasets, the proposed approach achieved competitive or superior results compared to existing methods reported in the literature, confirming its effectiveness and practical value for healthcare data analysis. These findings highlight the importance of integrated preprocessing strategies in improving the reliability and accuracy of machine learning–based healthcare prediction systems.

Despite these promising results, several limitations should be acknowledged. First, the evaluation was conducted on a limited number of healthcare datasets; therefore, the generalizability of the proposed framework to other medical domains and data sources warrant further investigation. Second, preprocessing steps were applied prior to train–test partitioning, which may introduce a degree of information leakage and potentially lead to optimistic performance estimates. Future studies should address this limitation by adopting strict data separation protocols and nested cross-validation schemes, ensuring that all preprocessing and optimization steps are performed exclusively on training data. Additionally, future work will incorporate confidence intervals and formal statistical significance tests to strengthen the robustness of performance comparisons. Finally, the real-world applicability of the proposed framework could be further validated through clinical studies or deployment in practical healthcare settings.

Author's Contributions

Hatice Nizam-Ozogur: Conceptualization of the study, implementation of software and coding, interpretation of results, and manuscript writing.

Zeynep Orman: Guidance on research design, methodological support, interpretation of results, and critical revision of the manuscript.

Ethics

There are no ethical issues after the publication of this manuscript.

References

[1]. García-Laencina, PJ, Sancho-Gómez, JL, et al. 2010. Pattern classification with missing data: a review. Neural Computing and Applications; 19: 263-282. https://doi.org/10.1007/s00521-009-0295-6

[2]. Lin, W-C, Ke, S-W, et al. 2017. When should we ignore examples with missing values? International Journal of Data Warehousing and Mining (IJDWM); 13(4): 53-63. https://doi.org/10.4018/ijdwm.2017100104

[3]. Bertsimas, D, Pawlowski, C, et al. 2018. From predictive methods to missing data imputation: an optimization approach. Journal of Machine Learning Research; 18(196): 1-39. https://www.jmlr.org/papers/v18/17-073.html

[4]. Lin, W-C, Tsai, C-F, et al. 2022. Deep learning for missing value imputation of continuous data and the effect of data discretization. Knowledge-Based Systems; 239: 108079. https://doi.org/10.1016/j.knosys.2021.108079

[5]. Nizam-Ozogur, H, Orman, Z. 2024. A heuristic-based hybrid sampling method using a combination of smote and enn for imbalanced health data. Expert Systems; 41(8): e13596. https://doi.org/10.1111/exsy.13596

[6]. Parhi, SK, Patro, SK. 2023. Compressive strength prediction of pet fiber-reinforced concrete using dolphin echolocation optimized decision tree-based machine learning algorithms. Asian Journal of Civil Engineering; 25(1): 977-996. https://doi.org/10.1007/s42107-023-00826-8

[7]. Zhou, X-H, Eckert, GJ, et al. 2001. Multiple imputation in public health research. Statistics in Medicine; 20(9-10): 1541-1549. https://doi.org/10.1002/sim.689

[8]. Khan, SI, Hoque, ASML. 2020. Sice: an improved missing data imputation technique. Journal of Big Data; 7(1): 37. https://doi.org/10.1186/s40537-020-00313-w

[9]. Raja, PS, Thangavel, K. 2020. Missing value imputation using unsupervised machine learning techniques. Soft Computing; 24(6): 4361-4392. https://doi.org/10.1007/s00500-019-04199-6

[10]. Madhu, G, Bharadwaj, BL, et al. 2019. A novel algorithm for missing data imputation on machine learning. In: 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, pp 173-177. https://doi.org/10.1109/icssit46314.2019.8987895

[11]. Rani, P, Kumar, R, et al. 2021. Hioc: a hybrid imputation method to predict missing values in medical datasets. International Journal of Intelligent Computing and Cybernetics; 14(4): 598-616. https://doi.org/10.1108/ijicc-03-2021-0042

[12]. Wang, Q, Cao, W, et al. 2019. Dmp mi: an effective diabetes mellitus classification algorithm on imbalanced data with missing values. IEEE Access; 7: 102232-102238. https://doi.org/10.1109/access.2019.2929866

[13]. Yadav, S, Maravi, YPS, et al. 2021. A neural network based diabetes prediction on imbalanced data. In: 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT). IEEE, pp 515-521. https://doi.org/10.1109/csnt51715.2021.9509732

[14]. Mienye, ID, Sun, Y. 2021. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. Informatics in Medicine Unlocked; 25: 100690. https://doi.org/10.1016/j.imu.2021.100690

[15]. Alruily, M, El-Ghany, SA, et al. 2023. A-tuning ensemble machine learning technique for cerebral stroke prediction. Applied Sciences; 13(8): 5047. https://doi.org/10.3390/app13085047

[16]. Jing, Y. 2022. Machine learning performance analysis to predict stroke based on imbalanced medical dataset. In: CAIBDA 2022; 2nd International Conference on Artificial Intelligence, Big Data and Algorithms. VDE, pp 1-7. https://doi.org/10.48550/arXiv.2211.07652

[17]. Rana, C, Chitre, N, et al. 2021. Stroke prediction using smote-tomek and neural network. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, pp 1-5. https://doi.org/10.1109/icccnt51525.2021.9579763

[18]. Sailasya, G, Kumari, GLA. 2021. Analyzing the performance of stroke prediction using ml classification algorithms. International Journal of Advanced Computer Science and Applications; 12(6). https://doi.org/10.14569/ijacsa.2021.0120662

[19]. Maisha, SJ, Biswangri, E, et al. 2022. An approach to detect chronic kidney disease (ckd) by removing noisy and inconsistent values of uci dataset. In: Proceedings of the Third International Conference on Trends in Computational and Cognitive Engineering: TCCE 2021. Springer, pp 457-472. https://doi.org/10.1007/978-981-16-7597-3_38

[20]. Wibowo, MR, Palupi, I. 2023. Enhancing accuracy on chronic-kidney disease detection using machine learning with technique of resampling and missing value treatment. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control; 8(3). https://doi.org/10.59287/kinetik.v8i3.1674

[21]. Al Imran, A, Amin, MN, et al. 2018. Classification of chronic kidney disease using logistic regression, feedforward neural network and wide & deep learning. In: 2018 International Conference on Innovation in Engineering and Technology (ICIET). IEEE, pp 1-6. https://doi.org/10.1109/ciet.2018.8660844

[22]. Shams, MY, Tarek, Z, et al. 2025. A novel rfe-gru model for diabetes classification using pima indian dataset. Scientific Reports; 15(1): 982. https://doi.org/10.1038/s41598-024-82420-9

[23]. Rahman, MM, Al-Amin, M, et al. 2024. Machine learning models for chronic kidney disease diagnosis and prediction. Biomedical Signal Processing and Control; 87: 105368. https://doi.org/10.1016/j.bspc.2023.105368

[24]. Eram, AF, Mahmud, AS, et al. 2025. Beyond the numbers: app-enabled stroke prediction system for high-risk individuals in imbalanced datasets. Neuroscience Informatics; 5(3): 100215. https://doi.org/10.1016/j.neuri.2025.100215

[25]. Jiang, K, Lu, J, et al. 2016. A novel algorithm for imbalance data classification based on genetic algorithm improved smote. Arabian Journal for Science and Engineering; 41: 3255-3266. https://doi.org/10.1007/s13369-016-2179-2

[26]. Cervantes, J, Garcia-Lamont, F, et al. 2017. Pso-based method for svm classification on skewed data sets. Neurocomputing; 228: 187-197. https://doi.org/10.1016/j.neucom.2016.10.041

[27]. DeMaris, A. 1995. A tutorial in logistic regression. Journal of Marriage and the Family; 57(4): 956-968. https://doi.org/10.2307/353415

[28]. Dunteman, GH, Ho, M-RH. 2006. An Introduction to Generalized Linear Models. Sage, Thousand Oaks. https://books.google.com.tr/books/about/An_Introduction_to_Generalized_Linear_Mo.html?id=iHvg0QEACAAJ&redir_esc=y

[29]. Biau, G, Scornet, E. 2016. A random forest guided tour. Test; 25(2): 197-227. https://doi.org/10.1007/s11749-016-0481-7

[30]. Breiman, L. 2001. Random forests. Machine Learning; 45(1): 5-32. https://doi.org/10.1023/a:1010933404324

[31]. Ghosh, S, Dasgupta, A, et al. 2019. A study on support vector machine based linear and non-linear pattern classification. In: 2019 International Conference on Intelligent Sustainable Systems (ICISS). IEEE, pp 24-28. https://doi.org/10.1109/iss1.2019.8908018

[32]. Cheng, W-C, Bai, X-D, et al. 2020. Identifying characteristics of pipejacking parameters to assess geological conditions using optimisation algorithm-based support vector machines. Tunnelling and Underground Space Technology; 106: 103592. https://doi.org/10.1016/j.tust.2020.103592

[33]. Costa, VG, Pedreira, CE. 2023. Recent advances in decision trees: an updated survey. Artificial Intelligence Review; 56(5): 4765-4800. https://doi.org/10.1007/s10462-022-10275-5

[34]. Wickramasinghe, I, Kalutarage, H. 2021. Naive bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Computing; 25(3): 2277-2293. https://doi.org/10.1007/s00500-020-05297-6

[35]. Nguyen, TT, Pham, TD, et al. 2022. A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and sar data fusion. Science of the Total Environment; 804: 150187. https://doi.org/10.1016/j.scitotenv.2021.150187

[36]. Asselman, A, Khaldi, M, et al. 2021. Enhancing the prediction of student performance based on the machine learning xgboost algorithm. Interactive Learning Environments; 31(6): 3360-3379. https://doi.org/10.1080/10494820.2021.1928235

[37]. Kaggle. Kaggle dataset. https://www.kaggle.com/datasets (accessed 16.01.2024).

[38]. UCI Machine Learning Repository. Pima indians diabetes dataset. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (accessed 19.09.2023).

[39]. Kaggle. Stroke prediction dataset. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (accessed 19.09.2023).

[40]. Kaggle. Chronic kidney disease dataset. https://www.kaggle.com/datasets/mansoordaku/ckdisease (accessed 19.09.2023).

[41]. Rabby, AKMSA, Mamata, R, et al. 2019. Machine learning applied to kidney disease prediction: comparison study. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, Kanpur, India, pp 1-7. https://doi.org/10.1109/ICCCNT45670.2019.8944799

[42]. Senan, EM, Al-Adhaileh, MH, et al. 2021. Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. Journal of Healthcare Engineering; 2021(1): 1004767. https://doi.org/10.1155/2021/1004767

[43]. Metun, P, Geethanjali, P, et al. 2024. Predictive analysis in nephrology: evaluating machine learning models for chronic kidney disease prediction. International Journal for Research in Applied Science and Engineering Technology (IJRASET); 12(3). https://doi.org/10.22214/ijraset.2024.58102