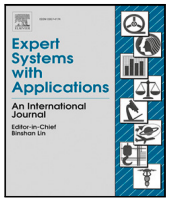
Expert Systems With Applications 255 (2024) 124886

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Enhancing cardiovascular risk assessment with advanced data balancing and domain knowledge-driven explainability

Fan Yang [a,b,c], Yanan Qiao [a], Petr Hajek [d], Mohammad Zoynul Abedin [e,*]

[a] School of Computer Science and Technology, Xi'an Jiaotong University, No. 28, Xianning West Road, Xi'an, Shaanxi, 710049, PR China
[b] Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin 541004, PR China
[c] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin, 130012, PR China
[d] Faculty of Economics and Administration, University of Pardubice, Studentska 95, Pardubice, 53210, Czech Republic
[e] Department of Accounting and Finance, School of Management, Swansea University, Bay Campus, Fabian Way, Swansea SA1 8EN, Wales, United Kingdom

ARTICLE INFO

Keywords:
Heart disease risk
Data balancing
Performance discrepancy
Explainability
Expert system
Domain knowledge

ABSTRACT

In medical risk prediction, such as predicting heart disease, machine learning (ML) classifiers must achieve high accuracy, precision, and recall to minimize the chances of incorrect diagnoses or treatment recommendations. However, real-world datasets often have imbalanced data, which can affect classifier performance. Traditional data balancing methods can lead to overfitting and underfitting, making it difficult to identify potential health risks accurately. Early prediction of heart attacks is of paramount importance, and researchers have developed ML-based systems to address this problem. However, much of the existing ML research is based on a single dataset, often ignoring performance evaluation across multiple datasets. As the demand for interpretable ML models grows, model interpretability becomes central to revealing insights and feature effects within predictive models. To address these challenges, we present a novel data balancing technique that uses a divide-and-conquer strategy with the K-Means clustering algorithm to segment the dataset. The performance of our approach is highlighted through comparisons with established techniques, which demonstrate the superiority of our proposed method. To address the challenge of inter-dataset discrepancies, we use two different datasets. Our holistic pipeline, strengthened by the innovative balancing technique, effectively addresses performance discrepancies, culminating in a significant improvement from 81% to 90%. Furthermore, through advanced statistical analysis, it has been determined that the 95% confidence interval for the AUC metric of our method ranges from 0.8187 to 0.8411. This observation serves to underscore the consistency and reliability of our approach, demonstrating its ability to achieve high performance across a range of scenarios. Incorporating Explainable AI (XAI), we examine the feature rankings and their contributions within the best performing Random Forest model. While the domain expert feedback is consistent with the explanatory power of XAI, some differences remain. Nevertheless, a remarkable convergence in feature ranking and weighting is observed, bridging the insights from XAI tools and domain expert perspectives.

1. Introduction

Classification models in machine learning (ML) often struggle with the conundrum of imbalanced datasets, where instances of the majority class significantly outnumber those of the minority class, hindering the model's learning efficiency during training (Tarawneh, Hassanat, Altarawneh, & Almuhaimeed, 2022). This imbalance becomes critical in scenarios such as disease risk diagnosis, where the contributions of the minority class are crucial (Brito, Chen, Wise, & Mortimore, 2022). Historically, oversampling the minority class or undersampling the majority class have been conventional remedial strategies. However, each approach has inherent drawbacks. Oversampling, which involves random replication of a subset of the minority class, fails to provide new insights (Douzas & Bacao, 2018), while undersampling, which involves random elimination from the majority class, incurs the penalty of data loss. In highly imbalanced scenarios, oversampling can lead to an overabundance of synthetic minority class data, reducing class variance and potentially introducing bias into classification processes (Amin et al., 2016). Conversely, undersampling potentially weakens classifier performance through information erosion, and oversampling occasionally culminates in model overfitting (Park & Park, 2021). Thus, the

* Corresponding author.
E-mail addresses: f.yangcs@xjtu.edu.cn (F. Yang), qiaoyanan@mail.xjtu.edu.cn (Y. Qiao), petr.hajek@upce.cz (P. Hajek), m.z.abedin@swansea.ac.uk (M.Z. Abedin).

https://doi.org/10.1016/j.eswa.2024.124886
Received 29 April 2024; Received in revised form 26 June 2024; Accepted 23 July 2024
Available online 1 August 2024

formulation of an innovative data equalization technique that avoids these entrenched limitations is imperative.

Although existing hybrid oversampling and undersampling techniques attempt to ameliorate imbalance problems, they fail in specific domains such as health informatics, bioinformatics, and biostatistics, where class instances are close and sometimes overlap, sowing seeds of ambiguity during the ML model learning phase and misdirecting classifiers during categorization. Dataset balancing is a powerful preprocessing technique in ML that is used in a variety of fields, especially where high precision and recall for all classes are essential (Ching et al., 2018).

Despite occasionally acceptable accuracy, the pronounced biases in balanced data can lead to unstable classification results, with precision and recall experiencing significant inter-class variation, compromising the consistency of the classifier's performance. To achieve congruent performance in ML classifiers — ensuring parity in precision, recall, and F1 scores across classes — careful dataset balancing is paramount to prevent bias during model training and testing (Liu, Fan, & Wu, 2019).

Artificial Intelligence (AI) has permeated various facets of healthcare, finding applications in predictive medicine, healthcare administration, diagnostics, and clinical decision-making, among others (Ahsan & Siddique, 2022; Marabelli, Vaast, & Li, 2021; Wallace, Mullarkey, & Hevner, 2023). Despite progress in achieving human-like performance, AI models are often underutilized, particularly in medical scenarios, due to their inherent opacity and consequent mistrust among practitioners (Mikalef, Conboy, Lundström, & Popovič, 2022). In response to this concern, Explainable Artificial Intelligence (XAI) has emerged to increase the transparency of model predictions by explaining the logical progression that led to them. This initiative aims to foster an environment conducive to the integration of AI systems into the healthcare industry by increasing user confidence in such technologies (Das, Sultana, Bhattacharya, Sengupta, & De, 2023). Within the XAI framework, the success of AI is measured not only by its predictive accuracy but also, and importantly, by its ability to provide understandable explanations for its conclusions. Improved 'explainability' aims to enable more timely, cost-effective, and contextually appropriate healthcare solutions than are traditionally available, particularly in the hospital environment.

ML algorithms within XAI, powered by rich data, evidence-based learning, validated protocols, and compelling post-action reasoning, navigate clinical pathways. This fosters collaboration and strengthens the doctor-patient relationship—critical elements in delivering high-quality, cost-effective healthcare. In today's environment, traditional ML training and testing paradigms are insufficient to unravel the nuanced narratives embedded in medical informatics and the broader healthcare sector. As a result, XAI tools have gained traction, serving to decipher the impact and significance of features within the performance matrices of ML models (Das et al., 2023). Establishing both local and global explicability of models is paramount to building a robust computational healthcare system. To cultivate trust among end-users and support domain experts in healthcare, ML models must not only be interpretable but also ensure that the impact and contributions of individual features are transparently accessible.

A careful examination of the congruence between domain expert knowledge and AI tools, particularly in healthcare XAI applications, is imperative. This need arises from the critical necessity to validate and cross-validate the functionalities and outcomes of widely used XAI tools. To address this crucial issue, we conducted a survey of healthcare experts and correlated their perspectives with the interpretability of XAI tools to unravel and understand the intrinsic narratives embedded in healthcare dilemmas. This paper revolves around several key contributions, which are outlined below:

- We present a robust data balancing technique strategically designed to regulate the stability of classifier performance, mitigate performance discrepancies, and avoid overfitting and underfitting scenarios. The proposed data balancing technique is based on a divide-and-conquer strategy using the K-means clustering algorithm. The dataset is segmented into multiple clusters, within which we independently apply oversampling and undersampling to balance the class distributions. This approach not only mitigates the drawbacks of traditional methods but also improves classifier performance by preserving class variance and reducing bias, ultimately leading to more accurate and reliable predictions in imbalanced classification scenarios.
- We identify and dissect the problem of data discrepancies and propose a structured pipeline aimed at mitigating this problem, thereby aiming to achieve enhanced accuracy within both single-dataset and inter-dataset frameworks.
- We use XAI techniques to uncover the 'inner story' hidden in black-box ML algorithms, focusing on elucidating the local and global explainability of ML models and assessing the impact of features on classifier performance.
- We synergize domain knowledge with XAI explainability through a detailed survey of domain experts, exploring the relationship between domain knowledge and XAI tool results, while identifying inconsistencies and exploring their potential causal factors.

The subsequent sections of this paper are organized as follows: Section 2 describes related work in the field, while Section 3 explains the proposed methodology along with the experimental setup. The experimental results and their respective analyses are comprehensively presented in Section 4, and the results are discussed in Section 5. Section 6 concludes the paper by providing insights into possible future work.

2. Related work

2.1. Data balancing techniques

The pivotal role of dataset balancing, an effective preprocessing approach in ML, has seen its application in a variety of domains, highlighting its importance in dealing with class imbalance problems. A review of related work provides insights into the different methodologies adopted by researchers and the diverse contexts of their applications.

In a study by Batista, Prati, and Monard (2004), a meticulous comparison was made between ten techniques across thirteen UCI datasets in an attempt to address class imbalance issues. Interestingly, their empirical findings underscored that discrepancies between classes do not always undermine the performance of learning systems. One particular investigation used ML to detect code smells and identified suboptimal performance due to pronounced dataset imbalance characteristics. Despite incorporating SMOTE in the preprocessing phase, the researchers found that data balancing did not significantly improve model performance (Pecorelli, Di Nucci, De Roover, & De Lucia, 2019). Extending this research, the same cohort (Pecorelli, Di Nucci, De Roover, & Lucia, 2020) investigated five different data balancing techniques, assessed their impact on code smell detection in object-oriented systems, and found that omitting the balancing phase did not adversely affect accuracy.

An insightful offering by Lemaître, Nogueira, and Aridas (2017) introduced the "imbalanced-learn API", a Python toolbox tailored for managing imbalanced datasets in ML. The research juxtaposed binary and multiclass data balancing models, traversed different data balancing methodologies, and provided insights into oversampling and undersampling techniques. Another investigation by Nagavelli, Samanta, and Chakraborty (2022) used a hybrid method that combined SMOTE and edited nearest neighbor (ENN) to balance datasets in heart disease prediction. By training ML models using this balancing technique on ECG data, they contrasted the results from balanced and unbalanced datasets, highlighting a significant improvement in classifier performance using the hybrid SMOTE-ENN method (95.9% accuracy achieved for XGBoost), underscoring the importance of data balancing in healthcare scenarios.

In exploring appropriate data balancing techniques for classifying the Cleveland heart disease dataset, researchers used NearMiss, SMOTE, and SMOTETomek, coupling ML with ensemble methods to determine the effectiveness of these balancing techniques (Sahid, Hasan, Akter, & Tareq, 2022). The study highlights that using advanced imbalance data handling techniques like SMOTETomek can significantly improve the accuracy of heart disease prediction models up to 96% (Sahid et al., 2022). In another study, in the midst of a highly imbalanced dataset for stroke prediction in an elderly Chinese population, SMOTE was implemented during preprocessing and a significant increase in classifier performance was observed in terms of AUC (0.78 for random forest), ensuring consistent and reasonably accurate results (Wu & Fang, 2020).

As noted above, SMOTE is a widely used method for handling class imbalance in medical datasets. However, it has notable limitations when applied to medical data, including the potential introduction of noise, increased computational complexity, persistence of imbalance in highly skewed datasets, and challenges related to the quality and interpretability of synthetic samples. Addressing these limitations often requires combining SMOTE with other techniques or using advanced methods tailored to the specific characteristics of medical datasets (Woźniak, Wieczorek, & Siłka, 2023).

Most recently, the RLMD-PA (reinforcement learning-based myocarditis diagnosis combined with population-based algorithm) model offers a robust approach to myocarditis diagnosis, leveraging reinforcement learning and a population-based algorithm for effective and accurate classification (with a mean accuracy of 88.6%) (Moravvej et al., 2022). The model formulates the classification problem as a sequential decision-making process, which allows for continuous learning and adjustment based on rewards, enhancing the model's adaptability to new data. The model effectively addresses the issue of class imbalance by giving greater rewards for correctly classifying minority class samples. However, the complexity of the model can make it challenging to interpret and understand the decision-making process.

In a context where deep learning was applied to medical data, a specific study (Zhang, Zhang, Pirbhulal, Wu, & Albuquerque, 2020) applied data balancing techniques to ECG data and proposed the ABM (active balancing mechanism) data balancing technique. The approach used the Gaussian Naive Bayes algorithm to estimate the object sample, using entropy as a query function to evaluate the results. ABM achieved 92.23% accuracy with support vector machines and 97.52% with a modified convolutional neural network. Active deep learning models have also significantly enhanced the precision of medical image segmentation and classification, aiding in accurate diagnosis exceeding 90% and treatment planning (Mahmood, Rehman, Saba, Nadeem, & Bahaj, 2023). By utilizing active learning techniques, these models reduce the annotation burden, selectively choosing the most informative samples for training. A limitation is that the algorithmic bias of active learning methods can result in an increased number of false positives and negatives for minority classes, reducing the overall effectiveness of the model (Mahmood et al., 2023).

2.2. Heart disease risk prediction using machine learning

Cardiovascular diseases (CVDs), recognized as a major global health burden, are responsible for a substantial proportion of deaths worldwide and fundamentally alter cardiac and vascular function (Azmi et al., 2022; Jiang et al., 2022). The World Health Organization (WHO) states that CVDs cause approximately 17.9 million deaths annually, accounting for approximately 32% of global deaths (Tarawneh et al., 2022). In particular, heart attacks and strokes cumulatively account for a staggering 85% of these deaths (Douzas & Bacao, 2018), with determinants such as unhealthy lifestyles, obesity, hypercholesterolemia, and diabetes serving as precipitating factors (Rajkumar, Devi, & Srinivasan, 2022). Amidst the spectrum of sometimes confusing signs of aging, making a definitive diagnosis becomes a complex endeavor. Given the critical nature of heart disease, early detection is emerging as a potentially effective strategy to mitigate associated mortality. Diagnostic modalities such as ECG and coronary angiography (CA) are conventionally used; however, both have inherent limitations—CA is associated with significant costs, while ECG can intermittently fail to detect symptomatic manifestations of heart disease (Park & Park, 2021).

Navigating the complexities of heart disease diagnosis requires acute precision, necessitating the fusion of data derived from multiple sensors to enhance the accuracy of the dataset (Uddin, Rashid, Hasan, Hossain, & Fang, 2022). ML represents a powerful tool to increase diagnostic accuracy, using available and real-time data sets for accurate disease detection (Cutrì et al., 2017). The integration of computational technologies into diagnostic procedures has experienced an upsurge, simultaneously increasing the volume of medical data and underscoring ML as an indispensable diagnostic tool in modern healthcare. ML is useful in scenarios where large amounts of data require rigorous analysis and discrimination, such as interpreting genetic data, predicting pandemics, and transforming medical data into actionable knowledge (Sarumi & Leung, 2022; Tiwari, Bhati, Al-Turjman, & Nagpal, 2022; Weissler et al., 2021). A plethora of research efforts across disciplines have used datasets from the UCI ML repository to predict cardiac disease, but few investigations have addressed the key issue of inter-dataset discrepancies when using multiple datasets (Lin, Mak, Li, & Chien, 2018).

The study by Alshraideh et al. (2024) demonstrated that ML techniques significantly enhance the accuracy of heart attack predictions by analyzing a variety of risk factors, including high blood pressure, cholesterol levels, irregular pulse rates, and diabetes. The study demonstrated superior predictive performance of SVM with an accuracy rate of 94.3%, outperforming other machine learning techniques tested. However, even when enhanced with an effective feature selection method, SVM struggles with imbalanced datasets, a common issue in medical data. Similarly, Dalal et al. (2023) demonstrated a significant improvement in prediction accuracy for cardiovascular disease risk using ML models. The best-performing ensemble learning models achieved an accuracy of 99.1%, which is superior to many traditional methods. However, the lack of model transparency can be a barrier in clinical settings.

To shape a real-time predictive system, it is imperative to develop an ML model fed by diverse data that embodies versatility and generality for assimilating novel input sensor data directly from human subjects. Therefore, the path to constructing a globally applicable model from existing heart disease prediction datasets requires mitigation of emerging inter-dataset discrepancy issues. Following the resolution of such issues, ML models trained on diverse datasets can be molded to exhibit flexibility towards real-time multi-sensor data during predictive analysis.

2.3. Explainable artificial intelligence in healthcare decision making

Srinivasu, Sandhya, Jhaveri, and Raut (2022) explored the emerging trajectory of creating XAI systems in the healthcare sector, underscored by the strategic use of techniques such as attention mechanisms and surrogate models. Achieving XAI is fundamentally rooted in facilitating a full human understanding of the decision-making processes of AI models. The authors elucidate a range of strategies driven by XAI in healthcare, including both regional and global post hoc explainability toolkits, as well as explainability tools focused on the rational, data, and performance dimensions. They further articulate the prospective horizon of XAI in healthcare and highlight its potential dividends in enhancing research cognizance within the sector.

In a parallel vein, Dave, Naik, Singhal, and Patel (2020) illuminate several interpretability techniques, emphasizing the imperative that if AI fails to elucidate its predictions — particularly within the healthcare sector — it could potentially create more dilemmas than

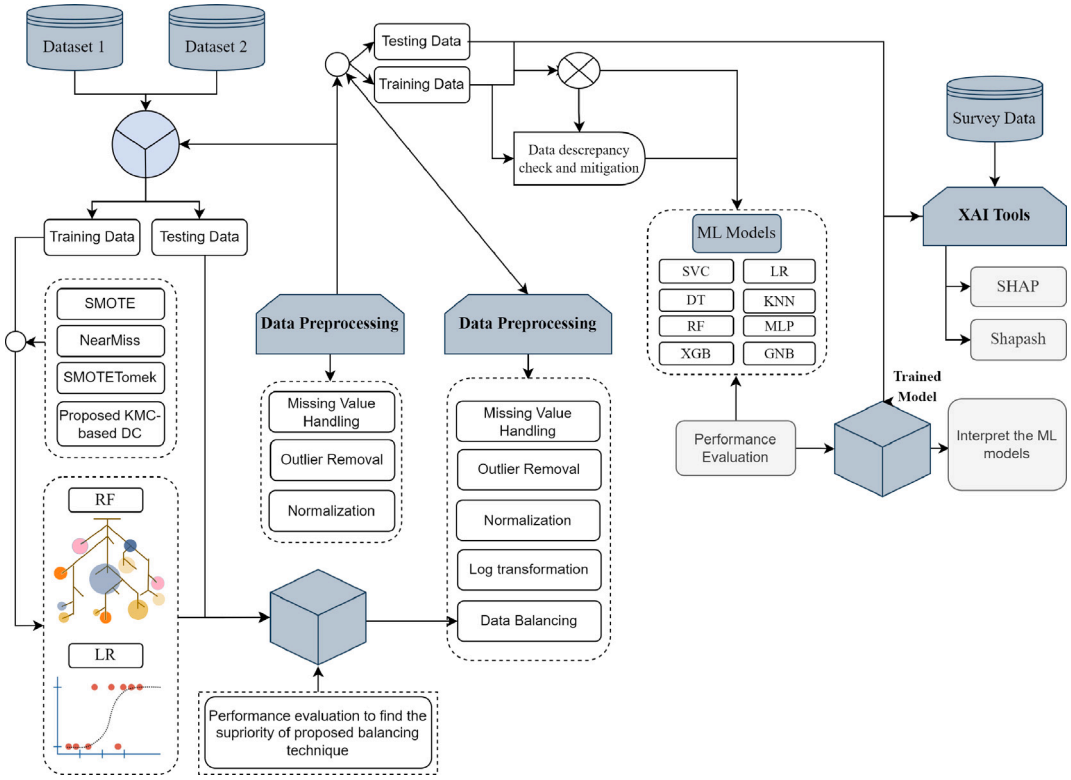


Fig. 1. An overview of the proposed methodology for predicting heart disease, including model development and XAI explanatory power.

solutions. Their experiments, conducted using the Cleveland Heart Disease dataset, showed that variables such as *ca*, *oldpeak*, and *thalach* were key contributors to the onset of heart disease. In contrast, Guleria, Naga Srinivasu, Ahmed, Almusallam, and Alarfaj (2022), using the same dataset but experimenting with different algorithms, found different results from their model interpretability tools. Their research showed that variables such as *sex*, *trestbps*, and *cp* had a significant impact on the manifestation of heart disease.

The findings of the above literature suggest that establishing both local and global explicability of models is crucial but challenging (Allgaier, Mulansky, Draelos, & Pryss, 2023; Dhar, Dey, Borra, & Sherratt, 2023). It requires ensuring that the impact and contributions of individual features are transparently accessible to cultivate trust among end-users and support domain experts in healthcare. There is a critical need to validate and cross-validate the functionalities and outcomes of widely used XAI tools through a survey of healthcare experts.

3. Methods

3.1. Approach overview

In the course of this work, certain datasets are selectively identified as representative instances of imbalanced data samples. Our proposed methodology passes through several stages, starting with data preprocessing, where each dataset is individually subjected to data balancing techniques. Subsequently, the balanced datasets are integrated into ML algorithms, specifically logistic regression (LR) and random forest (RF) classifiers, as visually depicted in Fig. 1.

The performance of each algorithmic ensemble is evaluated using metrics including accuracy, precision, recall, and F1 score, supplemented by the presentation of Receiver Operating Characteristic (ROC) curves. Furthermore, the methodology aims to reconcile inter-dataset discrepancies in heart disease risk prediction using two publicly recognized datasets. The stratagem is orchestrated into four main segments: (i) data preprocessing, (ii) ML classifier construction, (iii) model evaluation, and (iv) model explainability.

At the outset, the statistical properties of the discrete datasets are examined, after which various preprocessing techniques are employed to meticulously match the datasets to the requirements of the ML classifiers. A congruent dataset structure is established by standardizing the columns based on the Hungarian dataset. The performance metrics of each classifier through discrete stages are tabulated and visually manifested through various plots, shedding light on inter-dataset discrepancies and illustrating the effectiveness of our proposed methodology in overcoming these challenges.

In the area of model explainability, a ranking of features is performed, and their respective contributions to classifier performance are graphically illustrated. In addition, a survey was conducted to validate the accuracy of the explanations offered by XAI tools, and subsequently juxtaposed with insights extrapolated from domain experts, providing a holistic validation of XAI results.

In essence, our methodology provides a systematic way to navigate unbalanced datasets, improve classifier performance, and reduce inter-dataset discrepancies, thereby strengthening the robustness and reliability of heart disease risk prediction models.

3.2. Datasets description

In the context of this investigation, two main datasets are used, namely the Long Beach Veterans Affairs (VA) and the Hungarian heart disease datasets (Alizadehsani et al., 2019), which serve as the basic data sources for our experiments. Both datasets have identical characteristics and are binary labeled. They include 13 standard features (age, sex, chest pain type, cholesterol, resting blood pressure, fasting

Table 1
Common features of the primary datasets.

Feature name	Feature type	Detail
Age	Integer	Age of the patient.
Sex	String	M - Male and F - Female.
ChestPainType	String	ATA - Atypical Angina, ASY - Asymptomatic, TA - Typical Angina, and NAP - Non-Anginal Pain.
RestingBP	Integer	Resting Blood Pressure in mmHg.
Cholesterol	Integer	Serum Cholesterol in mm/dl.
FastingBS	Binary	1 - $if Fasting BS > 120mg/dl, 0 - otherwise.$
RestingECG	String	Resting Electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria.
Thalach	Integer	Maximum Heart Rate (between 60 and 202).
ExerciseAngina	String	Exercise - Induced Angina. Y - Yes, N - No.
Oldpeak	Float	Stress Test - ST depression induced by exercise relative to rest.
ST_Slope	String	The slope of the peak exercise ST segment. Flat - flat, Up - upsloping, Down - downsloping
Heart Disease	Binary	Output class: 0 - Normal and 1 - Heart disease.

blood glucose, maximum heart rate achieved, exercise-induced angina, resting electrocardiographic results, exercise-induced ST depression relative to rest, and peak exercise ST segment slope) in addition to a single target feature (heart disease), as shown in Table 1. The Long Beach VA dataset contains 200 instances (51 in class 0 and 149 in class 1), whereas the Hungarian dataset contains 294 instances (106 in class 0 and 188 in class 1).

Supplementary datasets, namely the Caesarean, Cervical, and Parkinson's datasets, are used to corroborate our experimental findings. The Caesarean dataset, provided by Campillo-Artero, Serra-Burriel, and Calvo-Pérez (2018), is derived from an exploration of 6,157 patient records collected in 2014 from four Spanish public hospitals and includes 161 features - 142 categorical and 19 numerical. These characteristics are divided into six groups, all of which, except for caesarean sections, are grouped together in a category called "normal delivery". Notably, there is an imbalance, with 692 records corresponding to caesarean deliveries and 5,465 to non-caesarean deliveries. The multivariate Parkinson's disease dataset, extracted from the UCI Machine Learning Repository (Sakar et al., 2019), consists of 188 Parkinson's disease patients (81 females and 107 males). Data acquisition consisted of recording each subject's articulation of the vowel /a/ three times, using a microphone calibrated at 44.1 kHz, resulting in a dataset of 756 instances and 754 attributes. The cervical cancer dataset, aimed at predicting cervical cancer indicators and diagnoses (Fernandes, Cardoso, & Fernandes, 2017), includes a conglomerate of demographic characteristics, lifestyle attributes, and historical medical records. It integrates data from 858 patients (840 in class 0 and 18 in class 1) covering 36 attributes.

To demonstrate the effectiveness of our proposed methodology, further validation was performed using the UNSW-NB 15, US Air Force LAN, CICDarknet 2020, and BETH datasets.

The UNSW-NB15 dataset (Moustafa & Slay, 2015) covers a range of nine network intrusion types, synthesizing a variety of real-world operational activities and contemporary attack methods. It includes 82,332 records in the test set and a substantial 175,341 records in the training set. Importantly, the training set ensures a balanced representation between attack scenarios and standard, non-intrusive operations. Conversely, the US Air Force LAN dataset, available from the Kaggle dataset repository (Dhanabal & Shantharajah, 2015), authentically replicates Local Area Network (LAN) attacks through a simulated flow of data between IP addresses adhering to specified protocols within a defined connection timeframe. Each connection, represented by a sequence of TCP packets, is categorically labeled as either 'normal' or a specific attack type. Analysis and data parsing facilitate the extraction of 41 attributes that encapsulate quantitative and qualitative dimensions of both typical and malicious connection scenarios. The CICDarknet2020 dataset (Aswad & Sonuç, 2020) presents a multi-faceted classification problem, exacerbated by the presence of data imbalance, with 141,530 observations across 85 columns. It is divided into two distinct labeling columns: Label-1, which identifies non-Tor users, non-VPN users, VPN users, and Tor users; and Label-2, which distinguishes between various usage categories, including browsing, audio streaming, chatting, file transfer, video streaming, email, VOIP, and additional specific use cases, each with a different number of cases. Lastly, the BETH dataset, also obtained from Kaggle (Highnam, Arulkumaran, Hanif, & Jennings, 2021), contains a robust 8,004,918 events. This dataset is derived from 23 honeypots, each strategically deployed on a major cloud provider and monitored at five-hour intervals. Preliminary process logs were judiciously selected for subsequent benchmarking and analysis efforts, with data subsets developed for training, validation, and testing based on host, log count, and activity metrics. Interestingly, the attack data is exclusively within the test subset.

3.3. Data preprocessing techniques

This research outlines a methodological approach to heart disease risk classification using ML classifiers, utilizing two elaborately curated datasets. Central to the analysis is the meticulous implementation of critical data preprocessing steps that provide a robust foundation for subsequent analysis stages. Firstly, missing value handling is undertaken to address potential gaps in the dataset and ensure that the subsequent analysis is based on a comprehensive data structure. Log-transformation is then applied, a technique crucial for stabilizing variance and making the data more amenable to the assumptions underlying many statistical and ML methods. This is followed by normalization, which ensures that different variables are made comparable by adapting them to a standard scale, thereby increasing the robustness and interpretability of the models developed. In addition, outlier detection is incorporated into the preprocessing stage to identify and

address anomalous values that may unduly bias the subsequent analysis. Finally, imbalanced data handling strategies are employed to ensure that the developed classifiers are not unduly influenced by the relative frequencies of the response variable categories.

3.3.1. Missing value handling

In real-world datasets, missing values are a primary cause of skewed results and significantly affect the performance of ML algorithms (Thomas, Bruin, Zhutovsky, & van Wingen, 2020). To resolve the problem of missing values in the Hungarian and Long Beach VA datasets, we treat the missing value column as the dependent variable and the other correlated columns as the independent variables. To replace any missing values with appropriate values, we use Random Forest (RF) as the regression model. We use the mean to handle the missing values of the other datasets.

3.3.2. Outlier detection

Outlier detection is crucial in ML algorithm development as outliers in a dataset decrease algorithm performance (Ramaswamy, Rastogi, & Shim, 2000). Outliers are identified using Tukey fences, which involve quartiles (Q1, Q2, Q3) to find extreme values (Zhou, Li, Li, Wang, & Wang, 2006). Q1 and Q3 are values below and above which 25% of data lies, respectively. Outliers fall below Q1 - 1.5 * (Q3 - Q1) or above Q3 + 1.5 * (Q3 - Q1). Outliers that are below the lower limit and above the upper limit are replaced with the lower limit and the upper limit, respectively.

3.3.3. Data balancing

Data balancing techniques are a crucial part of preprocessing because they assist classifiers in avoiding incorrect classifications due to imbalanced data. Data can be balanced by oversampling, undersampling, or combining the two techniques. In this research, we propose a novel technique that outperforms conventional methods.

3.3.4. Normalization

Normalization is the process of converting numerical column values in a dataset to a standard scale (García, Luengo, & Herrera, 2015). Normalization is essential when an ML model uses Euclidean distance for interpreting the inputs (Taunk, De, Verma, & Swetapadma, 2019). The Min-Max scaling method is used in this work to normalize the datasets. It divides the result by the range after subtracting the smallest value from the column's maximum value. Following normalization, each column's value ranges from 0 to 1.

3.4. Description of ML algorithms

In this research, we apply several benchmark ML algorithms to predict heart disease from secondary data. The short descriptions of the algorithms are listed below.

3.4.1. SVC

Support Vector Classifier (SVC) is a supervised ML algorithm used for binary classification (Sokoliuk, Kondratenko, Sidenko, Kondratenko, Khomchenko, & Atamanyuk, 2020). The key concept of this algorithm is to find a linear hyperplane that separates the two classes in the feature space with the largest margin. The margin is measured as the distance between the hyperplane and the nearest data points from each class. It is an efficient algorithm that can handle non-linearly separable data through the use of kernel functions to transform the data into a higher-dimensional space, enabling the discovery of a linear decision boundary (Prakash & Kanagachidambaresan, 2021).

3.4.2. DT

A Decision Tree (DT) classifier is a type of supervised learning algorithm used in ML for classification problems (Caruana & Niculescu-Mizil, 2006). The DT can be represented as a binary tree, where each node represents a test on an input feature, and each edge represents the outcome of that test. The leaves of the tree correspond to the class labels (Rokach & Maimon, 2005).

The DT classifier can be defined mathematically using the following equation:

$$h(x) = \sum_{i=1}^{L} y_i I(x \epsilon R_i) \tag{1}$$

where $h(x)$ is the predicted class label for input x, L is the number of leaves in the decision tree, R_i is the region of input space corresponding to the ith leaf, y_i is the class label assigned to the ith leaf, and $I(x \in R_i)$ is the indicator function that returns 1 if x is in R_i and 0 otherwise.

3.4.3. RF

RF is a popular ensemble learning algorithm employed in classification tasks. It is composed of a collection of decision trees, each built with a random subset of features and training data (Reis, Baron, & Shahaf, 2018). Each RF decision tree is built with a random subset of features and training data. The feature subsets are randomly selected at each node of the tree, and the training data subsets are created by bootstrapping the original dataset (Azar, Elshazly, Hassanien, & Elkorany, 2014). The criterion used to split the nodes of each tree is typically the Gini impurity or entropy, which measures the homogeneity of the class labels within each node. The final decision boundary is determined by the collective decision of all the trees in the forest.

RF is highly robust to noise and less prone to overfitting, especially important after data balancing, which can sometimes introduce synthetic noise. RF can also handle a large number of input features without requiring feature reduction techniques. This is particularly beneficial when dealing with complex medical datasets where numerous features may be relevant. Its ability to handle different types of data (numerical, categorical) and missing values makes it an excellent choice for real-world medical datasets, as confirmed in previous studies (Sumwiza, Twizere, Rushingabigwi, Bakunzibake, & Bamurigire, 2023).

3.4.4. XGBoost

XGBoost is an ensemble-based approach combining the strengths of gradient boosting and bagging techniques (Ferreira, Pilastri, Martins, Pires, & Cortez, 2021). The XGBoost algorithm creates a set of decision trees, each trained to fix the flaws of the previous one (Sagi & Rokach, 2021). A new tree is fitted to the negative gradient at the end of each iteration of the algorithm, which calculates the gradient of the loss function concerning the predictions made by the current model. The trees' predictions are then combined to give the final prediction. The algorithm also includes a regularization term to prevent overfitting and improve generalization.

3.4.5. LR

Logistic Regression (LR) is a statistical technique that models the relationship between a binary dependent variable and one or more independent variables (Zhu, Hu, Hou, & Li, 2021), making it a natural fit for predicting heart disease risk. Its inherent interpretability makes it easier to understand the impact of different features on heart disease risk. The model calculates the probability that, given the values of the independent variables, the dependent variable will take on the value 1. The logistic regression equation is given by:

$$p = \frac{1}{1 + \exp^{-z}} \tag{2}$$

where p is the predicted probability of the dependent variable being 1, e is the base of the natural logarithm, and z is the linear predictor.

3.4.6. KNN

The K-Nearest Neighbors (KNN) classifier is a simple and intuitive algorithm often used for classification tasks (Ali, Neagu, & Trundle, 2019). It works by labeling new instances according to their k-nearest neighbors' labels in the training set. Given a new instance x, the KNN algorithm first finds the k closest training instances to x using some distance metric, such as Euclidean distance. The algorithm then assigns the most frequent class label among these k neighbors to the new instance x. Using KNN may be computationally costly and sensitive to the choice of distance metric and value of k.

3.4.7. MLP

Multilayer Perceptron (MLP) is a type of feed-forward neural network comprising three layers - an input layer, a hidden layer, and an output layer (Japkowicz, 2001). The input layer receives the signal for processing, followed by the hidden layer capturing the nonlinear relationships between inputs and outputs. The necessary tasks, such as prediction and classification, are completed by the output layer. Within the deep MLP, the true computational engine consists of several hidden layers located between the input and output layers. MLPs can solve problems that cannot be linearly separated because they are designed to approximate any continuous function.

3.4.8. GNB

The Gaussian Naive Bayes (GNB) algorithm is a probabilistic classification algorithm based on Bayes' theorem (Leung et al., 2007). The term "naive" refers to the assumption that the features are independent of one another. The formula for GNB is given by:

$$p(y_k|x) = \frac{p(x|y_k)p(y_k)}{\sum_{j=1}^{K} p(x|y_j)p(y_j)} \tag{3}$$

where y_k is the class label, $\mathbf{x}$ is the input vector of features, $p(y_k|\mathbf{x})$ is the conditional probability of the class label given the input vector, $p(\mathbf{x}|y_k)$ is the conditional probability of the input vector given the class label, $p(y_k)$ is the prior probability of the class label, and K is the number of classes, $j = \{1, 2, \ldots, K\}$. A probability distribution, such as the Gaussian distribution for continuous features or the multinomial distribution for discrete features, is used to model the conditional probability $p(\mathbf{x}|y_k)$.

3.5. Existing data balancing techniques

To handle imbalanced class distribution for classification in ML, several existing methods are used. These methods employ oversampling and undersampling strategies to balance class distribution within a dataset.

SMOTE (Synthetic Minority Over-sampling Technique) is an oversampling technique that increases the number of instances in the minority class by generating synthetic samples that are similar to the existing minority samples (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). The SMOTE algorithm generates new samples by interpolating between existing minority class samples. Specifically, SMOTE selects k nearest neighbors from the minority class for each minority sample and generates new samples by interpolating between the sample and its neighbors.

NearMiss is an undersampling technique used to balance imbalanced datasets (Bao, Juan, Li, & Zhang, 2016). This technique selects the examples from the majority class that are closest to the examples of the minority class and retains only a subset of them. The subset is determined based on the parameter used for this technique.

SMOTE and Tomek Links are two resampling techniques used to address the issue of imbalanced datasets (Hasan, Islam, Sajid, & Hassan, 2022). SMOTE creates synthetic instances of the minority class by interpolating between existing samples, while Tomek Links identify pairs of examples that are closest to each other but belong to different classes and remove the majority class example (Chawla et al., 2002). SMOTETomek combines these two techniques to create a hybrid approach.

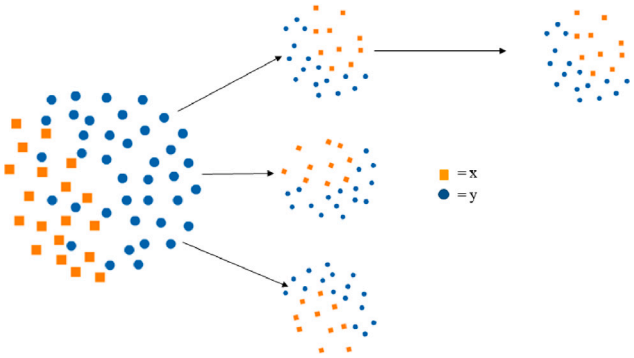


Fig. 2. Illustration of the proposed divide-and-conquer based data balancing technique.

3.6. Proposed data balancing technique

The above data-balancing techniques use either oversampling or undersampling to balance the dataset, considering the entire dataset as a single cluster for the balancing operation. Algorithm 1 and Fig. 2 represent our proposed balancing technique.

Initially, we divide the dataset into several clusters based on data characteristics determined by K-means clustering (Uddin et al., 2022). We balance each cluster separately and then merge the individual clusters to create the final balanced dataset. Within each cluster, we first identify the majority and minority classes and then apply the resampling techniques. Unlike existing approaches where the majority and minority classes are fixed, in our proposed technique, the majority and minority classes change based on the data sample of each cluster.

In each cluster, we randomly select data points from the minority class and compute the distances between these data points and their k nearest neighbors (Malangsa & Maravillas, 2017). We begin by multiplying the distance by a random number between 0 and 1, then incorporate the resultant data point into the minority class as a synthetic example. This step is repeated until the desired ratio is achieved. Subsequently, we randomly select another data point from the majority class and examine the nearest neighbors of the chosen item. If these neighbors belong to the minority class, we eliminate the randomly selected data point.

The selection of observations x and y should satisfy the following conditions.

- The nearest neighbors of observation x are y
- The nearest neighbors of observation y are x
- Both x and y belong to a different class. It means x and y belong to the majority and minority classes, and we select the two as a pair.

Mathematically, we define $d(x_i, x_j)$ as the Euclidean distance between the data point x_i and x_j. Here, x_i represents the minority class sample, while x_j represents the majority class sample. If there is no sample x_k satisfying the following conditions:

1. $d(x_i, x_k) < d(x_i, x_j)$
2. $d(x_j, x_k) < d(x_i, x_j)$

then the pair of $d(x_i, x_j)$ is the selected pair.

This technique can be employed to ascertain and remove data samples from the majority class that have the shortest Euclidean distance from the data in the minority class (i.e., the data from the majority class that is closest to the data from the minority class, making differentiation ambiguous).

In our clustering method, we explain the impact of some key parameters.

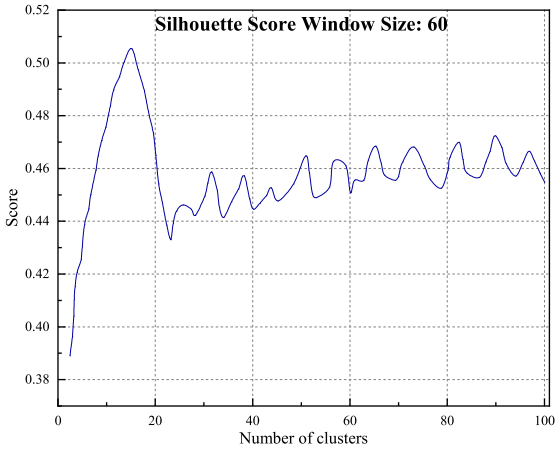


Fig. 3. Curves of silhouette scores with different values of K.

- Number of clusters (K): The number of clusters parameter, K, determines the granularity and structure of the clusters. A higher value of K may result in more clusters with smaller intra-cluster variation, while a lower value of K may generate larger, more generalized clusters.
- Distance metric: The choice of distance metric, such as Euclidean distance, impacts how proximity between data points is measured. Different distance metrics can lead to distinct cluster shapes and structures.
- Cluster initialization: The method used to initialize cluster centroids can influence the convergence speed and quality of the clustering results. Proper initialization can help avoid suboptimal solutions.
- Convergence criteria: Convergence criteria determine when the algorithm stops iterating. Setting appropriate convergence criteria is crucial to ensure the algorithm converges to a stable solution without unnecessary iterations.
- Cluster similarity measure: The similarity measure employed to assess the homogeneity within clusters and affects how data points are grouped together. Different similarity measures can lead to varying cluster structures.

The clustering method we propose involves validating clusters for homogeneity and separation by utilizing silhouette scores to determine the optimal clustering parameter K. In this process, we first assess the homogeneity and separation of clusters by calculating silhouette scores for different values of K. The silhouette score for each data point is calculated based on its distance to other points within the same cluster and to points in neighboring clusters. A higher silhouette score indicates that the data point is well-matched to its own cluster and poorly matched to neighboring clusters, implying better homogeneity and separation of clusters.

To choose the optimal value of K, we analyze the changes in silhouette scores as K varies. Fig. 3 illustrates the variation curve of silhouette scores. By iteratively adjusting the value of K, calculating silhouette scores, and observing how the scores change, we can identify the K value that maximizes the overall silhouette score. This optimal K value represents the number of clusters that best captures the underlying structure of the data, leading to well-defined clusters with distinct boundaries and meaningful separation between them.

By leveraging silhouette scores to validate clusters and select the appropriate value of K, our clustering method ensures that the resulting clusters are both internally coherent and well-separated from each other, enabling effective data segmentation and pattern discovery.

Algorithm 1 Proposed data balancing technique

Input: **Input:** Raw Dataset ($\boldsymbol{S}$)
Output: Balanced dataset, $\boldsymbol{S_b}$
Procedure Balancing Dataset ($\boldsymbol{S}$)
  Calculate the number of clusters $\boldsymbol{K_n}$ by applying Elbow Method in K-means clustering
  for $\boldsymbol{i}$=1 to $\boldsymbol{K_n}$ **do**
    Determine current majority samples
    Determine current minority samples
    Function Balance majority samples ()
    Select a random sample from the majority class
    Check the nearest neighbors of the sample
    if the neighbor's data are from a minority class **then**
      remove the random data point
    else
      do nothing
    end if
    end function
    Function Balance minority samples ()
    Choose a random data sample minority_point, $\boldsymbol{R_m}$ from the minority class.
    Check the k nearest neighbors of $\boldsymbol{R_m}$ within the minority class
    Distance $\boldsymbol{D_{mk}}$= Calculate the distances between $\boldsymbol{R_m}$ and its k nearest neighbors
    repeat
    Generate a random distance $\boldsymbol{D_{mr}}$ between 0 and 1.
    Generate synthetic sample by multiplying $\boldsymbol{D_{mk}}$ and $\boldsymbol{D_{mr}}$
    Until the minority class meets the desired proportion
    end function
  end for
  $\boldsymbol{S_b}$= Combine Balance majority samples () and Balance minority samples ()
end procedure

3.7. Explainable AI methods

3.7.1. SHAP

SHAP (SHapley Additive exPlanations) is an XAI technique that explains ML model outputs by decomposing them based on Shapley values from cooperative game theory (Zhang, Xu, & Zhang, 2020). This method breaks down predictions into input feature contributions, aiding comprehension and supporting tasks such as feature selection, debugging, and building trust (Zhang, Cho, & Vasarhelyi, 2022). SHAP values capture feature importance by considering feature interactions and ranking their impact on the model output. SHAP plots visualize these values, illustrating feature contributions to the model's output for a specific input (Lundberg, Erion, & Lee, 2018). Positive values indicate an increase in the feature's contribution, while negative values signify a decrease.

3.7.2. Shapash

Shapash is an open-source XAI tool that offers user-friendly and code-free interactive visualizations to explain ML outputs (Baniecki, Parzych, & Biecek, 2023). This tool empowers users to explore data, predictions, and generate specific prediction explanations. Visualizations depict input feature contributions, aiding comprehension of model output and revealing enhancement opportunities. Customized Shapash dashboards can seamlessly integrate into existing ML pipelines. Visualizations encompass global or local feature importance, dependence, and summary plots. Moreover, Shapash facilitates detailed prediction explanations by enabling sorting, filtering, identifying key input features, and exploring feature distributions.

3.8. Performance measure techniques for the classifiers

To find the best-performing ML algorithm, we need to measure the performance of the classifiers using various performance metrics such as accuracy, precision, recall, and F1 Score. Accuracy is a common performance measure used in classification tasks. It represents the proportion of correct predictions made by a model (Nabipour, Nayyeri, Jabani, Shahab, & Mosavi, 2020). Precision is the classifier's ability to not label a negative sample as positive (Gwetu, Tapamo, & Viriri, 2020). It is calculated by dividing the number of true positives by the sum of true positives and false positives. Recall, also called Sensitivity or the True Positive Rate, is defined as the number of positive predictions divided by the actual number of positive instances in the test data (Tharwat, 2020). F1 Score is interpreted as the harmonic mean of precision and recall. All these performance measurement tools can be expressed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precison = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{8}$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

We also use the Area Under the Curve (AUC) to determine the performance of the model. AUC represents a classifier's capacity to differentiate between positive and negative classifications (Deepak & Ameer, 2019). The AUC measures how efficiently the model differentiates between negative and positive classes.

The combination of these performance evaluation indicators provides a comprehensive and balanced assessment of the ML models' effectiveness. Our evaluation captures the models' ability to correctly identify both positive and negative cases, minimize false positives and false negatives, and maintain high discriminative power across different datasets.

Notably, in medical diagnosis, high precision means that the model has a low false positive rate, which is crucial to avoid unnecessary alarm and treatments. Recall is critical in medical contexts where identifying positive cases (such as heart disease) is vital for early intervention and treatment. High specificity indicates a low false positive rate, ensuring that healthy individuals are not incorrectly diagnosed with a condition. The F1 Score is more informative than accuracy for imbalanced datasets as it considers both false positives and false negatives. Unlike accuracy, precision, and recall, AUC provides a performance measure that is independent of the decision threshold, offering a more comprehensive evaluation.

Overall, this evaluation approach is particularly critical in the medical field, where the cost of misclassification can be high, and trust in predictive models must be firmly established.

4. Result and discussion

4.1. Results of the proposed data balancing technique

In this paper, a data balancing technique is proposed based on divide-and-conquer principles. We compare the performance of this new method against existing classifiers to determine its superiority. LR and RF are initially applied to the imbalanced dataset, followed by SMOTE. We then proceed to test NearMiss, SMOTETomek, and the proposed balancing techniques one by one to evaluate model performance.

Table 2
State of the data sample before and after data balancing.

Dataset	Class	Original	SMOTE	NearMiss	SMOTETomek	Proposed
Long beach VA	0	51	149	51	51	134
	1	149	149	51	149	134
Hungarian	0	106	188	106	106	173
	1	188	188	106	188	173
Cesarean	0	692	5465	692	692	5235
	1	5465	5465	692	5465	5235
Parkinsons	0	192	564	192	466	553
	1	564	564	192	466	553
Cervical	0	840	840	18	825	768
	1	18	840	18	56	768

Each experiment is run on an Intel Xeon E5-2683 CPU with 8 cores and 10 GB RAM.

Table 2 demonstrates the state of the data before and after being balanced, with both datasets shown in five different states. The resulting datasets are initially imbalanced; however, with the application of balancing techniques, the instances increase and decrease at random, resulting in a balanced dataset. The balanced dataset is then used independently of the ML classifiers, and the results are evaluated using the classification report.

4.1.1. Performance of the classifiers on imbalance datasets

On both datasets, we use LR and RF. Table 3 displays the outcomes for the Long Beach VA and Hungarian datasets, which resulted in an imbalanced state. The performance of the ML classifiers is poor, and the F1 scores for individual classes are unstable. In Table 3, the precision and recall for class 0 are lower than for class 1 for both classifiers. This table shows a significant contrast in performance. Because the Long Beach VA data are more imbalanced than the Hungarian dataset, we must use data balancing techniques to balance the data and obtain a stable output with good accuracy.

4.1.2. Performance of the classifiers after balancing in long beach VA and hungarian dataset

We applied balancing techniques to both the Long Beach and Hungarian datasets and presented the results in Table 3. The findings showed that the proposed data balancing approach outperformed other techniques in both classifiers. Although the accuracy of LR and RF was nearly identical in the imbalanced Long Beach dataset, other metrics were unstable for both classes. After implementing the data-balancing techniques, our proposed methods showcased a significant impact on the classification accuracy of both classes. The proposed data-balancing technique enhances the accuracy and stabilizes other performance measurement metrics. Our methods aid the classifiers, resulting in the accuracy evolving from 77% to 89% for LR and 77% to 91% for RF. We have plotted the ROC for all potential combinations in Figs. 4 and 5, which clearly exhibit that the proposed data balancing technique is superior, as shown in Table 3. Therefore, we can adopt RF as a classifier for predicting heart disease after utilizing the proposed data balancing technique.

4.1.3. Performance of the classifiers on other datasets

Our proposed balancing technique was applied to the Caesarean, Cervical, and Parkinson's datasets, as shown in Table 4. Results indicate a significant discrepancy between the two classes when utilizing existing balancing techniques. However, our proposed method achieved equal performance for both classes. The LR model achieved an F1 score of 9% for class 0 and 97% for class 1 in the Caesarean dataset. The RF model also delivered notable performance with an F1 score of 98% for both classes. Using our proposed method, the LR algorithm attained a remarkable 99% F1 score for both classes in the Cervical dataset. In the Parkinson's dataset, the LR model's F1 score stood at 95%, while the RF model showed an impressive F1 score of 97% for class 0 and 98% for class 1.

Table 3
Performance of the classifiers after balancing on the Long Beach VA dataset and Hungarian dataset.

Technique	Algor.	Class	Long Beach VA		Hungarian	
			F1 Score	Accuracy	F1 Score	Accuracy
Imbalance	LR	01	0.370.86	0.78	0.660.83	0.77
	RF	01	0.490.87	0.79	0.670.83	0.77
SMOTE	LR	01	0.720.71	0.72	0.860.87	0.87
	RF	01	0 850.85	0.85	0.860.85	0.86
NearMiss	LR	01	0.560 67	0.62	0.720.61	0.67
	RF	01	0.560.67	0.62	0.730.65	0.70
SMTomek	LR	01	0 330.81	0.70	0.790.83	0.81
	RF	01	0.600.87	0.80	0.810.84	0.83
Proposed	LR	01	0.760 79	0.78	0.880.89	0.89
	RF	01	0 900.92	0.91	0.910.92	0.91

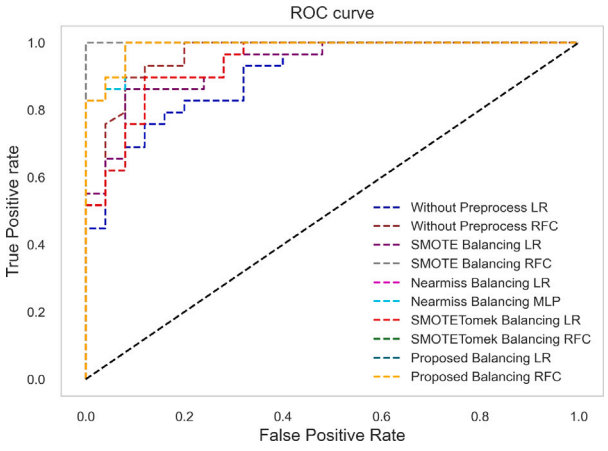


Fig. 4. ROC curve of all possible combinations in Long Beach VA dataset.

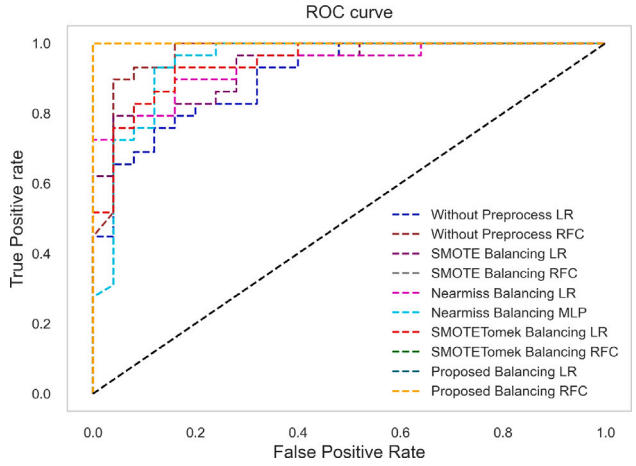


Fig. 5. ROC curve of all possible combinations in Hungarian dataset.

4.1.4. Validation of proposed data balancing technique

To assess the effectiveness of our balancing method, we applied it to four datasets: UNSW-NB 15, US Air Force LAN, CICDarknet 2020, and BETH. The experimental results are presented in Table 5, which demonstrate that our proposed technique outperforms the imbalanced state for all datasets. In the case of the UNSW-NB 15 dataset, the accuracy of the RF algorithm increased from 85% to 95%. Additionally, the RF algorithm achieved 96% accuracy on the US Air Force LAN dataset. The LR and RF algorithms of the CICDarknet 2020 dataset perform equally well with our method, exhibiting an enhanced accuracy of 99%, compared to the prior 88%. As for the BETH dataset, our method advances the classification accuracy from the previous 88% to 91% for the LR algorithm and from 89% to 98% for the RF algorithm.

Table 4
F1 scores for classifiers on imbalanced datasets and datasets balanced using existing methods and our proposed method.

Technique	Algor.	Class	Caesarean	Cervical	Parkinson's
Imbalanced	LR	0	0.78	0.93	0.76
		1	0.97	0.78	0.82
	RF	0	0.87	0.95	0.85
		1	0.98	0.88	0.90
SMOTE	LR	0	0.90	1.00	0.92
		1	0.94	0.92	0.93
	RF	0	0.88	1.00	0.94
		1	0.95	0.91	0.96
NearMiss	LR	0	0.89	0.56	0.65
		1	0.89	0.72	0.72
	RF	0	0.80	0.67	0.82
		1	0.82	0.88	0.83
SMOTETomek	LR	0	0.78	0.92	0.75
		1	0.97	0.75	0.82
	RF	0	0.87	0.92	0.80
		1	0.98	0.77	0.80
Proposed	LR	0	0.96	0.99	0.95
		1	0.97	0.99	0.95
	RF	0	0.98	1.00	0.97
		1	0.98	1.00	0.98

Table 5
Validation of proposed data balancing technique using four different datasets.

Dataset	Algor.	Imbalanced (Accuracy)	After Balancing (Accuracy)
UNSW-NB 15	LR	0.75	0.92
	RF	0.85	0.95
US Air Force LAN	LR	0.85	0.95
	RF	0.89	0.96
CICDarknet 2020	LR	0.88	0.99
	RF	0.87	0.99
BETH	LR	0.88	0.91
	RF	0.89	0.98

4.2. Result of heart disease prediction and performance discrepancy mitigation

The datasets were initially assessed without utilizing our suggested preprocessing pipeline to examine the inter-data discrepancy in heart disease risk prediction. Subsequently, each component of the preprocessing pipeline was used independently to ascertain its functionality. A summary of the results is presented through bar plots and ROC curves.

The hyperparameters associated with the algorithms were determined via Grid Search Cross Validation (CV) (Patil & Bhosale, 2021). After fitting the models to the datasets, the algorithm's output can be evaluated. The hyperparameters and their respective values for the Hungarian dataset are listed in Table 6.

Indeed, the careful adjustment and optimization of model parameters based on the specific characteristics of the Long Beach VA and Hungarian datasets were crucial in achieving high performance. Specifically, this tuning was necessary because the Long Beach VA dataset, with its higher imbalance, required stronger regularization to avoid overfitting, whereas the Hungarian dataset, being more balanced, benefited from a more flexible model with a higher complexity C value. For the Hungarian dataset, 'lbfgs' provided better convergence and stability, while for the Long Beach VA dataset, 'liblinear' was preferred due to its efficiency with smaller and more imbalanced datasets. Further, for the more complex Hungarian dataset, deeper trees were found to capture the complex relationships between features better. In contrast, shallower trees sufficed for the Long Beach VA dataset to prevent overfitting given its smaller size.

Table 6
Hyperparameters tuning of the classifiers using grid search CV.

ML Classifier	Parameter and Value
SVC	*probability*=True, *C* = 10, *gamma* = 0.1, *kernel* = linear
KNN	*n_neighbors* = 5, *algorithm* = 'ball_tree', *weights* = 'distance', *metric* = 'minkowski', *p* = 2
DT	*criterion*='gini', *splitter*='best'
LR	*penalty*='l2', *C*=1.0, *random_state*=None, *solver*='lbfgs', *max_iter*=100, *multi_class*='auto', *verbose*=0
RF	*n_estimators* = 100, *random_state* = 42
XGB	*n_estimators* = 100, *booster* = 'gbtree', *gamma* = 0
GNB	*priors*=None, *var_smoothing*=1e−09
MLP	*hidden_layer_sizes*=(100), *activation*='relu', *solver*='adam', *alpha*=0.0001, *batch_size*='auto', *learning_rate*='constant'

Table 7
Performance of the classifiers on the Hungarian and Long Beach VA dataset without preprocessing.

Model	Hungarian				Long Beach VA			
	Acc	F1 Score	Spec.	AUC	Acc	F1 Score	Spec.	AUC
SVC	0.64	0.50	0.96	0.78	0.80	0.71	0.00	0.61
DT	0.88	0.88	0.89	0.88	0.68	0.70	0.50	0.61
RF	0.90	0.90	0.92	0.95	0.83	0.81	0.38	0.83
XGB	0.92	0.91	0.95	0.96	0.70	0.70	0.25	0.69
LR	0.83	0.83	0.84	0.94	0.78	0.75	0.25	0.78
KNN	0.66	0.65	0.79	0.70	0.73	0.72	0.25	0.52
MLP	0.83	0.83	0.89	0.92	0.75	0.69	0.00	0.66
GNB	0.81	0.82	0.82	0.92	0.73	0.72	0.25	0.82

4.2.1. Performance of the classifiers on a single dataset

Initially, we analyzed the two individual datasets without utilizing our proposed preprocessing pipeline. We proceeded to train the machine learning models utilizing ratios of 80:20, 70:30, and 50:50 for training and testing data. Table 7 presents the classifiers' performance for the Long Beach VA and Hungarian datasets. Among all the methods experimented with in the Hungarian dataset, RF and XGB demonstrate the highest accuracy of 90% and 92%, respectively. The AUC scores for RF and XGB are 95% and 96%, suggesting high algorithm performance on both classes. However, considering the AUC scores and the minimum scores achieved by the different performance measurement techniques, RF and XGB exhibit more obvious superiority. In terms of accuracy, SVC displays lower performance than other algorithms but presents high specificity.

In some instances, the classifiers demonstrate low specificity, implying the existence of false positives while exhibiting a low number of true negatives. This results in all non-cases being erroneously identified as positive. Consequently, the collective occurrence of these phenomena implies that all instances were deemed positive regardless of their actual nature. See Fig. 6 for the ROC of this dataset.

The Long Beach VA dataset produces diverse results. The RF outperforms the XGB, but the latter experiences a decline. Conversely, SVC delivers favorable outcomes, though its precision, Cohen Kappa, and specificity are suboptimal. RF accuracy reaches 83%, but its specificity is merely 38%. For XGB, only 70% accuracy and 25% specificity are achieved. Comparatively, the classifiers fare worse than in the Hungarian dataset. The ROC in Fig. 7 illustrates the mean performances of the classifiers. It is essential to achieve high levels of accuracy, precision, and recall in this critical healthcare concern.

4.2.2. Checking the performance discrepancy issue on the heart disease datasets

We need to modify the training and testing phases to assess the performance variation in heart disease prediction. Initially, we used the Hungarian dataset for training and the Long Beach VA for testing, and we recorded the classifier performances in Table 8. Subsequently, we

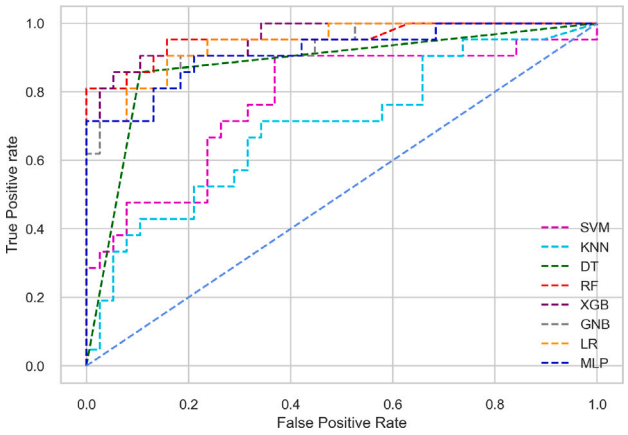


Fig. 6. Performance of the classifiers on the Hungarian dataset without preprocessing.

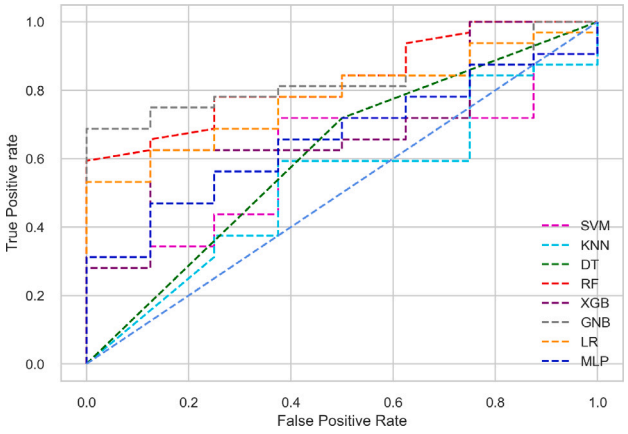


Fig. 7. Performance of the classifiers on the Long Beach VA dataset without preprocessing.

Table 8
Performance discrepancy on heart disease prediction (Hungarian: Long Beach VA) and (Long Beach VA: Hungarian).

Model	Hungarian:Long Beach				Long Beach:Hungarian			
	Acc	Fl Score	Spec	AUC	Acc	Fl Score	Spec	AUC
SVC	0.25	0.11	0.96	0.59	0.36	0.19	0.25	0.63
DT	0.52	0.54	0.63	0.55	0.70	0.69	0.84	0.65
RF	0.58	0.60	0.69	0.68	0.76	0.76	0.85	0.82
XGB	0.48	0.49	0.76	0.67	0.68	0.65	0.88	0.76
LR	0.57	0.59	0.75	0.69	0.79	0.77	0.95	0.86
KNN	0.46	0.48	0.71	0.53	0.43	0.38	0.19	0.57
MLP	0.72	0.72	0.41	0.69	0.61	0.62	0.54	0.75
GNB	0.75	0.74	0.37	0.69	0.81	0.81	0.85	0.85

inversed the setup and used the Long Beach VA for training and the Hungarian dataset for testing, assessing the classifiers' performance. In the initial case, the performance of the classifiers is inadequate, and the precision, recall, and other metrics are unreliable. GNB attains a peak of 75% accuracy and 69% AUC. RF displays merely 58% accuracy, while SVC only exhibits a minimum of 25% accuracy. The recorded performance metric values demonstrate a clear discrepancy in inter-dataset performance in this configuration. The ROC curve vividly displays the disparity in Fig. 8.

For the Long Beach VA case, using Hungarian as testing data and Long Beach VA as training data presented similar issues. The GNB algorithm demonstrated its superiority with 81% accuracy, precision, recall, and f1 score, as well as 85% specificity and AUC. On the other hand, the RF algorithm achieved an average accuracy of 76% with the same

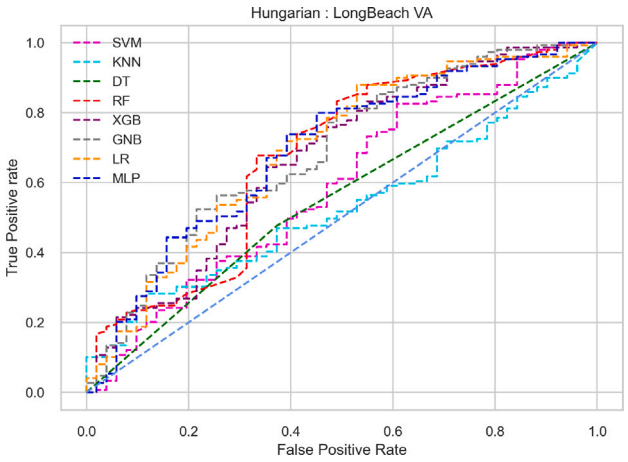


Fig. 8. Performance discrepancy on heart disease risk prediction (Hungarian: Long Beach VA).

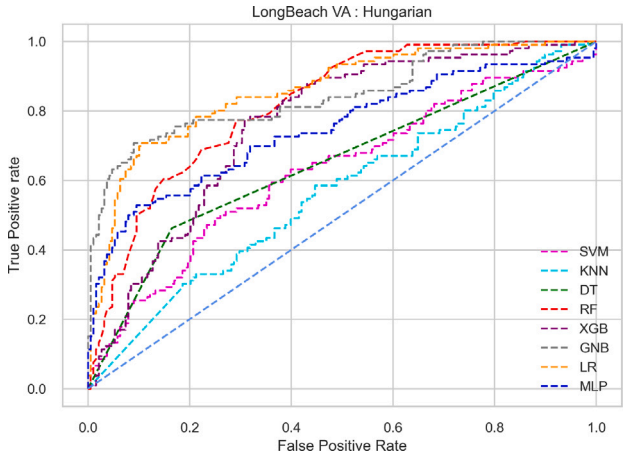


Fig. 9. Performance discrepancy on heart disease risk prediction (Long Beach VA: Hungarian).

precision, recall, and F1 score. SVC demonstrated a minimum of 36% accuracy, 13% precision, 36% recall, and 25% specificity. This also demonstrates the difference in performance within this configuration. The ROC of this configuration can be found in Fig. 9.

4.2.3. Performance of the classifiers after applying the proposed preprocessing pipeline on single datasets

To improve the accuracy of heart disease prediction, we established a preprocessing pipeline consisting of various techniques to enhance classifier performance and stability. The application of this pipeline led to a significant improvement in classifier performance, as demonstrated in Table 9. RF, XBG, and SVC exhibited a 92% accuracy rate, with RF displaying superior AUC performance. The algorithm's performance improves and reaches a stable state, which is the objective within the healthcare sector. The ROC in Fig. 10 displays a positive curve, indicating stable algorithm performance.

In a comparable setting utilizing the Long Beach VA dataset, the experimental results demonstrate that RF attains 93% accuracy, recall, and F1 score with 94% precision and 97% AUC, rendering it remarkably effective in sensitive domains such as healthcare. The majority of the algorithm's performance gain is due to the proposed preprocessing pipeline. The ROC depicted in Fig. 11 reveals the classifiers' stability in this situation.

Table 9
Performance of the classifiers on the Hungarian and Long Beach VA dataset with preprocessing.

Model	Hungarian				Long Beach VA			
	Acc	Fl Score	Spec	AUC	Acc	Fl Score	Spec	AUC
SVC	0.92	0.92	0.95	0.98	0.90	0.90	0.87	0.96
DT	0.86	0.86	0.87	0.86	0.90	0.90	0.87	0.90
RF	0.92	0.93	0.95	0.99	0.93	0.93	0.87	0.97
XGB	0.92	0.92	0.97	0.97	0.90	0.90	0.83	0.96
LR	0.88	0.88	0.89	0.94	0.63	0.63	0.57	0.76
KNN	0.89	0.89	0.89	0.98	0.90	0.90	0.90	0.96
MLP	0.88	0.88	0.92	0.96	0.65	0.65	0.60	0.75
GNB	0.83	0.83	0.82	0.88	0.72	0.71	0.57	0.75

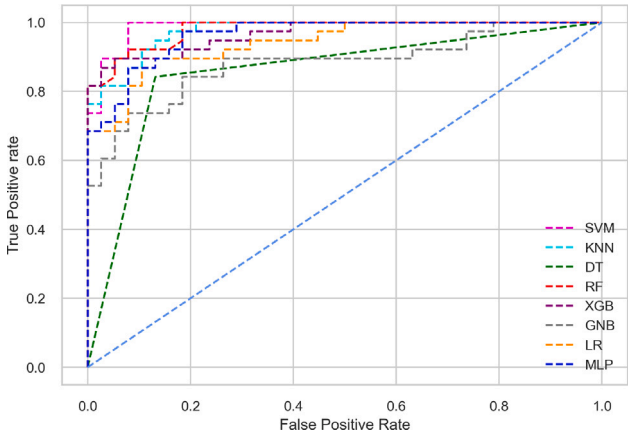


Fig. 10. Performance of the classifiers on the Hungarian dataset with preprocessing.

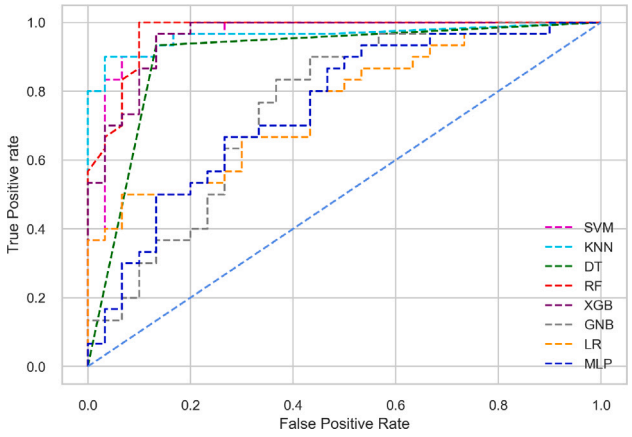


Fig. 11. Performance of the classifiers on the Long Beach VA dataset with preprocessing.

4.2.4. Performance discrepancy mitigation after applying the preprocessed pipeline

To address the issue of performance discrepancy, both datasets are processed through the pipeline using the previous setup, as shown in Table 10. KNN achieves 88% accuracy, precision, recall, and F1 score. RF outperforms KNN in terms of AUC. Table 10 demonstrates the mitigation of performance discrepancy in heart disease prediction. The ROC curve in Fig. 12 is more stable than in the previous setup.

In contrast, during training at the Long Beach VA and testing with Hungarian data, there is a marked increase in classifier performance. The experiment demonstrates that KNN achieves an accuracy of 90% alongside equal precision, recall, and F1 scores. In this setup, RF attains a maximum AUC of 95%. The ROC charted in Fig. 13 displays greater

Table 10
Performance discrepancy mitigation on heart disease prediction (Hungarian: Long Beach VA) and (Long Beach VA: Hungarian).

Models	Hungarian:Long Beach				Long Beach:Hungarian			
	Acc	F1 Score	Spec	AUC	Acc	F1 Score	Spec	AUC
SVC	0.85	0.85	0.81	0.94	0.87	0.87	0.78	0.94
DT	0.73	0.72	0.66	0.72	0.85	0.85	0.84	0.85
RF	0.83	0.83	0.87	0.95	0.87	0.87	0.78	0.95
XGB	0.86	0.86	0.91	0.95	0.86	0.86	0.82	0.95
LR	0.79	0.79	0.77	0.89	0.85	0.85	0.76	0.93
KNN	0.88	0.88	0.87	0.91	0.90	0.90	0.86	0.95
MLP	0.81	0.81	0.79	0.90	0.83	0.83	0.73	0.93
GNB	0.77	0.77	0.74	0.86	0.81	0.81	0.73	0.90

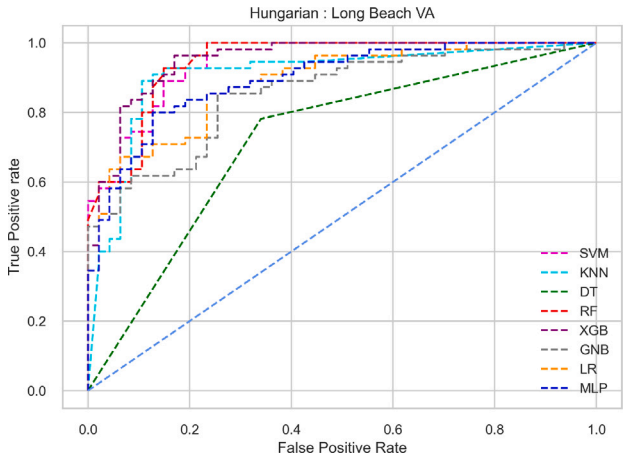


Fig. 12. Performance discrepancy mitigation on heart disease prediction (Hungarian and Long Beach VA).

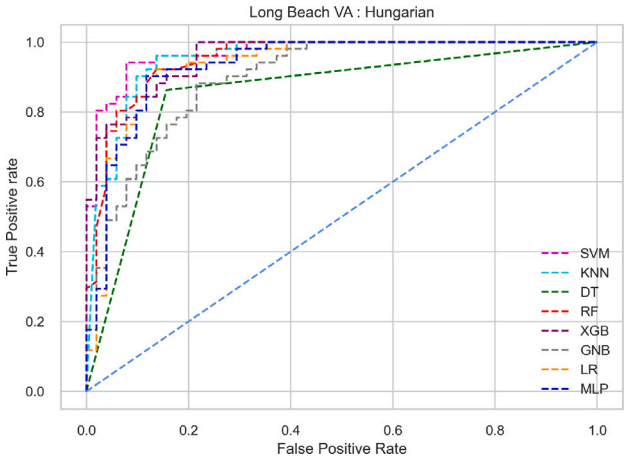


Fig. 13. Performance discrepancy mitigation on heart disease prediction (Long Beach VA and Hungarian).

stability, indicating the reduction of performance disparities in heart disease prognosis.

This analysis highlights that appropriate preprocessing techniques can considerably alter classifier performance. Once we have addressed inter-dataset performance discrepancies, a comprehensive healthcare dataset and model can be established.

4.3. Comparison and analysis of experiments

This analysis highlights that appropriate preprocessing techniques can considerably alter classifier performance. Once we have addressed inter-dataset performance discrepancies, a comprehensive healthcare dataset and model can be established.

Table 11
Performance comparison with other methods.

Model	AUC	Precision
Our method	0.793	0.765
Multivariate Regression	0.714	0.727
Naive Bayes	0.707	0.717
Bagged Trees	0.745	0.732
AdaBoost	0.786	0.713
Ensemble machine learning algorithms (MLAs)	0.787	0.743
Framingham Score	0.760	0.737

When comparing our approach with the latest methods, we observed a significant improvement in the AUC and precision metric for model classification in Table 11. This enhancement can be attributed to the introduction of data balancing and an efficient clustering mechanism in our method. By applying data balancing techniques, we were able to address issues related to imbalanced datasets and improve the overall performance of the model. Additionally, the implementation of an efficient clustering mechanism helped in identifying distinct patterns within the data, leading to more accurate classification results. Our method's effectiveness in enhancing model performance is underscored by these key factors, highlighting the importance of incorporating data balancing and efficient clustering strategies in machine learning processes.

The performance of each algorithm was evaluated using cross-validation, specifically employing k-fold cross-validation, a widely recognized technique for model assessment. This method helps prevent overfitting and ensures that models can be applied to new, independent datasets. K-fold cross-validation involves randomly dividing the sample data into k equal-sized subsamples. In this study, the data was divided into 10 folds, a commonly used number. Each of the 10 folds was then used as a test set, with the data in the remaining 9 folds serving as training data. Performance metrics were calculated for each fold, and the mean and standard deviation were computed to evaluate the overall performance of each algorithm. The results of the mean and standard deviation for each algorithm are displayed in Table 12.

Further statistical analysis of our method about comparison of confidence intervals on AUC metric in Table 13 revealed that, when compared to other models, our approach exhibited shorter confidence intervals. We chose to compare using a 95% confidence interval, which is constructed by sample statistics to estimate the parameter. This interval gives us 95% confidence that it contains the true population parameter value when conducting multiple experiments. The results indicate a higher level of precision and reliability in the estimated results. The tighter confidence intervals suggest that our method produces more consistent and accurate outcomes, contributing to greater confidence in the model's performance and predictive capabilities. By reducing the variability in the estimates, our approach enhances the robustness of the results and underscores its efficacy in providing reliable and dependable insights for decision-making processes. The narrower confidence intervals signify the method's ability to generate more stable and trustworthy predictions, thereby reinforcing its value in practical applications and research contexts.

In order to further evaluate the performance of the model in terms of runtime, Table 14 presents a comparative analysis of the runtime of our method and four other approaches. All runtime values are obtained by running the models 20 times and calculating the average. Upon testing on two distinct datasets, it was observed that our proposed method consistently exhibits lower runtime durations. This can be attributed to our enhanced data balancing strategy, which effectively minimizes the interference caused by imbalanced data during cardiovascular disease risk analysis. By optimizing the data balance, our method significantly boosts model performance and efficiency. The superior runtime of our approach underscores its effectiveness in handling data intricacies

Table 12
10-fold cross-validation results.

Algorithm	Measure	Accuracy	Precision	Recall	F1 Score	AUC
Our method	Mean	0.808	0.817	0.611	0.699	0.833
	Standard Deviation	0.004	0.015	0.020	0.016	0.004
Decision Jungle	Mean	0.882	0.853	0.340	0.521	0.801
	Standard Deviation	0.004	0.025	0.023	0.022	0.008
Locally Deep Support Vector Machine	Mean	0.892	0.853	0.463	0.569	0.804
	Standard Deviation	0.006	0.039	0.031	0.027	0.009
Neural Network	Mean	0.859	0.664	0.632	0.643	0.811
	Standard Deviation	0.007	0.053	0.066	0.015	0.006

Table 13
Comparison of confidence intervals on AUC metric.

Classification	Confidence intervals	
	Minimum	Maximum
Our method	0.819	0.841
Maximum Likelihood	0.726	0.776
SVM	0.716	0.796
RF	0.704	0.778

Table 14
Comparison of running time (s).

Data Sets	KNN	Fuzzy KNN	Locally deep SVC	RF	Pro-posed Method
Hungarian	15.36	11.74	13.52	15.21	7.21
Long Beach VA	16.35	12.34	12.69	14.22	6.16

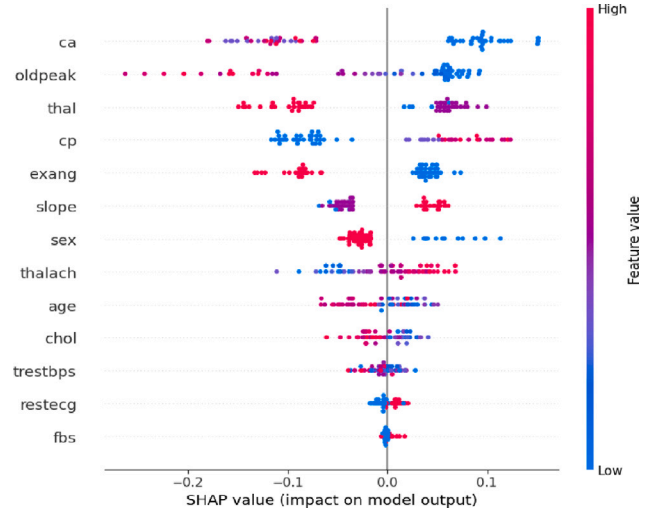


Fig. 14. Impact of the features on model output using SHAP 1 value.

and highlights its ability to deliver accurate and efficient predictions. This efficiency is crucial in real-world applications where timely and precise analysis of cardiovascular disease risk is essential for effective decision-making and healthcare interventions.

4.4. Model explainability

The significance of each feature on a particular sample is shown in Fig. 14 in a sorted manner. This plot tells us that features like *ca, oldpeak, exang, thal* are negatively correlated with the target variable, and *cp, slope* are positively correlated.

We first identified the most crucial traits, *ca, thal*, and *cp*, and then we looked at each of the top 3 features independently in Figs. 15 to 17 to comprehend them better. The value 0 has a positive predictive influence on the model, whereas the other values have a negative

Table 15
Rank and frequency of the domain expert's opinion.

Feature	D1	D2	D3	D4	D5	D6	D7	Freq	Rank
age	1	0	0	1	0	1	1	4	2
sex	0	1	1	1	1	0	1	5	2
cp	1	1	1	1	1	1	1	7	1
trest bps	0	1	0	0	1	1	0	3	3
chol	1	0	1	0	0	1	1	4	2
fbs	0	1	1	0	1	0	0	3	3
restecg	0	1	1	0	1	0	0	3	3
thalach	1	0	0	1	1	1	1	5	2
exang	1	1	1	1	1	1	1	7	1
old peak	1	1	1	1	1	1	1	7	1
slope	1	1	1	1	1	1	1	7	1
ca	1	1	1	1	1	1	1	7	1
thal	1	1	1	1	1	1	1	7	1

predictive influence, according to the feature contribution graph of *ca*. The risk of developing heart disease increases when fewer main vessels are seen with fluoroscopy. In the case of features *thal*, the *normal* and *fixed defect* qualities impact the model to predict positively. Having typical angina in the context of feature *cp* lowers the risk of developing heart disease.

The effects of the features on some selected data points are illustrated in Fig. 18. Using Patient 198 as an example, we can observe that no significant factors, in this case, might contribute to heart disease. Contrarily, Patient 283's *thal* value did not impact much, but their *ca, oldpeak chol* value significantly influenced her 96% likelihood of getting heart disease.

To further comprehend the effects of our examined features, we look at Figs. 19 and 20. For patient 246, the most important factors contributed negatively, indicating just a 0.38% probability of having heart disease. On the other hand, a patient with ID 57 had major factors like *ca, thal, age, exang, slope, and oldpeak* that were positively contributing; as a result, the patient had a 92% chance of having heart disease.

4.4.1. Ground truth explanation

We surveyed the domain experts (Doctors and Medical senior students) to find the weight and rank of the features. Firstly, we conducted face-to-face interviews with seven doctors and asked them to select variables from the dataset list. According to their opinion, we count the frequency and rank the features in Table 15. We get 6 rank 1 features, 4 rank 2 features, and 3 rank 3 features. The rank 1 features are much more important, and we put weight 3 to the features and then give weight 2 for the rank 2 features and weight 1 to the rank 3 features. Then we conducted another survey of the seven medical doctors and 67 students (MBBS) with a questionnaire and asked them some questions and put the value corresponding to each feature from 1 to 5, where 5 means the features are more important and 1 indicates less importance.

From the survey, it is clearly identified that heart disease is one of the major problems worldwide. 100% of the participants agreed with the question, "Do you think Heart disease is now a common problem worldwide?". Among the 67 participants, all believe our unhealthy lifestyle is the main cause of heart disease. About 73.33% of them

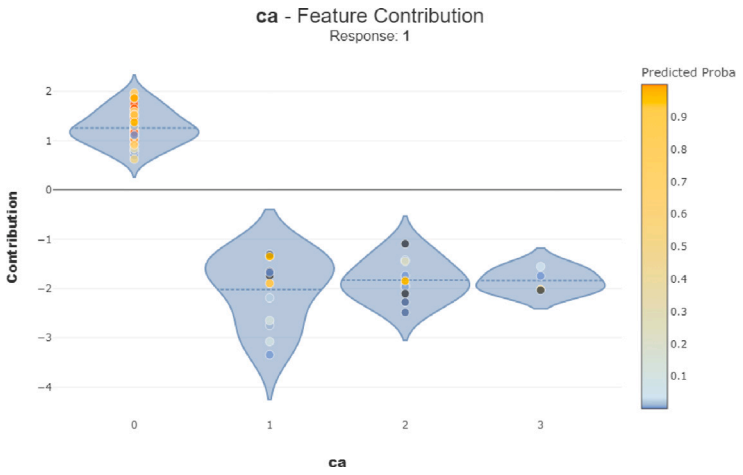


Fig. 15. Importance of the top features (ca) on the model.

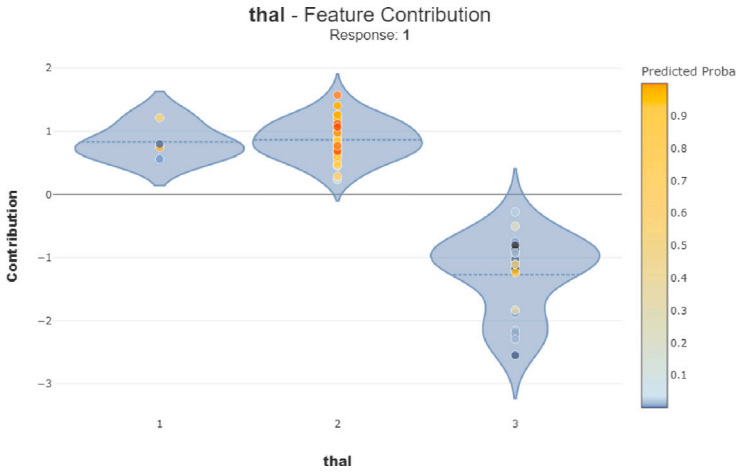


Fig. 16. Importance of the second top features (thal) on the model.

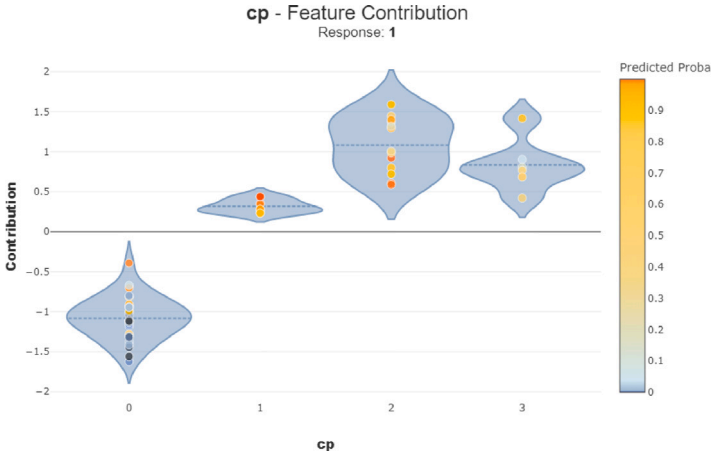


Fig. 17. Importance of the third top features (cp) on the model.

believe food habit is also responsible for this disease, and only 13.33% of them agree that it can be caused by a Genomic Problem.

After adding weight to the survey values, we get the final output that is meaningful to the output of XAI tools. The features are shown

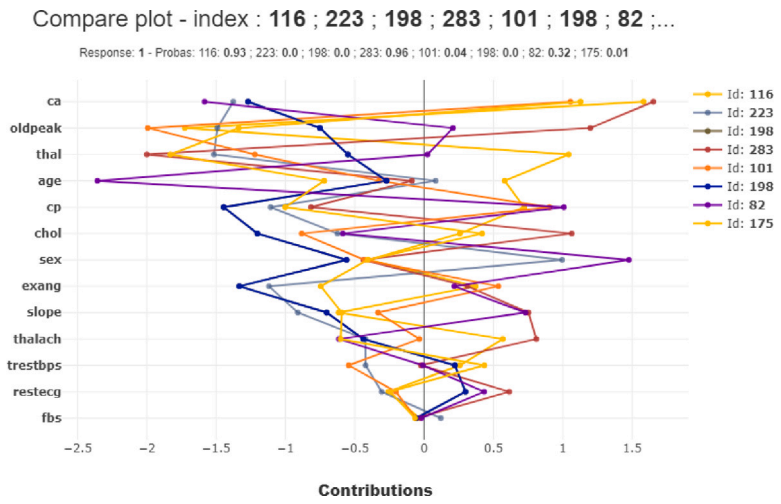


Fig. 18. Comparison of the effect of features on some data points.

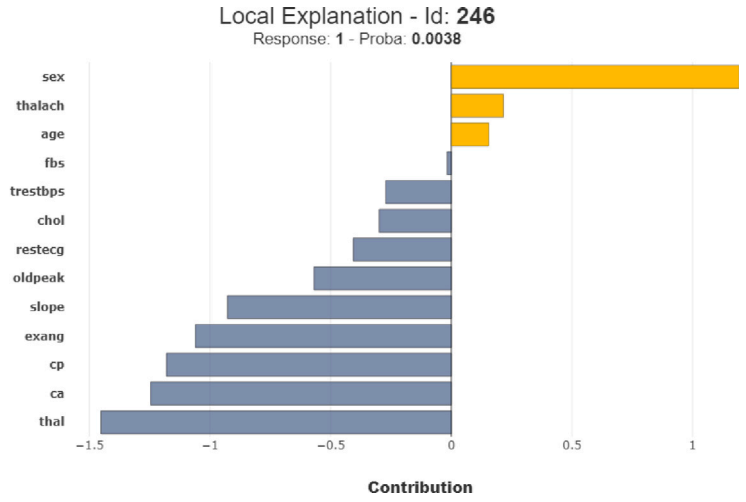


Fig. 19. Local explainability of non-heart disease instance.

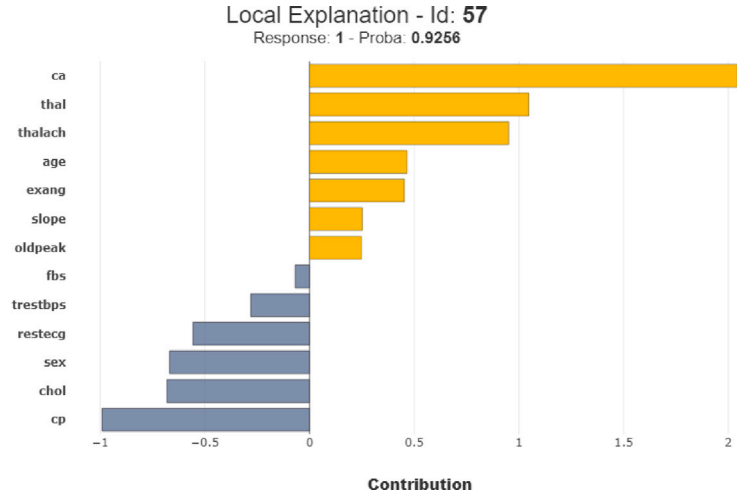


Fig. 20. Local explainability of heart disease instance.

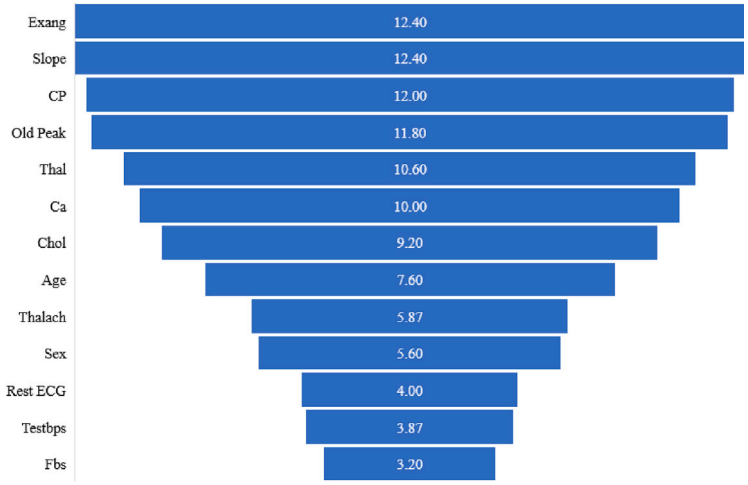


Fig. 21. Rank and score achieved of the features from the survey.

in decreasing order in Fig. 21. The top 6 features show a different order than XAI ranked features. The number of major vessels is ranked top by the XAI tools, but in the ground truth, it gets 6th position but is still in category 1. In XAI tools, explainability, and ground truth explainability, the features are in the same categories, but the orders differ. It can be a lack of XAI tools understanding or the data differences between the foreign countries and Bangladeshi survey.

5. Discussion

The current study advances the field of cardiovascular risk assessment by introducing a novel data balancing technique, a comprehensive preprocessing pipeline, and the integration of XAI methods. Our proposed approach significantly improves the predictive accuracy and stability of ML classifiers when applied to imbalanced datasets, specifically for early risk prediction of heart disease.

The proposed divide-and-conquer strategy using K-means clustering allows oversampling and undersampling to be applied independently. This method addresses the shortcomings of traditional balancing techniques, such as overfitting in oversampling and data loss in undersampling, leading to improved classifier performance by preserving class variance and reducing bias. By using two different datasets (Long Beach VA and Hungarian), our holistic pipeline mitigates performance discrepancies commonly seen in single dataset studies. This approach ensures that ML models are robust and generalizable across different datasets. In addition. by incorporating XAI methods such as SHAP and Shapash, we provide detailed insight into the importance of features and their contribution to predictive models. This transparency bridges the gap between model predictions and domain expert knowledge, fostering trust and interpretability in clinical settings.

Our study shows that the proposed data balancing technique significantly outperforms traditional methods such as SMOTE, NearMiss and SMOTETomek. Notably, the proposed method increased classification accuracy from 77% to 89% for LR and from 77% to 91% for RF on the Long Beach VA dataset. Furthermore, the performance metrics, including precision, recall and F1 score, are more stable across different classes when using our proposed technique. This can be attributed to the limitations of the compared methods, such as susceptibility to overfitting (SMOTE), data loss (NearMiss), and potential introduction of noise (SMOTETomek).

The current study also identified several critical features for early prediction of heart disease risk, highlighted through the use of XAI tools. These features include: (a) age (as a primary risk factor in most cardiovascular models), (b) cholesterol levels (another well-known risk factor for heart disease), (c) resting blood pressure (strongly correlated with increased heart disease risk), (d) exercise-induced angina (presence/absence of angina during physical activity seems to be a critical diagnostic indicator), (e) resting ECG results (abnormal ECG readings can signal underlying heart issues). These features were consistently ranked high in importance by both the XAI tools and domain experts, confirming their relevance and impact in predicting heart disease.

The current study has several significant implications for the field of cardiovascular risk assessment, ML applications in healthcare, and the broader domain of medical data analysis. The improved accuracy and stability of the ML models can lead to more reliable early detection of heart disease. The introduction of the novel data balancing technique contributes to the ML community by providing a method that addresses the limitations of existing oversampling and undersampling methods. The integration of XAI techniques and their validation against domain expert knowledge also sets a precedent for the importance of model interpretability in healthcare applications.

Despite the promising results, the study has some limitations. First, the proposed technique requires more computational resources compared to traditional methods, which may limit its applicability in resource-constrained environments. Second, the effectiveness of the proposed method depends on the quality and relevance of the features used. Poor feature selection may still lead to suboptimal model performance. Finally, while the method showed robustness on two datasets, additional testing on more diverse datasets is required to confirm its universal applicability. These limitations suggest areas for future research, including optimizing the computational efficiency of the proposed technique and exploring its performance on a wider range of datasets.

To further the progress made in enhancing model explainability, several specific future studies can be outlined. These studies could involve the exploration of different XAI techniques and their application to more diverse or complex datasets. The goal would be to deepen our understanding of model interpretability and its practical utility in various healthcare contexts. XAI techniques to explore include anchors (to generate high-precision rules that explain the model's decision, integrated gradients, counterfactual explanations), and DeepLIFT (deep learning important features, to compare the activation of each neuron to a reference activation). Other future research directions may focus on assessing the practical utility and impact of XAI techniques when integrated into real-world clinical workflows, and creating and validating domain-specific XAI techniques tailored to particular medical specialties. These could address temporal explanations of patient monitoring data or hierarchical explanations of models that utilize multi-level data, such as patient history combined with real-time monitoring.

6. Conclusion and future work

This investigation introduces a novel data-balancing approach for forecasting the risk of heart disease from two established datasets. We suggest a preprocessing scheme that boosts the predictive potential of the classifier in assessing heart disease risk. To address the problem of performance incongruity among inter-dataset setups, we have developed a preprocessing pipeline supplemented with diverse methods and our data balancing techniques. In relation to the proposed data balancing stage, we utilize LR and RF ML algorithms to assess the classifiers' performance within the imbalanced dataset. The precision and recall of the classifiers could be enhanced, while their level also needs improvement in both classes. Subsequently, we apply three established data-balancing techniques and evaluate the classifier's performance. The overall performance sometimes declines, but precision and recall prove more robust. The results demonstrate that the suggested data balancing techniques are superior to the three alternative approaches and that RF presents greater accuracy than LR. The proposed technique yielded an F1 score increase from 0.61 to 0.69, and AUC improved from 0.801 to 0.833. Specifically, our method achieved an accuracy of 93% and an AUC of 0.97 when using RF on the Long Beach VA dataset, significantly outperforming traditional techniques such as SMOTE and NearMiss.

In the second phase, we predict the occurrence of heart disease using two secondary datasets and utilize benchmark ML classifiers for classification. In most cases, the performance of the classifiers does not meet satisfactory levels. Subsequently, we assess performance discrepancy through the use of one dataset for training and another for testing. Results reveal a decrease in classifier performance, signifying the existence of performance discrepancy. Using our proposed data preprocessing pipeline, we observe a significant boost in classifier performance for both cases. For example, when comparing the performance metrics on the Hungarian and Long Beach VA datasets, the accuracy improved from 79% to 90%, and the AUC increased from 0.89 to 0.96 with KNN classifier.

Subsequently, we utilize XAI tools SHAP and Shapash on the datasets to determine the explainability of the top-performing model, RF. Subsequently, we utilize XAI tools SHAP and Shapash on the datasets to determine the explainability of the top-performing model, RF. We obtain a feature hierarchy and individually assess their contributions to the model's performance. In addition to the global explainability, we also examine the model's local explainability. Then the survey results indicate that the features in XAI-driven explainability and the domain experts' explanation fall within the same category.

This study centers on the binary classification problem and presents a suitable data balancing technique for binary classification. Subsequent research will examine multi-label classification in different domains. We intend to incorporate existing heart disease datasets in the future, creating a global model and dataset that can accurately predict heart disease in real-time using wearable sensor data from the human body. Further research will explore the discrepancy between domain experts and XAI explainability, and extend to discovering more personalized and targeted interpretable machine learning methods such as those developed by Jang, Kim, and Yoon (2023), Rajpal et al. (2023) and Bonifazi et al. (2024).

CRediT authorship contribution statement

Fan Yang: Conceptualization, Methodology, Data curation, Investigation, Supervision, Software, Writing – original draft, Writing – review & editing, Funding acquisition. **Yanan Qiao:** Conceptualization, Methodology, Data curation, Investigation, Software, Writing – original draft, Funding acquisition. **Petr Hajek:** Methodology, Data curation, Investigation, Supervision, Writing – original draft, Validation, Visualization. **Mohammad Zoynul Abedin:** Conceptualization, Writing – review & editing, Formal analysis, Methodology, Validation, Visualization.

Declaration of competing interest

There is no competing interest among the authors.

Data availability

Data will be made available on request.

Acknowledgments

We will be grateful to the anonymous reviewers who will comment on this manuscript.

Funding

This research is supported by the Natural Science Basic Research Program of Shaanxi [Program No. 2023-JC-YB-490]. This research is also supported by the Research Fund of Guangxi Key Lab of Multi-source Information Mining & Security (MIMS24-06). This research is also supported by "the Fundamental Research Funds for the Central Universities, JLU" (93K172024K12).

References

Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, *128*, Article 102289.

Ali, N., Neagu, D., & Trundle, P. (2019). Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Applied Sciences*, *1*, 1–15.

Alizadehsani, R., Roshanzamir, M., Abdar, M., Beykikhoshk, A., Khosravi, A., Panahiazar, M., et al. (2019). A database for using machine learning and data mining techniques for coronary artery disease diagnosis. *Scientific Data*, *6*(1), 227.

Allgaier, J., Mulansky, L., Draelos, R. L., & Pryss, R. (2023). How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. *Artificial Intelligence in Medicine*, *143*, Article 102616.

Alshraideh, M., Alshraideh, N., Alshraideh, A., Alkayed, Y., Al Trabsheh, Y., & Alshraideh, B. (2024). Enhancing heart attack prediction with machine learning: A study at jordan university hospital. *Applied Computational Intelligence and Soft Computing*, *2024*(1), Article 5080332.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., et al. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access*, *4*, 7940–7957.

Aswad, S. A., & Sonuç, E. (2020). Classification of VPN network traffic flow using time related features on apache spark. In *2020 4th international symposium on multidisciplinary studies and innovative technologies* (pp. 1–8). IEEE.

Azar, A. T., Elshazly, H. I., Hassanien, A. E., & Elkorany, A. M. (2014). A random forest classifier for lymph diseases. *Computer Methods and Programs in Biomedicine*, *113*(2), 465–473.

Azmi, J., Arif, M., Nafis, M. T., Alam, M. A., Tanweer, S., & Wang, G. (2022). A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. *Medical Engineering & Physics*, Article 103825.

Baniecki, H., Parzych, D., & Biecek, P. (2023). The grammar of interactive explanatory model analysis. *Data Mining and Knowledge Discovery*, 1–37.

Bao, L., Juan, C., Li, J., & Zhang, Y. (2016). Boosted near-miss under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets. *Neurocomputing*, *172*, 198–206.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29.

Bonifazi, G., Cauteruccio, F., Corradini, E., Marchetti, M., Terracina, G., Ursino, D., et al. (2024). A model-agnostic, network theory-based framework for supporting XAI on classifiers. *Expert Systems with Applications*, *241*, Article 122588.

Brito, M. P., Chen, Z., Wise, J., & Mortimore, S. (2022). Quantifying the impact of environment factors on the risk of medical responders' stress-related absenteeism. *Risk Analysis*, *42*(8), 1834–1851.

Campillo-Artero, C., Serra-Burriel, M., & Calvo-Pérez, A. (2018). Predictive modeling of emergency cesarean delivery. *PLoS One*, *13*(1), Article e0191248.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning* (pp. 161–168).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface*, *15*(141), Article 20170387.

Cutrì, E., Meoli, A., Dubini, G., Migliavacca, F., Hsia, T.-Y., & Pennati, G. (2017). Patient-specific biomechanical model of hypoplastic left heart to predict post-operative cardio-circulatory behaviour. *Medical Engineering & Physics*, *47*, 85–92.

Dalal, S., Goel, P., Onyema, E. M., Alharbi, A., Mahmoud, A., Algarni, M. A., et al. (2023). Application of machine learning for cardiovascular disease risk prediction. *Computational Intelligence and Neuroscience*, *2023*(1), Article 9418666.

Das, S., Sultana, M., Bhattacharya, S., Sengupta, D., & De, D. (2023). XAI–reduct: accuracy preservation despite dimensionality reduction for heart disease classification using explainable AI. *Journal of Supercomputing*, 1–31.

Dave, D., Naik, H., Singhal, S., & Patel, P. (2020). Explainable AI meets healthcare: A study on heart disease dataset. arXiv:2011.03195.

Deepak, S., & Ameer, P. (2019). Brain tumor classification using deep CNN features via transfer learning. *Computers in Biology and Medicine*, *111*, Article 103345.

Dhanabal, L., & Shantharajah, S. (2015). A study on NSL-KDD dataset for intrusion detection system based on classification algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(6), 446–452.

Dhar, T., Dey, N., Borra, S., & Sherratt, R. S. (2023). Challenges of deep learning in medical image analysis—Improving explainability and trust. *IEEE Transactions on Technology and Society*, *4*(1), 68–75.

Douzas, G., & Bacao, F. (2018). Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*, *91*, 464–471.

Fernandes, K., Cardoso, J., & Fernandes, J. (2017). Cervical cancer (Risk Factors). http://dx.doi.org/10.24432/C5Z310, UCI Machine Learning Repository.

Ferreira, L., Pilastri, A., Martins, C. M., Pires, P. M., & Cortez, P. (2021). A comparison of automl tools for machine learning, deep learning and xgboost. In *2021 international joint conference on neural networks* (pp. 1–8). IEEE.

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining: vol. 72*, Springer.

Guleria, P., Naga Srinivasu, P., Ahmed, S., Almusallam, N., & Alarfaj, F. K. (2022). XAI framework for cardiovascular disease prediction using classification techniques. *Electronics*, *11*(24), http://dx.doi.org/10.3390/electronics11244086, URL https://www.mdpi.com/2079-9292/11/24/4086.

Gwetu, M. V., Tapamo, J.-R., & Viriri, S. (2020). Random forests with a steepend gini-index split function and feature coherence injection. In *International conference on machine learning for networking* (pp. 255–272). Springer.

Hasan, M., Islam, M. M., Sajid, S. W., & Hassan, M. M. (2022). The impact of data balancing on the classifier's performance in predicting cesarean childbirth. In *2022 4th international conference on electrical, computer & telecommunication engineering* (pp. 1–4). IEEE.

Highnam, K., Arulkumaran, K., Hanif, Z., & Jennings, N. R. (2021). Beth dataset: Real cybersecurity data for anomaly detection research. *Training*, *763*(66.88), 8.

Jang, H., Kim, S., & Yoon, B. (2023). An eXplainable AI (XAI) model for text-based patent novelty analysis. *Expert Systems with Applications*, *231*, Article 120839.

Japkowicz, N. (2001). Supervised versus unsupervised binary-learning by feedforward neural networks. *Machine Learning*, *42*(1–2), 97.

Jiang, Z., Cui, X., Qu, P., Shang, C., Xiang, M., & Wang, J. (2022). Roles and mechanisms of puerarin on cardiovascular disease: A review. *Biomedicine & Pharmacotherapy*, *147*, Article 112655.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(1), 559–563.

Leung, K. M., et al. (2007). Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, *2007*, 123–156.

Lin, W.-W., Mak, M.-W., Li, L., & Chien, J.-T. (2018). Reducing domain mismatch by maximum mean discrepancy based autoencoders.. In *Odyssey* (pp. 162–167).

Liu, T., Fan, W., & Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artificial Intelligence in Medicine*, *101*, Article 101723.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

Mahmood, T., Rehman, A., Saba, T., Nadeem, L., & Bahaj, S. A. O. (2023). Recent advancements and future prospects in active deep learning for medical image segmentation and classification. *IEEE Access*, *11*, 113623–113652.

Malangsa, R. D., & Maravillas, E. A. (2017). Performance comparison of naïve bayes and K-NN algorithms on contamination grading for abaca tissue culture (in vitro). *International Journal of Computer Science & Information Technology*, *5*, 5–10.

Marabelli, M., Vaast, E., & Li, J. L. (2021). Preventing the digital scars of COVID-19. *European Journal of Information Systems*, *30*(2), 176–192.

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible ai and 'the dark side' of AI.

Moravvej, S. V., Alizadehsani, R., Khanam, S., Sobhaninia, Z., Shoeibi, A., Khozeimeh, F., et al. (2022). RLMD-PA: a reinforcement learning-based myocarditis diagnosis combined with a population-based algorithm for pretraining weights. *Contrast Media & Molecular Imaging*, *2022*(1), Article 8733632.

Moustafa, N., & Slay, J. (2015). UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 military communications and information systems conference* (pp. 1–6). IEEE.

Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., & Mosavi, A. (2020). Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. *IEEE Access*, *8*, 150199–150212.

Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine learning technology-based heart disease detection models. *Journal of Healthcare Engineering*, *2022*.

Park, S., & Park, H. (2021). Combined oversampling and undersampling method based on slow-start algorithm for imbalanced network traffic. *Computing*, *103*(3), 401–424.

Patil, S., & Bhosale, S. (2021). Hyperparameter tuning based performance analysis of machine learning approaches for prediction of cardiac complications. In *International conference on soft computing and pattern recognition* (pp. 605–617). Springer.

Pecorelli, F., Di Nucci, D., De Roover, C., & De Lucia, A. (2019). On the role of data balancing for machine learning-based code smell detection. In *Proceedings of the 3rd ACM SIGSOFT international workshop on machine learning techniques for software quality evaluation* (pp. 19–24).

Pecorelli, F., Di Nucci, D., De Roover, C., & Lucia, D. (2020). A large empirical assessment of the role of data balancing in machine-learning-based code smell detection. *Journal of Systems and Software*, *169*, Article 110693.

Prakash, K. B., & Kanagachidambaresan, G. (2021). Pattern recognition and machine learning. In *Programming with tensorflow: Solution for edge computing applications* (pp. 105–144). Springer.

Rajkumar, G., Devi, T. G., & Srinivasan, A. (2022). Heart disease prediction using IoT based framework and improved deep learning approach: Medical application. *Medical Engineering & Physics*, Article 103937.

Rajpal, S., Rajpal, A., Saggar, A., Vaid, A. K., Kumar, V., Agarwal, M., et al. (2023). XAI-MethylMarker: Explainable AI approach for biomarker discovery for breast cancer subtype classification using methylation data. *Expert Systems with Applications*, *225*, Article 120130.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data* (pp. 427–438).

Reis, I., Baron, D., & Shahaf, S. (2018). Probabilistic random forest: A machine learning algorithm for noisy data sets. *Astronomical Journal*, *157*(1), 16.

Rokach, L., & Maimon, O. (2005). Decision trees. In *Data mining and knowledge discovery handbook* (pp. 165–192). Springer.

Sagi, O., & Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. *Information Sciences*, *572*, 522–542.

Sahid, M. A., Hasan, M., Akter, N., & Tareq, M. M. R. (2022). Effect of imbalance data handling techniques to improve the accuracy of heart disease prediction using machine learning and deep learning. In *2022 IEEE region 10 symposium* (pp. 1–6). IEEE.

Sakar, C. O., Serbes, G., Gunduz, A., Tunc, H. C., Nizam, H., Sakar, B. E., et al. (2019). A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. *Applied Soft Computing*, *74*, 255–263.

Sarumi, O. A., & Leung, C. K. (2022). Adaptive machine learning algorithm and analytics of big genomic data for gene prediction. In *Tracking and preventing diseases with artificial intelligence* (pp. 103–123). Springer.

Sokoliuk, A., Kondratenko, G., Sidenko, I., Kondratenko, Y., Khomchenko, A., & Atamanyuk, I. (2020). Machine learning algorithms for binary classification of liver disease. In *2020 IEEE international conference on problems of infocommunications. science and technology (PIC s&t)* (pp. 417–421). IEEE.

Srinivasu, P. N., Sandhya, N., Jhaveri, R. H., & Raut, R. (2022). From blackbox to Explainable AI in healthcare: Existing tools and case studies. In S. Hakak (Ed.), *Mobile Information Systems*, *2022*, Article 8167821. http://dx.doi.org/10.1155/2022/8167821, Publisher: Hindawi.

Sumwiza, K., Twizere, C., Rushingabigwi, G., Bakunzibake, P., & Bamurigire, P. (2023). Enhanced cardiovascular disease prediction model using random forest algorithm. *Informatics in Medicine Unlocked*, *41*, Article 101316.

Tarawneh, A. S., Hassanat, A. B., Altarawneh, G. A., & Almuhaimeed, A. (2022). Stop oversampling for class imbalance learning: A review. *IEEE Access*, *10*, 47643–47660.

Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019). A brief review of nearest neighbor algorithm for learning and classification. In *2019 international conference on intelligent computing and control systems* (pp. 1255–1260). IEEE.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*.

Thomas, R. M., Bruin, W., Zhutovsky, P., & van Wingen, G. (2020). Dealing with missing data, small sample sizes, and heterogeneity in machine learning studies of brain disorders. In *Machine learning* (pp. 249–266). Elsevier.

Tiwari, D., Bhati, B. S., Al-Turjman, F., & Nagpal, B. (2022). Pandemic coronavirus disease (Covid-19): World effects analysis and prediction using machine-learning techniques. *Expert Systems*, *39*(3), Article e12714.

Uddin, M. M., Rashid, M. M., Hasan, M., Hossain, M. A., & Fang, Y. (2022). Investigating corporate environmental risk disclosure using machine learning algorithm. *Sustainability*, *14*(16), 10316.

Wallace, J., Mullarkey, M. T., & Hevner, A. (2023). Patient health locus of control: the design of information systems for patient-provider interactions. *European Journal of Information Systems*, *32*(1), 52–63.

Weissler, E. H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., et al. (2021). The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, *22*(1), 1–15.

Woźniak, M., Wieczorek, M., & Siłka, J. (2023). BiLSTM deep neural network model for imbalanced medical data of IoT systems. *Future Generation Computer Systems*, *141*, 489–499.

Wu, Y., & Fang, Y. (2020). Stroke prediction with machine learning methods among older Chinese. *International Journal of Environmental Research and Public Health*, *17*(6), 1828.

Zhang, C. A., Cho, S., & Vasarhelyi, M. (2022). Explainable Artificial Intelligence (XAI) in auditing. *International Journal of Accounting Information Systems*, *46*, Article 100572.

Zhang, K., Xu, P., & Zhang, J. (2020). Explainable AI in deep reinforcement learning models: A shap method applied in power system emergency control. In *2020 IEEE 4th conference on energy internet and energy system integration* (pp. 711–716). IEEE.

Zhang, H., Zhang, H., Pirbhulal, S., Wu, W., & Albuquerque, V. H. C. D. (2020). Active balancing mechanism for imbalanced medical data in deep learning–based classification models. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *16*(1s), 1–15.

Zhou, Q., Li, S., Li, X., Wang, W., & Wang, Z. (2006). Detection of outliers and establishment of targets in external quality assessment programs. *Clinica Chimica Acta*, *372*(1–2), 94–97.

Zhu, R., Hu, X., Hou, J., & Li, X. (2021). Application of machine learning techniques for predicting the consequences of construction accidents in China. *Process Safety and Environmental Protection*, *145*, 293–302.