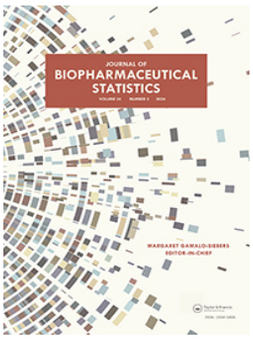

Taylor & Francis
Taylor & Francis Group

Journal of Biopharmaceutical Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/lbps20

Revolutionizing cardiovascular disease classification through machine learning and statistical methods

Tapan Kumar Behera, Siddhartha Sathia, Sibarama Panigrahi & Pradeep Kumar Naik

To cite this article: Tapan Kumar Behera, Siddhartha Sathia, Sibarama Panigrahi & Pradeep Kumar Naik (24 Nov 2024): Revolutionizing cardiovascular disease classification through machine learning and statistical methods, Journal of Biopharmaceutical Statistics, DOI: 10.1080/10543406.2024.2429524

To link to this article: https://doi.org/10.1080/10543406.2024.2429524

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
https://doi.org/10.1080/10543406.2024.2429524

Revolutionizing cardiovascular disease classification through machine learning and statistical methods

Tapan Kumar Behera [a], Siddhartha Sathia [b], Sibarama Panigrahi [c], and Pradeep Kumar Naik [a]

[a]Centre of Excellence in Natural Products and Therapeutics, Department of Biotechnology and Bioinformatics, Sambalpur University, Jyoti Vihar, Burla, Sambalpur, Odisha, India; [b]Department of Cardiothoracic Surgery (CTVS), All India Institute of Medical Sciences, Sijua, Patrapada, Bhubaneswar, Odisha, India; [c]Department of Computer Science & Engineering (CSE), National Institute of Technology, Rourkela, Odisha, India

ABSTRACT

Background: Cardiovascular diseases (CVDs) include abnormal conditions of the heart, diseased blood vessels, structural problems of the heart, and blood clots. Traditionally, CVD has been diagnosed by clinical experts, physicians, and medical specialists, which is expensive, time-consuming, and requires expert intervention. On the other hand, cost-effective digital diagnosis of CVD is now possible because of the emergence of machine learning (ML) and statistical techniques.

Method: In this research, extensive studies were carried out to classify CVD via 19 promising ML models. To evaluate the performance and rank the ML models for CVD classification, two benchmark CVD datasets are considered from well-known sources, such as Kaggle and the UCI repository. The results are analysed considering individual datasets and their combination to assess the efficiency and reliability of ML models on the basis of various performance measures, such as precision, kappa, accuracy, recall, and the F1 score. Since some of the ML models are stochastic, we repeated the simulation 50 times for each dataset using each model and applied nonparametric statistical tests to draw decisive conclusions.

Results: The nonparametric Friedman – Nemenyi hypothesis test suggests that the Extra Tree Classifier provides statistically superior accuracy and precision compared with all other models. However, the Extreme Gradient Boost (XGBoost) classifier provides statistically superior recall, kappa, and F1 scores compared with those of all the other models. Additionally, the XGBRF classifier achieves a statistically second-best rank in terms of the recall measures.

ARTICLE HISTORY
Received 19 December 2023
Accepted 9 November 2024

KEYWORDS
Cardiovascular disease (CVD); machine learning (ML); artificial intelligence (AI); artificial neural network (ANN); adaptive boosting (ADB); bagging classifier (BC); decision tree (DT); extra trees (ETC); gradient boosting (GB); Gaussian Naïve Bayes (GNB); K-nearest neighbor (KNN); logistic regression (LR); linear support vector classifier (LSVC); multilayer perceptron (MLP); passive aggressive classifier (PAC); ridge classifier (RC); random forest (RF); stochastic gradient decent (SGD); support vector classifier (SVC); extra tree (TC); voting classifier (VC); extreme gradient boost (XGBoost); extreme gradient boost random forest (XGBRF)

1. Introduction

Global death due to cardiovascular disease (CVD) worldwide has peaked over the last few decades. Approximately 18 million people died in 2019, 85% of whom died because of CVD. Based on current trends and examination of pertinent data, the World Health Organization (WHO) has forecasts for the future of CVD up to 2030. A statistical report estimates that the global deaths due to CVD worldwide will increase to 24 million by 2030 (Tuli et al. 2020). This prediction shows a strong upward trend in mortality due to CVD. In clinical terms, CVD can be divided into various types: ischaemic heart disease or coronary artery disease, heart attack, stroke, arrhythmias,

CONTACT Sibarama Panigrahi panigrahi.sibarama@gmail.com Department of Computer Science & Engineering (CSE), National Institute of Technology, Rourkela, Odisha 769008, India; Pradeep Kumar Naik pknaik1973@gmail.com Centre of Excellence in Natural Products and Therapeutics, Department of Biotechnology and Bioinformatics, Sambalpur University, Jyoti Vihar, Burla, Sambalpur, Odisha 768019, India

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10543406.2024.2429524.

cardiomyopathy, rheumatic disease, congenital heart disease (CHD), vascular heart disease, and aortic disease (Chen et al. 2019). Various risk factors for CVD are related, which adds to the complexity of CVD diagnosis. For instance, both diabetes and obesity are significant, and obesity is related to the development of CVD in people with type 2 diabetes. Most CVD cases are caused by certain high-risk aspects like smoking, high blood pressure (high), age, high cholesterol, obesity, diabetes, alcohol consumption, and many other parameters. As a result, developing any form of the summative equation for predicting CVD risk on the basis of risk variables is problematic.

The procedure used to foresee and analyse CVDs largely depends on investigating cardiac patients' clinical records, indications and physical assessment records by specialists. Moreover, the current diagnostic methods lack early detection, integrate data, conduct point-of-care testing, and provide real-time diagnosis and timely prevention. Improving these aspects could increase the chances of reducing mortality from cardiovascular diseases. As a result, for reliable and precise CVD prediction in health industries, an automated intelligent system is needed. Notably, the type of raw information used can strongly influence the performance and quality of Artificial Intelligence (AI) approaches (Alizadehsani et al. 2021). The implementation of ML in the automated diagnosis of a disease involves (a) automated image analysis, which includes the imaging data of the patient, i.e., X-rays, MRIs, and CT scans (Madabhushi and Lee 2016), and (b) predictive analysis of ML models, which analyse patient data, including medical history, symptoms, laboratory tests, and genetic information, to predict the likelihood of certain diseases or conditions. This can help healthcare providers prioritize patients for further testing or intervention, leading to faster diagnosis and treatment (Ramesh et al. 2022), (c) Natural Language Processing (NLP) techniques can extract relevant information from unstructured text data, such as electronic health records, clinical notes, and medical literature. ML algorithms can analyse this information to assist healthcare providers in making faster and more accurate diagnoses by summarizing patient histories, identifying relevant research findings, and suggesting potential treatment options (Nagarhalli et al. 2021), (d) Decision Support System: ML-based decision support systems can assist healthcare providers in interpreting complex diagnostic tests, such as genetic sequencing, and recommend appropriate follow-up actions on the basis of patient-specific factors and clinical guidelines. These systems can help streamline the diagnostic process and reduce the time it takes to reach a diagnosis (Arena et al. 2022). (e) Remote Monitoring and Telemedicine: ML algorithms can analyse data from wearable devices, sensors, and remote monitoring systems to continuously monitor patients' health status and detect early signs of deterioration or disease progression. This enables timely intervention and reduces the need for frequent in-person visits, leading to faster diagnosis and treatment adjustments (Kumari et al. 2022). For cardiovascular disease (CVD) data, various machine learning (ML) models can be suitable depending on the specific task and the nature of the data. There are several widely used ML models, such as LR, RF, GB, SVM, MLP and DT for their better efficacy in identifying CVDs. These classifiers are most often used for binary classification and regression tasks and can handle mixed categorical, sequential, and complex clinical data. These algorithms are even much more effective, especially when dealing with high-dimensional data, and can handle nonlinear relationships through the kernel trick and provide a greater understanding of feature importance for CVD prediction (Ammar et al. 2021). CVD datasets are openly available at different portals for observing derived prognosis models. By implementing AI and ML techniques, experimenters can design predictive models that can better predict disease through the available datasets. The novelty of this study is to improve the performance of the ML classifiers in response to sensitivity, specificity, accuracy, precision, f1 and kappa measures. Deep learning (DL) focuses on artificial neural networks with multiple layers in it. Despite this, DL is a subset of ML that provides scope in the fields of image recognition, natural language processing and speech recognition. Choosing ML encompasses a broad range of techniques that enable computers to learn from the available data and make predictions or decisions based on the same data (Janiesch et al. 2021). The datasets are available on open-source platforms such as UCI, and the Kaggle science repository and are in numerical form. Despite there being a bunch of applications of ML for classification tasks, still no systematic studies have been made to evaluate the true potential of ML classifiers such as Support

Vector Classifier (SVC), Random Forest (RF), K-nearest neighbour (KNN), Logistic regression (LR), Decision tree (DT), Bagging, Ridge, Passive Aggressive, Gaussian Naïve Bayes (GNB), Extra Trees (ETC), voting Classifiers (VC), Extreme Gradient Boost (XGBoost), Extreme Gradient Boost Random Forest (XGBRF), ADABoost (ADB), Multilayer Perceptron (MLP), Stochastic Gradient Descent (SGD), Linear SVC (LSVC), Gradient Boost (GB) in classifying CVD (Ali et al. 2021; Pham et al. 2021; Swathy and Saruladha 2021). These algorithms will offer improvements in diagnostic accuracy and better healthcare delivery in the successful translation of these algorithms into clinical practice.

Additionally, since some of the ML classifiers (Decision Tree, Multilayer Perceptron, Bagging, Passive Aggressive, Stochastic Gradient Descent, Extra Trees) are stochastic, no systematic study has been conducted to evaluate these ML models by employing statistical analysis of the obtained results. This study uses a systematic analysis to estimate the true potential of stochastic and deterministic ML models employing bagging, boosting, and voting in classifying CVD. Two benchmark cardiovascular datasets are considered to make the analysis and conclusions reliable, and the obtained results are compared statistically using individual datasets and all datasets together.

2. Literature review

Maini et al. (2021) proposed an ML approach that effectively and economically predicts CVD. A total of 1670 unknown medical records were collected from hospitals and used to train the ML models. The authors considered the five best models as KNN, NB, LR, ADB, and RF using the python platform. Among these methods, the RF method achieves better accuracy, with a value of 94%. Statistical tests like, t-test and chi-square tests are carried out to determine the statistical significance of the differences between the models. Based on the result, it was inferred that male patients had a higher risk of CVD in comparison to females. This study portrays the true potential of ML algorithms in identifying CVD in the Indian population. In reference to this study, the author only deemed the best ML models but did not consider hybrid ML models such as stacking, voting, bagging, and boosting, which may improve the performance and lead to more decisive conclusions.

Sharma et al. (2020) proposed a modified artificial plant optimization (MAPO) algorithm and applied it to forecast the heart rate using feature selection together with other ML models. This study considered fingertip videos of patients aged between 20 and 50 years and a heart disease dataset from the Kaggle repository for identification of the presence of Coronary Heart Disease (CHD) in individuals. The ML algorithms used were Extreme Gradient Boosting (XGB), LR, GNB, and ANN. The authors achieved Pearson correlation and standard error of estimation scores of 0.95 and 2.41, respectively. The authors concluded that MAPO, with its optimal feature selection, acquires highest performance when tested over the CHD dataset.

Pires et al. (2020) proposed a singleton ML method to uniquely identify the heart disease in the population. In this study, the heart disease dataset is obtained from the open access UCI ML scientific repository. This dataset is used by a good number of ML models, such as Neural Network, DT, KNN, combined nomenclature (CN2) rule inducer, SVC, and SGD to automatically identify the disease accurately and efficiently. In this study, the authors reported that the early detection efficiency of CHD using the SGD model is 87.6%. This study can be expanded to include additional ML models, and by stacking, boosting, voting, and bagging, the accuracy in prediction of CVD can be achieved. Additionally, the authors used only a single CVD dataset for model evaluation, which may have resulted in biased conclusions towards the dataset. Hence, to obtain decisive conclusions, a greater number of CVD datasets and performance metrics need to be used along with the application of statistical tests to the obtained results.

Pavithra and Jayalakshmi (2021) presented a novel K-means neighbor classifier that provides better classification of heart disease than the other classifiers. In this work, the authors considered various ML classifiers, such as LR, GNB, ANN, DT, and KNN, which are implemented via the Python language. Four performance measures, namely, specificity, sensitivity, precision, and accuracy, are used to evaluate the models. By using this novel k-mean neighbor, the model achieved better accuracy than the other models did. In this study, despite the use of stochastic models such as ANNs and DTs, no statistical tests are employed on the obtained results to draw decisive conclusions. Additionally, only a single dataset with less data is used to evaluate the performance of the models, which may result in biased conclusions towards the dataset.

Table 1 summarizes the datasets, ML models, and highest accuracy achieved by the models in classifying CVD. ML models such as Support Vector Classifier (Ammar et al. 2021; Pires et al. 2020; Reddy et al. 2021; Swathy and Saruladha 2021), Decision Tree (Ali et al. 2021; Pavithra and Jayalakshmi 2021; Pires et al. 2020), Random Forest (Ali et al. 2021; Ammar et al. 2021; Jamthikar et al. 2020; Oh et al. 2021), Logistic Regression (Lip et al. 2021, Nathala et al. 2022), K-nearest Neighbour (Ali et al. 2021; Maini et al. 2021; Pavithra and Jayalakshmi 2021; Pires et al. 2020), Gaussian Naïve Bayes (Chen et al. 2019; Pavithra and Jayalakshmi 2021; Sharma et al. 2020), XGBoost (Pham et al. 2021; Sharma et al. 2020), ADABoost (Maini et al. 2021; Reddy et al. 2021), Extra Tree (Reddy et al. 2021), Multilayer Perceptron (MLP) (Ammar et al. 2021; Pavithra and Jayalakshmi 2021; Sharma et al. 2020) methods have been used in the

Table 1. Systematic approach to cardiovascular diagnosis using ML models.

Authors	Datasets	Methods	Accuracy
Chen et al. (2019)	98 Patients with Severe DCM from Two Centres	NB	88.7
Ali et al. (2021)	Heart Disease Dataset	KNN, DT and RF	100
Maini et al. (2021)	1670 Unknown Medical Records from Hospital	KNN, NB, LR, ADB, and RF	93.8
Reddy et al. (2021)	Heart Failure Dataset	Mel-Frequency Cepstral Coefficient (MFCC), SVC, ETC, ADB and Feed-Forward Neural Network (FFNN)	-
Sharma et al. (2020)	Heart Disease Dataset (kaggle)	MAPO (Modified Artificial Plant Optimization), LR, NB, XGBoost, Artificial Neural Network	-
Lip et al. (2021)	2,80,592 Cohort patients with COVID-19 cases from Hospital	LR	72.9
Nathala et al. (2022)	700 CVD patients Data	Multiple Logistic Regression	-
Pires et al. (2020)	UCI Heart Disease Dataset	Neural Network, DT, KNN, Combined Nomenclature Rule Inducer, SVC, SGD	87.6
Pavithra et al. (Razeghi et al. 2020)	UCI Heart Disease Dataset	HRFLC (Random Forest + AdaBoost + Pearson Coefficient)	-
Uddin et al. (Hagan et al. 2021)	Long Beach Heart Disease datasets	RF, NB, GB, KNN	99
Hagan et al. (Swathy and Saruladha 2021)	Cardiac Arrhythmia, Cardiovascular disease Dataset	SVC, MLP, DT	96
Park et al. (Oh et al. 2021)	4019 Cohort patients' data	LR, Classification and Regression Tree (CART), RF and Conditional Inference Tree (CIT)	71.3
Oh et al. (Pham et al. 2021)	Korean Cohort Dataset with Chronic Kidney Disease Patients data	MLP, XGB, LR, SVC and RF	-
Ghosh et al. (Mohan et al. 2019)	Heart Disease Cleveland,Long_Beach VA, Switzerland, Hungarian, Stat-log Dataset	DTBM, RFBM, KNNBM, ABBM, GBBM	99.05
Mohan et al. (Bertsimas et al. 2021)	Heart Disease Dataset Cleveland	HRFLM- Hybrid Random Forest Linear Model	88.7

literature to classify the CVD. The authors used the Heart Failure (Razeghi et al. 2020; Reddy et al. 2021), Cardiovascular Study, Cardiovascular Disease (Ali et al. 2021; Pavithra and Jayalakshmi 2021; Pires et al. 2020; Sharma et al. 2020; Swathy and Saruladha 2021), and Heart Statlog (Bertsimas et al. 2021; Hagan et al. 2021; Tuli et al. 2020) datasets and employed a fixed ratio of training and test percentages. Although some ML models are stochastic in nature, only some authors have applied statistical tests to obtain results. Hence, a robust systematic study on evaluating the efficiency of different ML models for CVD classification is still lacking in the literature. In this study, we statistically evaluated the performance of 19 ML models for classifying CVD by employing four benchmark datasets and five accuracy measures.

3. Contribution

The handout of this paper is outlined as follows:

(i) We statistically compared the true potential of 19 major ML models in classifying CVD. We have additionally ranked the models on the basis of different performance measures.
(ii) Two benchmark cardiovascular datasets, namely, the heart failure dataset (Razeghi et al. 2020; Reddy et al. 2021) and Heart Statlog Cleveland Hungary dataset (Bertsimas et al. 2021; Hagan et al. 2021; Tuli et al. 2020) having diverse characteristics and features, are considered.
(iii) The five performance metrics, precision, recall, accuracy, Cohen's kappa coefficient and F1 score, were considered for better analysis of the CVD classification performance of the considered models.
(iv) Since some of the ML models are stochastic in nature, the simulations are repeated 50 times on each dataset using each model, and extensive statistical analysis of the derived results is carried out to make a firm decision. The Friedman-Nemenyi hypothesis and Wilcoxon Signed Rank tests are used to statistically distinguish the models considering individual datasets and all datasets, respectively.

4. Materials and methodology

A total of 19 ML models were used in this study to evaluate their performance in classifying CVD. It has three subsections, (1) data preprocessing, (2) ML classifiers, and (3) statistical analysis, which are explained in Figure 1.

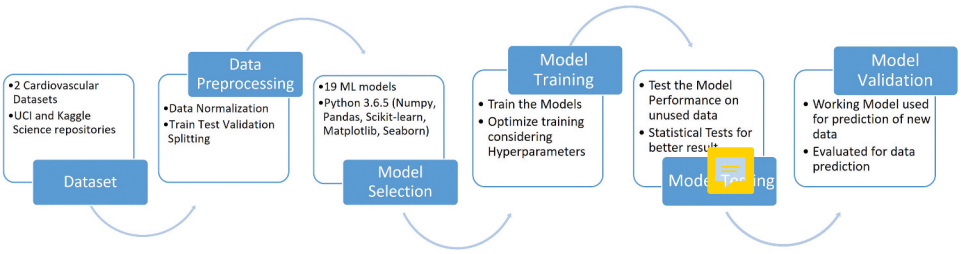

Figure 1. Flow diagram of the study.

4.1. Feature importance

Feature importance refers to the significance or contribution of each feature (or independent variable) towards predicting the target variable (or dependent variable). It helps in understanding which features have the greatest influence on the model's predictions and can be crucial for feature selection, model interpretation, and identifying key factors in the classification outcome. The pairplot in Figure 2 shows a grid of scatter plots and histograms for each pair of variables in the dataset, which indicates the correlations and distributions among the dataset features. This plot revealed the relationships among all the features when sex was considered a marker. The diagonal displays a kernel density estimate (KDE) of each feature. By involving the feature importance plots such as pair plots and heatmaps, the significant relationships among the features can be found and elaborated.

In Figure 2(a), the pairplot illustrates the mortality attributed to Heart_Failure_Dataset on the basis of the following features: age, anaemia (0: absent, 1: present), creatinine_phosphokinase, ejection_-fraction, platelets, serum_creatinine, serum_sodium, and time. Considering Death_Event as a marker, we draw the feature correlation in the dataset.

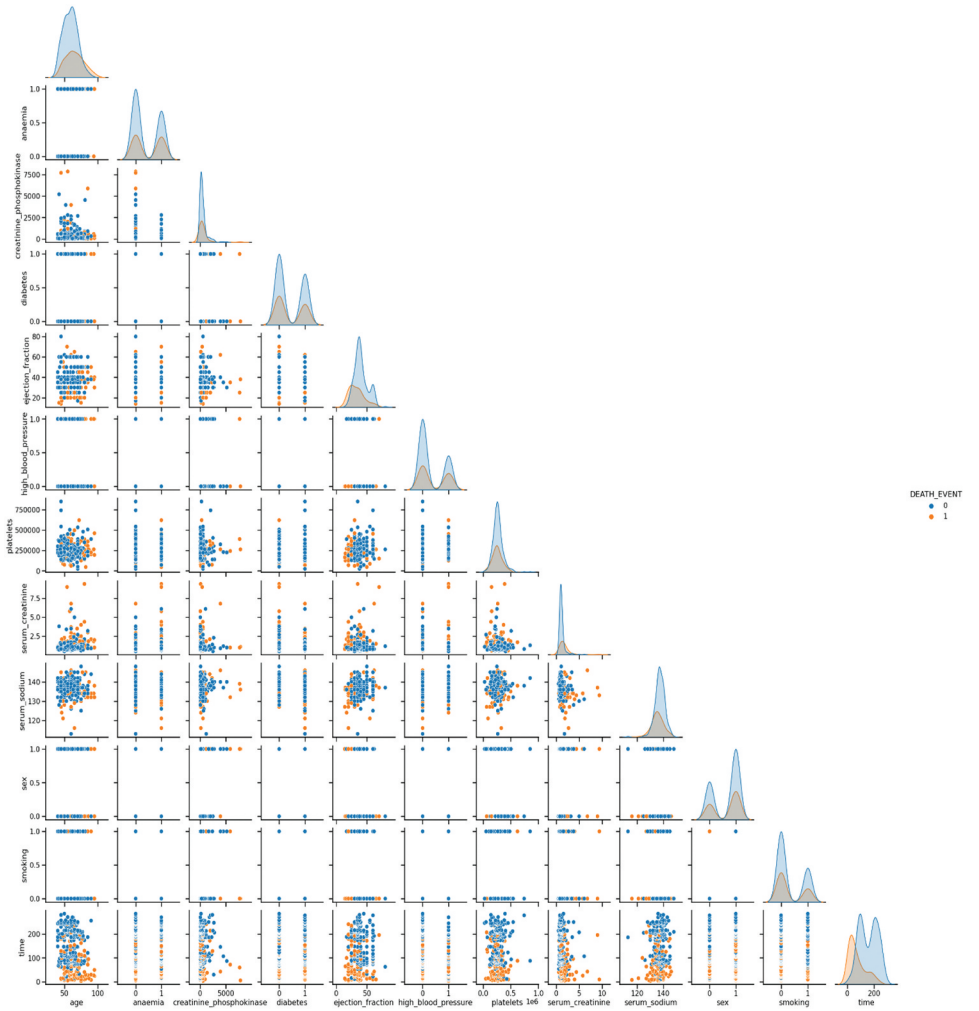

Figure 2. (a) Potential correlation between the features of the Heart_Failure dataset.

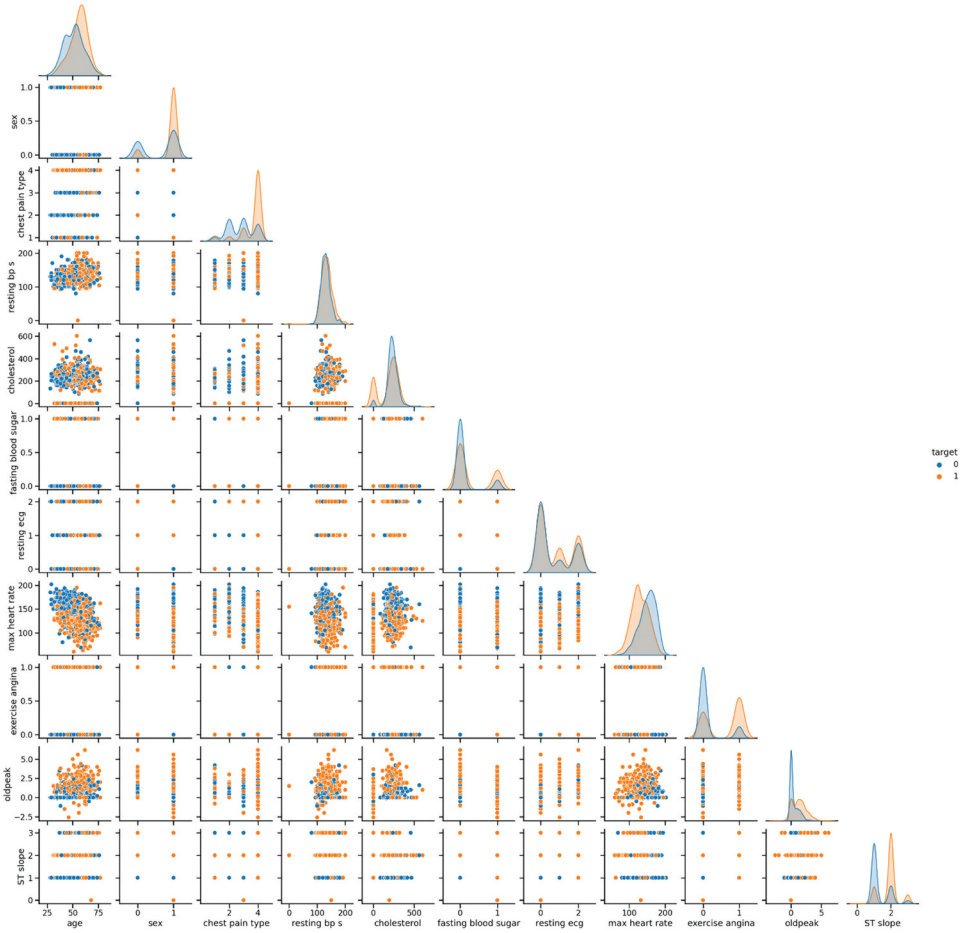

Figure 2b. Potential correlation between the features of the Heart_Statlog_Cleveland_Hungary dataset.

The pairplot illustrates the mortality attributed to Heart_Statlog_Cleveland_Hungary_Dataset for the features, namely, age, sex, chest_pain_type (1: stable angina, 2: unstable angina, 3: microvascular angina, 4: variant angina), resting_bp_s, cholesterol, fasting_blood_sugar (0: absent, 1: present), resting_ecg, max_heart_rate, exercise_angina (0: no, 1: yes), olpeak, and ST_slope. In Figure 2(b), by considering the target as a marker, the feature correlation in the dataset is determined.

Figure 3(a) It illustrates the relationship between the features and their interconnections, aiding in the identification of the most crucial attributes within the Heart_Failure dataset. The generated values exhibited a lower level of dependency on each other. Figure 3(a) shows that sex and smoking are highly correlated with each other at 0.45, whereas time and death_event are least correlated with each other at −0.53.

Figure 3(b), it illustrates the relationship between the features and their interconnections, aiding in the identification of the most crucial attributes within the Heart_Statlog_Cleveland_Hungary_Dataset. The generated values exhibited a lower level of dependency on each other. Here, the instances, namely, the old peak and ST_slope, have a correlation of 0.52 with each other, whereas max_heart_rate has the least correlation with almost all instances.

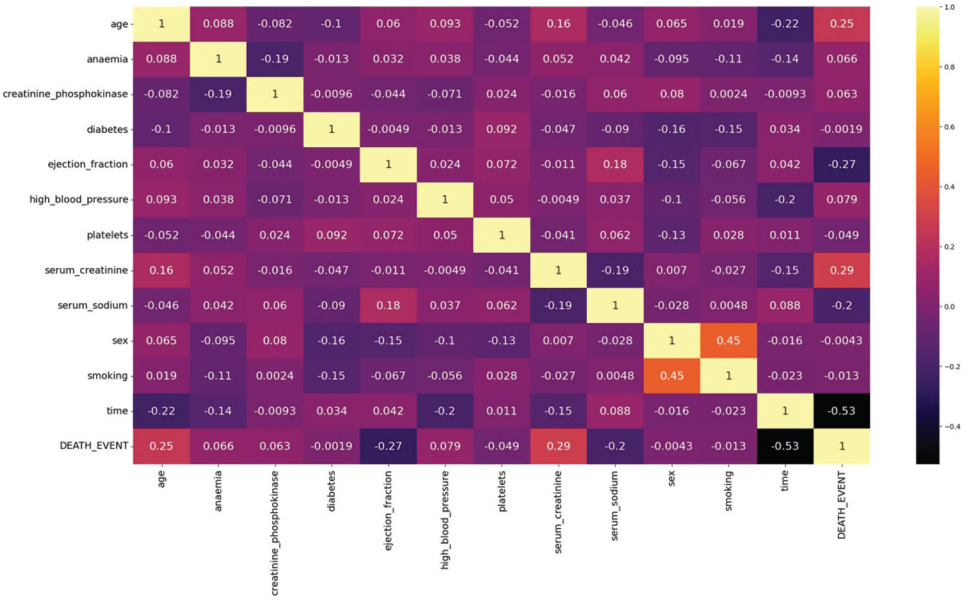

Figure 3a. Heatmap showing the correlations between the instances in the Heart_Failure dataset.

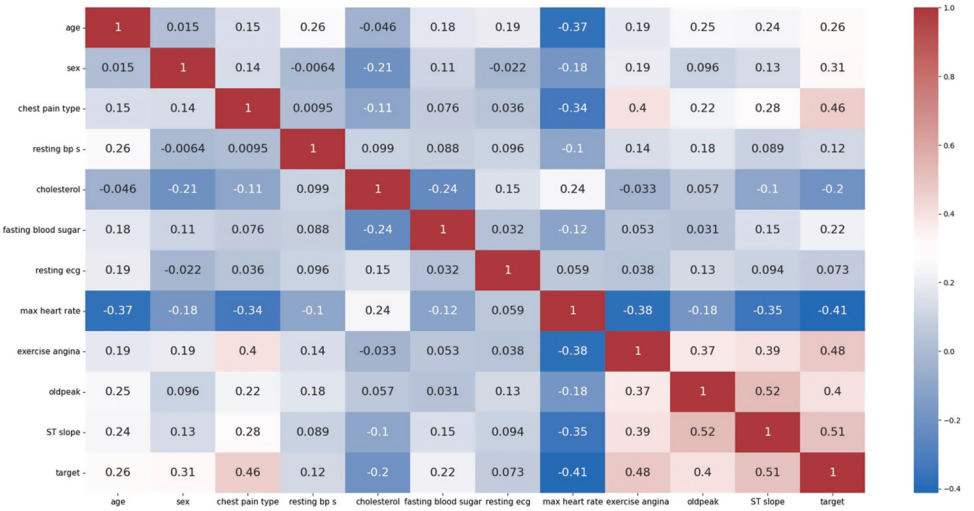

Figure 3b. Heatmap showing the correlation between the instances of the Heart_Statlog_Cleveland_Hungary dataset.

4.2. Data handling

The ability to handle missing data in machine learning models depends on the nature of the missingness (MCAR, MAR, or MNAR) and the specific algorithm being used.

4.2.1. Missing completely at random (MCAR)

In the case of MCAR data, one common and appropriate approach is to remove the rows containing missing values (known as listwise deletion). This method avoids introducing bias if

the missingness is genuinely random. Alternatively, imputation techniques such as the mean or median imputation, or the filling of missing values with a constant (e.g., zero), can also be employed.

4.2.2. Missing at random (MAR)

MAR data often require the use of multiple imputation techniques. This process entails generating multiple complete versions of the dataset by imputing missing values repeatedly. Each complete dataset is then analysed separately, and the results are combined. Various techniques, such as the mean or median imputation, as well as more advanced methods, such as k-nearest neighbors (KNN) imputation, can be utilized for this purpose.

4.2.3. Missing not at random (MNAR)

Handling MNAR data is more challenging because the missingness is related to unobserved variables. Sensitivity analysis can evaluate how different assumptions about the missing data mechanism might impact the results. Model-based imputation techniques can also be effective, using a model to estimate missing values on the basis of observed data. However, these models must consider the underlying missing data mechanism to be accurate.

Moreover, certain machine learning algorithms are naturally resilient to missing data, whereas others may need preprocessing to manage missing values. Tree-based algorithms such as Random Forests and Gradient Boosting Machines can internally handle missing values by utilizing surrogate splits.

4.2.3.1. Imputation. Imputation involves filling in missing values with estimated values. The common techniques are mean/median imputation, mode imputation, constant imputation, and interpolation.

4.2.3.2. Deletion. Deleting missing values can be done either for rows or columns with missing data.

4.2.3.3. Predictive imputation. Predictive imputation methods involve the use of machine learning algorithms to predict missing values on the basis of other features:

K-nearest neighbor imputation: Predict missing values by averaging the values of the nearest neighbors.

Regression imputation: Predicting missing values using regression models trained on the non-missing values of the feature and other features.

4.2.3.4. Advanced imputation technique. More advanced imputation methods involve utilizing the structure of the data and relationships between features: matrix factorization and multiple imputation.

4.2.3.5. Masking. For neural networks and deep learning models, masking layers can be used to ignore missing values during training.

4.3. Data preprocessing

Data preprocessing is required to clean the raw data and make it valuable for the ML model, increasing its efficiency. It is the first and crucial step in which missing and categorical data are processed and broken down into training sets, and testing sets and feature scaling are carried out. In our methodology, we normalize and split the data before processing them through ML models. Data normalization of variables is required when we perform a multivariate analysis, and each variable is expected to contribute equally to the result. The data were normalized to values between 0 and 1 using Equation (1).

$$x_i^{'} = (x_i - x_{min})/(x_{max} - x_{min}) \quad (1)$$

where
$x_i^{'}$: is the ith normalized value of the taken dataset.
$\boldsymbol{x_i}$: is the ith value of the taken dataset.
$\boldsymbol{x_{max}}$: is the maximum value of the taken dataset.
$\boldsymbol{x_{min}}$: is the minimum value of the taken dataset.

4.4. Machine learning classifiers

In ML, a classifier is a program that categorizes the data autonomously into one or more sets of "classes". In this research, we consider 19 promising ML classifiers, which are briefly explained below.

4.4.1. Decision tree classifier (DT)

The Decision Tree Classifier is one of the most promising classifiers, which recursively divides the input data into two or more branches to construct a tree. The parameters used were "gini" for the Gini impurity, best "random" split, the minimum number of samples required at the leaf node was "2", and the minimum number of samples required at split was "5" for binary classification.

4.4.2. K-nearest neighbor classifier (KNN)

KNN is a simple and optimistic supervised ML algorithm used to resolve classification and regression problems (Hambali et al. 2019). The parameters used were, as follows: nearest neighbors point "5", a leaf size of 30 for speed up construction, and a power parameter of 2 for the Minkowski metric using the Euclidian distance. To compute the nearest neighbors, the ball_tree algorithm was used.

4.4.3. Support vector machine classifier (SVC)

SVC is the most encountered supervised ML algorithm that thrives well for classification and regression problems (Ammar et al. 2021; Skandha et al. 2020). This study considers the 'linear' kernel version with regularization parameters C = 1 and gamma = 'scale', and the decision function shape is "ovr: one vs. rest" for the binary classification strategy.

4.4.4. Multilayer perceptron classifier (MLP)

The MLP model is a supervised Feed Forward Artificial Neural Network applied for classification tasks. The MLP consists of three layers first: an input layer; second, a hidden layer, and finally: an output layer (Ammar et al. 2021; Ciumărnean et al. 2022; Hassan et al. 2022; Swathy and Saruladha 2021). In this algorithm, with a hidden layer of 100 nodes, the activation function was a hyperbolic tan: "tanh": weight optimization using the solver "sgd", weight update with a learning rate of "invscaling", and a maximum number of epochs is 10.

4.4.5. Logistic regression classifier (LR)

LR is a statistical ML approach that is applied to analyse and build the relationship between one or more independent and binary dependent variables. In this study, we have set the classification multi_class = 'multinomial' with the faster optimization problem solver = 'newton-cg', which handles multinomial loss and random_state = 1, the norm of penalty was l2, the maximum number of iterations was 100, and the inverse of regularization was 109.85.

4.4.6. Random forest classifier (RF)

The RF is based on the election voting method. In a forest, each tree is represented as a voter, and the percentage of votes from the set of trees is made for the final decision. In this study, we have considered 1000 trees, 1 random state, a maximum depth of 5, and five features selected at

random to avoid overfitting "log2" for the randomness of each tree and six leaf nodes for maximum voting and better prediction.

4.4.7. Bagging classifier (BC)

Bagging is an ensemble classifier approach that works by constructing multiple decision trees that are used as meta-estimators for regression and classification. Different decision tree outcomes were accumulated together to form a strong prediction. The common parameters used were base_estimator=None, n_estimator = 10, and random_state=None.

4.4.8. Extra tree classifier (TC)

Extre Tree is one of the most accurate and efficient ML ensemble learning-based approaches that generates a classification result by combining several de-correlated decision trees, resulting in a "forest" (Alizadehsani et al. 2019). This algorithm is best suited for both classification and regression purposes. The parameter used was n_estimator = 100.

4.4.9. Ridge classifier (RC)

RC is a well-known ridge regression-based binary classification algorithm that uses the ridge regularization technique, which penalizes the model parameters to control complexity and avoid overfitting. The parameter used was alpha = 2.0, the degree of regularization.

4.4.10. Passive aggressive classifier (PAC)

Passive-Aggressive algorithm gets its name from the fact that if the forecast is right, the model is kept, and if not right, it adjusts the weight vector by a small amount. The amount of adjustment was varied on the basis of the degree of error, and for greater accuracy, a large amount of adjustments were made. The regularization parameter used was C = 1.0, which was the size of the margin and the number of mispredictions.

4.4.11. AdaBoost classifier (ADB)

Adaptive Boost is a decision tree-based optimal ML algorithm that is applied to boost the performance of any ML model by combining the decisions of weak learners. The algorithm best handles noisy and complex data and is resistant to overfitting. The common parameters involved were learning_rate = 10, n and estimators = 50.

4.4.12. Stochastic gradient descent classifier (SGD)

SGD has the capability of retarding local minima by converging them to global minima. The algorithm uses an important parameter, i.e., the learning rate (ε), such that when ε is minimal, the classifier has to go through many iterations to converge, which takes much more time. Nevertheless, if ε is too high, it may oscillate. This parameter helps in fine-tuning the performance of the model.

4.4.13. Linear support vector classifier (LSVC)

LSVC works well with large datasets and can use a linear kernel function for classification. LSVC is identical to SVC with the kernel='linear' option. However, it is applied in terms of liblinear, which is more flexible than libsvm in terms of penalties and loss functions, along with the ability to scale to large samples. The LSVC model is faster than the SVC model is.

4.4.14. ExtraTrees classifier (ETC)

ETC forms multiple decision trees, and each of these decision trees is trained via features randomly picked features from the training dataset. Some significant factors can affect the model's outcome, i.e., the number of trees in the ensemble, the maximum depth of the tree, and the number of features taken at each split. The common parameters were n_estimator = 100 and random_state = 0.

4.4.15. Gradient boosting classifier (GB)

GB is a special type of ensemble ML algorithm that is suitable for classification and regression problems. This algorithm logic is to assemble the weak classifier results that are repeatedly trained on a training dataset to create a robust classifier for accurate prediction. The parameter used was learning_rate = 0.1, the value ranges between 0 and 1, which scales the classifier of every weak learner.

4.4.16. Gaussian Naïve Bayes classifier (GNB)

This GNB is an advanced version of the naive Bayes classifier and is a robust ML probabilistic algorithm. The algorithm efficiently performs many classification tasks, especially when the number of features is relatively small.

4.4.17. Extreme gradient boost classifier (XGBoost)

XGBoost has the advantage of dealing with missing values and categorical features, which permits the algorithm to learn the nonlinear relationship between the target and categorical value. In this study, we have used two arguments: eval_metric = 'mlogloss', which measures the error over the considered estimators, with the use_label_encoder = False.

4.4.18. Extreme gradient boost random forest (XGBRF)

XGBRF delivers a good balance between bias and variance that minimizes overfitting by increasing the accuracy of the model. In XGBRF, we use the eval_metric = 'logloss' loss function for classification to measure the cost of poor prediction accuracy in classification tasks with no label encoding use_label_encoding = 'False'.

4.4.19. Voting classifier

The VC is one of the majority-based optimal ML algorithms that learns from an ensemble of countable models and predicts an outcome (class) based on the highest likelihood of the results being selected as the target class. VC is ensemble of ETC, RF, XGB, BC, and VC with voting="soft", the most likely class, by adding up the probabilities for each class from all the models and selecting the one with the highest sum.

5. Simulation results

5.1. Performance measures

Five performance measures were used to evaluate the true potential of the ML models in classifying CVD: recall (Eq. 2), F1 score (Eq. 3), accuracy (Eq. 4), precision (Eq. 5) and Cohen's kappa (Eq. 6).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2{*}(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \tag{3}$$

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

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Kappa} = (\text{p}_\text{o} - \text{p}_\text{e})/(1 - \text{p}_\text{e}) \tag{6}$$

where true positive (TP) denotes the values that are designated as true and were true, false positive (FP) denotes the erroneous values that are recognized as true, false negative (FN) occurs when a value is true but incorrectly labelled negative, and true negative (TN) denotes the values that are negative and are correctly detected as such. Where p_o represents the correlated observed acceptance between annotators and where p_e represents the theoretical probability of agreement between annotators. The Cohen's kappa value ranges between 0 and 1, where

- 0 = no agreement between the annotators
- 1 = perfect agreement between annotators

5.2. Statistical tests

In this study, we have compared all the ML classifiers dataset-wise using the Wilcoxon Signed-Rank test and compared the ML classifiers considering all the datasets at once using the Friedman and Nemenyi hypothesis tests. Both tests are nonparametric and are conducted with a 95% confidence level.

5.3. Datasets

This study considers two benchmark CVD datasets from the UCI and Kaggle repositories, which are split into training, validation, and test set ratios of 80:10:10. These CVD datasets are open source and have been evaluated by several authors. Dataset_1 was donated to the repository on 02/04/2020. The authors of Chicco et al. (Chicco and Jurman 2020) had already expanded the applications of this dataset. Similarly, the dataset_2 was updated to the Kaggle repository in the year 2020 and was expanded by the author Alizadehsani et al. (Alizadehsani et al. 2019). The CVD datasets are presented in Tables 2, and 3 presents the heart failure dataset, Heart Statlog Cleveland Hungary dataset. Each dataset has several distinct features with varied instances. Among these datasets, some features are found to be more pertinent than others. This section explains various parameters related to patients' demographic data (age, gender, time, education), blood reports (creatinine, cholesterol, ejection fraction, sodium, cholesterol, fasting blood sugar), disease types (diabetes, anemia), major risk factors (smoking), resting BP, chest pain, maximum, exercise angina, Oldpeak, and ST slope.

Figure 4 represents the mortality ratio between the two datasets, the Heart Failure Dataset and the Heart Statlog Cleveland Hungary Dataset. The mortality of males (1: male) was greater than that of females (0: female) in both scenarios. In Figure 4(a), out of 299 cases, the mortality rate is 100

Table 2. Feature descriptions of heart failure dataset.

Attributes	n	Min	Max	Mean	Standard deviation	Attribute type	Attribute description
Age	47	40	95	60.8	11.89	Integer	–
Anemia	2	0	1	0.43	0.49	Integer	0:absent, 1:present
Creatinine_phosphokinase	208	23	7861	581.8	970.2	Integer	It is an enzyme expressed by various tissues and cell types
Diabetes	2	0	1	0.41	0.49	Integer	0:absent, 1:present
Ejection_fraction	17	14	80	38	11.83	Integer	amount of blood that pump out of the heart each time it beats
High_blood_pressure	2	0	1	0.35	0.47	Integer	0:no, 1:yes
Platelets	176	25100	850000	263358	97804	Float	Small colourless blood fragments that form clots
Serum_creatinine	40	0.5	9.4	1.3	1	float	Level of creatinine present in the blood
Serum_sodium	27	113	148	136.6	4.4	Integer	Level of sodium present in the blood
Sex	2	0	1	0.64	0.47	Integer	0:female, 1:male
Smoking	2	0	1	0.32	0.46	Integer	0:no, 1:yes
Time	148	4	285	130	77.61	Integer	–
Death_event	2	0	1		0.47	Integer	0:no risk, 1:risk

Table 3. Feature descriptions of the Heart_Statlog_Cleveland_Hungary dataset.

Attributes	n	Min	Max	Mean	Standard deviation	Attribute type	Attribute description
Age	50	28	77	53.7	9.3	Integer	-
Gender	2	01	1	0.7	0.4	Integer	0:female, 1:male
Chest_pain_type	4	0	4	3.2	0.9	Integer	1:stable angina, 2:unstable angina, 3:microvascular angina, 4:variant angina
Resting_BP	67	0	200	132.1	18	Integer	Blood pressure at body rest
Cholesterol	222	0	603	210.3	101	Integer	Blood cholesterol in body
FBS	2	0	1	0.2	0.4	Integer	Blood sugar level before having food
Resting_ECG	3	0	2	0.6	0.87	Integer	ECG at body rest
Max_Heart_rate	119	60	202	139.7	25.5	Integer	Maximum times heart beats
Exercise_Angina	2	0	1	0.3	0.4	Integer	0:absent, 1:present
OldPeak	53	−2.6	6.2	0.9	1	Float	ST depression induced by exercise relative to rest.
ST_Slope	3	0	3	1.6	0.6	Integer	1:upsloping, 2:flat, 3:downsloping
Target	2	0	1	0.5	0.4	Integer	0:no risk, 1:risk

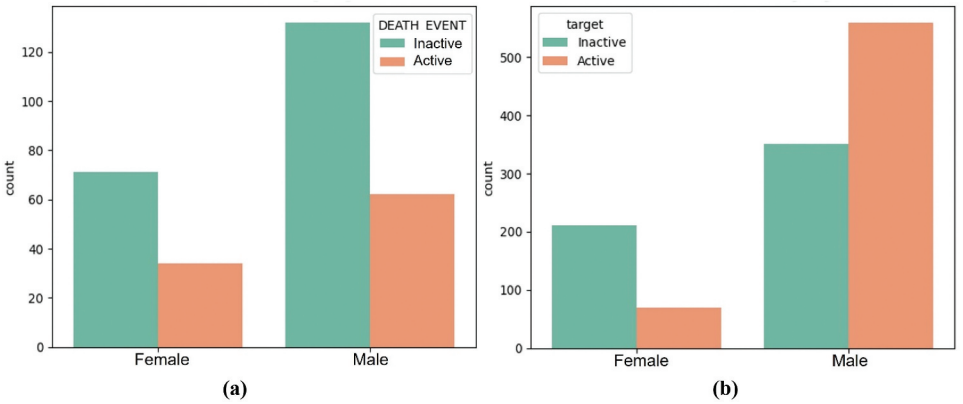

Figure 4. (a,b) Comparing mortality ratio in male and female for both the datasets.

(60 males and 40 females), which accounts for one-third of the total population of the dataset. Similarly, in Figure 4(b), out of 1191 cases, 610 (560 males and 50 females) represent 51% of the total population of the dataset. Hence, the number of male deaths in the two datasets was greater than the number of female deaths.

5.4. Results analysis considering individual datasets

Table 4 shows the mean and standard deviation (Std. Dev) of the accuracy, precision, recall, F1 score, and kappa score obtained over 50 independent simulations for each of the 19 ML models for the heart failure dataset. Among the 19 algorithms, it is observed that the Gradient Boosting (GB), Extreme Gradient Boosting (XGB), Support Gradient Boosting (GB), Multi-Layer Perceptron (MLP), and Ridge Classifier (RC) models provide the best means of Kappa, F1-Score, Recall, Precision, and Accuracy, respectively, for the heart failure dataset. Since some of the classifiers are stochastic, the Wilcoxon signed-rank test is applied. It compares the best model for a performance measure with the remaining models to obtain a statistically better (+), worse (-), and equivalent (≈) model with respect to the corresponding best model (Sibarama Panigrahi and Behera 2020). The test results are described in Table 5. Table 5 shows that the superior model for the respective performance measure statistically outperforms all other models in that particular performance measure.

Figure 5(a) represents the highest accuracy of the 19 ML models in three basic forms, i.e., normal data splitting with train_test with a ratio of 80:20, best validation splitting with a ratio of

Table 4. Mean and standard deviation of performance measures of ML models considering heart failure dataset (best performances in boldface).

	Accuracy Mean ± Std. Dexv	Precision Mean ± Std. Dev	Recall Mean ± Std. Dev	F1-Score Mean ± Std. Dev	Kappa Mean ± Std. Dev
ADB	0.8333 ± 0	0.611 ± 0	0.9167 ± 0	0.733 ± 0	0.649 ± 0
BC	0.9333±0.057	0.655 ± 0.0627	0.885 ± 0.0731	0.8 ± 0	0.676 ± 0.0701
DT	0.74 ± 0.084	0.5451 ± 0.039	0.8667 ± 0.0607	0.667 ± 0.0376	0.559 ± 0.051
ETC	0.73 ± 0.0453	0.775 ± 0.0446	0.8867± 0.0689	0.826 ± 0.047	0.7792 ± 0.0589
GB	0.8333 ± 0	0.75 ± 0	**1 ± 0**	0.857 ± 0	**0.8148 ± 0**
GNB	0.7667 ± 0	0.4545 ± 0	0.4167 ± 0	0.434 ± 0	0.3011 ± 0
KNN	0.8 ± 0	0.7143 ± 0	0. 4167 ± 0	0.526 ± 0	0.444 ± 0
LR	0.8333 ± 0	0.7692 ± 0	0.833 ± 0	0.8 ± 0	0.7475 ± 0
LSVC	0.533 ±0.204	0.7857±0	0.9167 ± 0	0.846 ± 0	0.803 ± 0
MLP	0.666 ± 0	**0.7981 ± 0.055**	0.915 ± 0.0117	0.851 ± 0.0331	0.810 ± 0.0436
PAC	0.534 ± 0.135	0.5430 ±0.225	0.765±0.2628	0.581 ± 0.173	0.458 ± 0.195
RC	**0.966 ± 0**	0.733 ± 0	0.9167 ± 0	0.814 ± 0	0.7619 ± 0
RF	0.7567 ± 0.041	0.733 ± 0	0.9167 ±0	0.814 ± 0	0.7619 ± 0
SGD	0.606 ± 0.124	0.6737 ± 0.223	0.6983 ± 0.268	0.616 ± 0.145	0.5272 ± 0.145
SVC	0.9 ± 0	0.75 ± 0	0.75 ± 0	0.5455 ± 0	0.6875 ± 0
TC	0.7587 ± 0.074	0.4371 ± 0.087	0.64 ± 0.1402	0.514 ± 0	0.3626 ± 0.1248
VC	0.7527 ± 0.031	0.708 ± 0.0228	0.9183 ± 0.0117	0.799 ± 0.015	0.7407 ± 0.0203
XGB	0.833 ± 0	0.75 ± 0	1 ± 0	**0.8571 ± 0**	0.8148 ± 0
XGBRF	0.7667 ± 0	0.66 ± 0	1 ± 0	0.8 ± 0	0.7368 ± 0

Table 5. Wilcoxon signed-rank test results on heart failure dataset denoting the better (+), worse (−) and equivalent (≈) model to that of the corresponding best model for a particular performance measure.

	Accuracy	Precision	Recall	F1-Score	Kappa
Best Model	BC	LR	SVC	SVC	GB
ADB	−	−	−	−	−
BC	−	−	−	−	−
DT	−	−	−	−	−
ETC	−	−	−	−	−
GB	−	−	NA	−	NA
GNB	−	−	−	−	−
KNN	−	−	−	−	−
LR	−	−	−	−	−
LSVC	−	−	−	−	−
MLP	−	NA	−	−	−
PAC	−	−	−	−	−
RC	NA	−	−	−	−
RF	−	−	−	−	−
SGD	−	−	−	−	−
SVC	−	−	−	−	−
TC	−	−	−	−	−
VC	−	−	−	−	−
XGB	−	−	−	NA	−
XGBRF	−	−	−	−	−

80:10:10 and validation splitting with the best hyperparameter optimization. It seems that some models have better performance, whereas some are not when hyper-optimized with the dataset train_validation_test split. The models, namely, DT and ETC, achieve a superior accuracy of 100% when hyper-optimized. Similarly, the models namely ETC, and LSVC achieve a score of 1.0 for the precision measure and PAC, and SGD achieves a score of 1.0 with respect to the recall score in dataset 1.

Table 6 presents the mean and standard deviation (Std. Dev) of the accuracy, precision, recall, F1 score, and kappa scores obtained over 50 independent simulations for each of the 19 ML models for the Heart_Statlog_Cleveland_Hungary dataset. In Table 6, the ETC, ETC, XGB, ETC, and ETC models provide the best mean Kappa, F1-Score, Accuracy, Recall, and Precision, respectively,

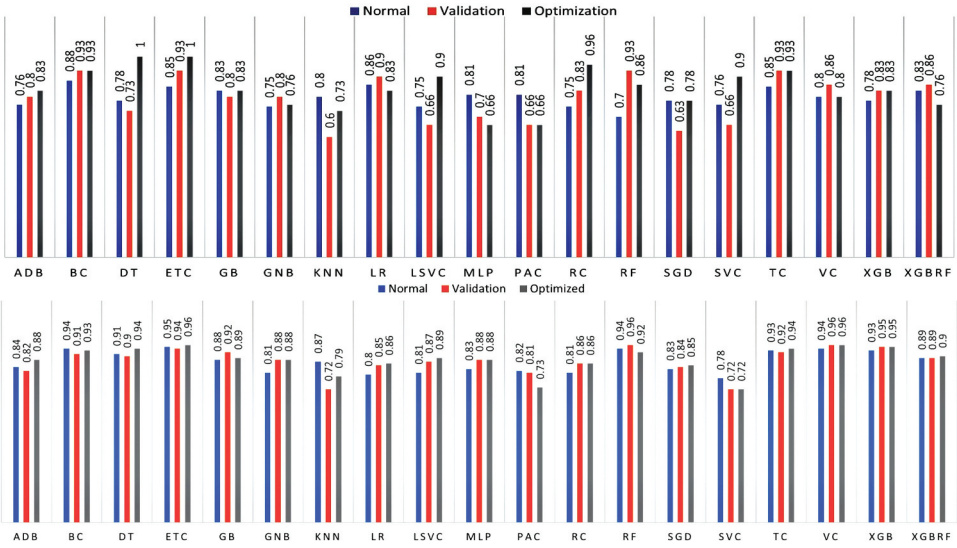

Figure 5. Performance of the 19 ML classifiers on three bases (1: normal split with no optimization, 2: split the dataset with train_validation_test with no optimization, 3: split the dataset with train_validation_test with optimized hyperparameters) for the Heart Failure dataset.

Table 6. Means and standard deviation of performance measures of ML models considering Heart_Statlog_Cleveland_Hungary dataset (best performances in boldface).

	Accuracy Mean ± Std. Dev	Precision Mean ± Std. Dev	Recall Mean ± Std. Dev	F1-Score Mean ± Std. Dev	Kappa Mean ± Std. Dev
ADB	0.882 ± 0	0.981 ± 0	0.806 ± 0	0.885 ± 0	0.7669 ± 0
BC	0.904 ± 0.0163	0.927 ± 0.0189	0.9 ± 0.0229	0.913 ± 0.0151	0.806 ± 0.0329
DT	0.88 ± 0.0048	0.902 ± 0.0045	0.86 ± 0.0079	0.882 ± 0.0051	0.764 ± 0.0096
ETC	**0.93 ± 0.0048**	**0.912 ± 0.0044**	0.955 ± 0.0078	**0.933 ± 0.0047**	**0.86 ± 0.0097**
GB	0.88 ± 0.0008	0.873 ± 0.0002	0.902 ± 0.0016	0.887 ± 0.0008	0.764 ± 0.0016
GNB	0.82 ± 0	0.827 ±0	0.827 ± 0	0.827 ± 0	0.6468 ± 0
KNN	0.84 ± 0	0.819 ± 0	0.893 ± 0	0.854 ± 0	0.688 ± 0
LR	0.81 ± 0	0.785 ± 0	0.868 ± 0	0.824 ± 0	0.62 ± 0
LSVC	0.806 ± 0	0.783 ±0	0.86 ± 0	0.82 ± 0	0.612 ± 0
MLP	0.827 ± 0.0048	0.813 ± 0.0064	0.861 ± 0.0084	0.836 ± 0.0046	0.654 ± 0.0097
PAC	0.744 ± 0.067	0.772 ± 0.093	0.77 ± 0.214	0.737 ± 0.141	0.487 ± 0.132
RC	0.806 ± 0	0.783 ± 0	0.86 ± 0	0.82 ± 0	0.6122 ± 0
RF	0.924 ± 0	0.912 ± 0	0.942 ± 0	0.927 ± 0	0.8485 ± 0
SGD	0.785 ± 0.0237	0.771 ± 0.0472	0.841 ± 0.089	0.799± 0.3	0.5703 ± 0.048
SVC	0.802 ± 0	0.809 ± 0	0.806 ± 0	0.6829 ± 0	0.604 ± 0
TC	0.873 ± 0.0184	0.868 ± 0.0221	0.887± 0.0238	0.877 ± 0.017	0.746 ± 0.036
VC	0.923 ± 0.0047	0.915 ± 0.0078	0.94 ± 0.0091	0.926 ± 0.0048	0.8463 ± 0.0094
XGB	0.907 ± 0	0.909 ± 0	**0.909 ±0**	0.909 ± 0	0.815 ± 0
XGBRF	0.84 ± 0	0.818 ± 0	0.885 ± 0	0.85 ± 0	0.679 ± 0

among all the models considered in this study. Since some of the classifiers are stochastic, the Wilcoxon signed-rank test is utilized, and the test outcomes are described in Table 7, which shows a statistically better (+), worse (-), and equivalent (≈) model with respect to the corresponding best model. In the Recall column, the RF and VC models are statistically equivalent to XGB. In all other cases, the best model for a performance measure statistically outperforms all other models in that particular performance measure.

Figure 5(b) represents the highest achieved accuracy of the 19 machine learning models and is depicted in three fundamental formats: standard data division using the train_test method with an 80:20 ratio,

Table 7. Wilcoxon signed-rank test results on the Heart_Statlog_Cleveland_Hungary dataset denoting the better (+), worse (−), and equivalent (≈) model to that of the corresponding best model for a particular performance measure.

	Accuracy	Precision	Recall	F1-Score	Kappa
Best Model	ETC	ETC	XGB	ETC	ETC
ADB	−	−	−	−	−
BC	−	−	−	−	−
DT	−	−	−	−	−
ETC	NA	NA	−	NA	NA
GB	−	−	−	−	−
GNB	−	−	−	−	−
KNN	−	−	−	−	−
LR	−	−	−	−	−
LSVC	−	−	−	−	−
MLP	−	−	−	−	−
PAC	−	−	−	−	−
RC	−	−	−	−	−
RF	−	−	≈	−	−
SGD	−	−	−	−	−
SVC	−	−	−	−	−
TC	−	−	−	−	−
VC	−	−	≈	−	−
XGB	−	−	NA	−	−
XGBRF	−	−	−	−	−

optimal validation division with an 80:10:10 ratio, and validation partitioning with the finest hyperparameter optimization. Certain models exhibit better performance compared to others when subjected to hyperparameter optimization with the dataset train_validation_test split. Specifically, the VC and ETC models demonstrate exceptional accuracy, achieving a perfect score of 96% when hyper optimized for classifying dataset2. Similarly, the model PAC achieves a precision and recall score of 1.0 for dataset 2.

5.5. Hyperparameter optimization

Table 8 shows the higher best hyper parameters for 19 ML algorithm optimizations (Saboor et al. 2022). Among these classifiers, most have improved results in response to performance measures. There are few classifiers that yield superior outcomes, namely, DT and ETC, in terms of accuracy (i.e., 100%) when optimized with the best combination of hyperparameters.

5.6. Result analysis considering all datasets

In this portion, to rank the performance of the ML models measure wise, the outcomes are analysed considering both CVD datasets at a time. The Friedman and Nemenyi hypothesis tests are carried out separately for five performance measures, namely, accuracy, precision, recall, F1 score, and kappa, to rank the 19 ML models in classifying CVD.

Figure 6(a) shows the Friedman and Nemenyi hypothesis test results by studying the accuracy performance metric is studied. ETC and XGB acquire Rank-1 values and are statistically equivalent, as the mean rank difference value is below the critical distance (19.8). The BC model is statistically equivalent to the MLP, as the mean rank difference value is below the critical distance (19.8). Similarly, SGD and DT are statistically equivalent to GNB, RC and TC, ADB are statistically equivalent to KNN, and LR, as the mean rank difference value is below the critical distance, i.e., 19.8.

Figure 6(b) shows the Friedman and Nemenyi hypothesis test results by studying the Precision performance measure. ETC achieves Rank-1 among all 19 ML classifiers to classify CVD since it has the highest mean rank (819.88), which is larger than all other models' means, with a difference of more than the critical distance (19.8). The DT, KNN, and MLP models are statistically equivalent to SVC, LSVC, and BC. Similarly, PAC and RC are statistically equivalent to ADB and SGD, as the mean rank difference value is below the critical distance (19.8).

Table 8. Best hyper parameters for 19 ML algorithms optimization.

Model	Best hyperparameters values
ADB	(n_estimators = 50, learning_rate = 10, algorithm='SAMME')
BC	(max_features = 10, max_samples = 25, n_estimators = 100)
DT	(criterion='gini', splitter='random', max_depth = 10, min_samples_leaf = 10, min_samples_split = 5, max_features = 10)
ETC	(n_estimators = 100, random_state = 0)
GB	(learning_rate = 0.1, n_estimators = 100, subsample = 1.0, criterion='friedman_mse', min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0, max_depth = 3, min_impurity_decrease = 0.0, verbose = 0, warm_start=False, validation_fraction = 0.1, tol = 0.0001, ccp_alpha = 0.0)
GNB	(priors=None, var_smoothing = 1e-09)
KNN	(n_neighbors = 5, weights='distance', algorithm='ball_tree', leaf_size = 30, p = 2, metric='minkowski')
LR	(dual=False, fit_intercept=True, intercept_scaling = 1, max_iter = 100, multi_class='auto', penalty='l2', solver='newton-cg', tol = 0.0001, verbose = 0, warm_start=False)
LSVC	(tol = 0.0001, C = 1.0, multi, class='crammer_singer', fit_intercept=True, intercept_scaling = 1, verbose = 0, max_iter = 1000)
MLP	(hidden_layer_sizes = 100, activation='tanh', solver='sgd', alpha = 0.0001, batch_size = 10, learning_rate='invscaling', learning_rate_init = 0.001, power_t = 0.5, max_iter = 200, shuffle=True, tol = 0.0001,verbose=False, warm_start=False, momentum = 0.9,nesterovs_momentum=True, early_stopping=False, validation_fraction = 0.1, beta_1 = 0.9, beta_2 = 0.999, epsilon = 1e-08, n_iter_no_change = 10, max_fun = 20000)
PAC	(C = 1.0, fit_intercept=True, max_iter = 1000, tol = 0.001, early_stopping=False, validation_fraction = 0.1, n_iter_no_change = 5, shuffle=True, verbose = 0, loss='hinge', warm_start=False, average=False)
RC	(alpha = 2.0, fit_intercept=True, copy_X=True, tol = 0.0002, solver='sparse_cg', positive=False)
RF	(max_depth = 3, max_features='log2', max_leaf_nodes = 6, n_estimators = 100)
SGD	(loss='squared_hinge', penalty='l1', alpha = 0.0001, l1_ratio = 0.15, fit_intercept=True, max_iter = 1000, tol = 0.001, shuffle=True, verbose = 0, epsilon = 0.1, learning_rate='optimal', eta0 = 0.0, power_t = 0.6, early_stopping=False, validation_fraction = 0.1, n_iter_no_change = 5, warm_start=False, average=False)
SVC	(C = 1.0, kernel='rbf', degree = 3, gamma='scale', coef0 = 0.0, shrinking=True, probability=False, tol = 0.001, cache_size = 200, verbose=False, maxiter = −1, decision_function_shape='ovr', break_ties=False)
TC	(min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0, max_features = 1.0, min_impurity_decrease = 0.0, ccp_alpha = 0.0)
VC	Ensemble of ETC, RF, XGB, BC, VC (estimators=[('etc', clf1), ('rf', clf2), ('xgb', clf3), ('bc', clf4)], voting='soft')
XGB	(learning_rate = 0.2, n_estimators = 1000, max_depth = 15, min_child_weight = 1, alpha = 5, subsample = 0.8, colsample_bytree = 0.8, objective='binary:logistic', nthread = 4, scale_pos_weight = 1, estimators = 123)
XGBRF	(n_estimators = 100, max_depth = 100)

Figure 6(c) shows the Friedman and Nemenyi hypothesis test results by studying the Recall performance measure. In Figure 6(c), XGB acquires Rank-1 values among all the considered 19 ML classifiers to classify CVD since it has the highest mean rank (844.75), which is greater than all other models' means, with a difference more than the critical distance (19.8). The RF model is statistically equivalent to VC. Similarly, DT and RC are statistically equivalent to LSVC and MLP. Both KNN and TC are statistically equivalent to LR and SGD, as the mean rank difference value is below the critical distance (19.8).

Figure 6(d) shows the Friedman and Nemenyi hypothesis test results considering the F1 Score performance measure. In Figure 6(d), XGB achieves Rank-1 among all 19 ML classifiers to classify CVD since it has the highest mean rank (801.25), which is larger than all other models' means, with a difference of more than the critical distance (19.8). The XGBRF and DT models are statistically equivalent to the MLP and BC models, as the mean rank difference is below the critical distance (19.8). Similarly, LR and TC are statistically equivalent to KNN. Furthermore, ADB and PAC, GNB and SGD are statistically equivalent in classifying CVD.

Figure 6(e) shows the Friedman and Nemenyi hypothesis test results by studying the Kappa performance measure. In Figure 6(e), XGB achieves Rank-1 among all 19 ML classifiers to classify CVD considering Cohen's Kappa measure since it has the highest mean rank (825.5) greater than all other models' means, with a difference greater than the critical distance (19.8). The VC, GBRF and BC are statistically equivalent to the GB, and the MLP, as the mean rank difference value, is below the critical distance (19.8). Similarly, TC, LR, and DT are statistically equivalent to ADB and RC. Moreover, SGD is statistically equivalent to the GNB model in classifying CVD according to Cohen's Kappa measure.

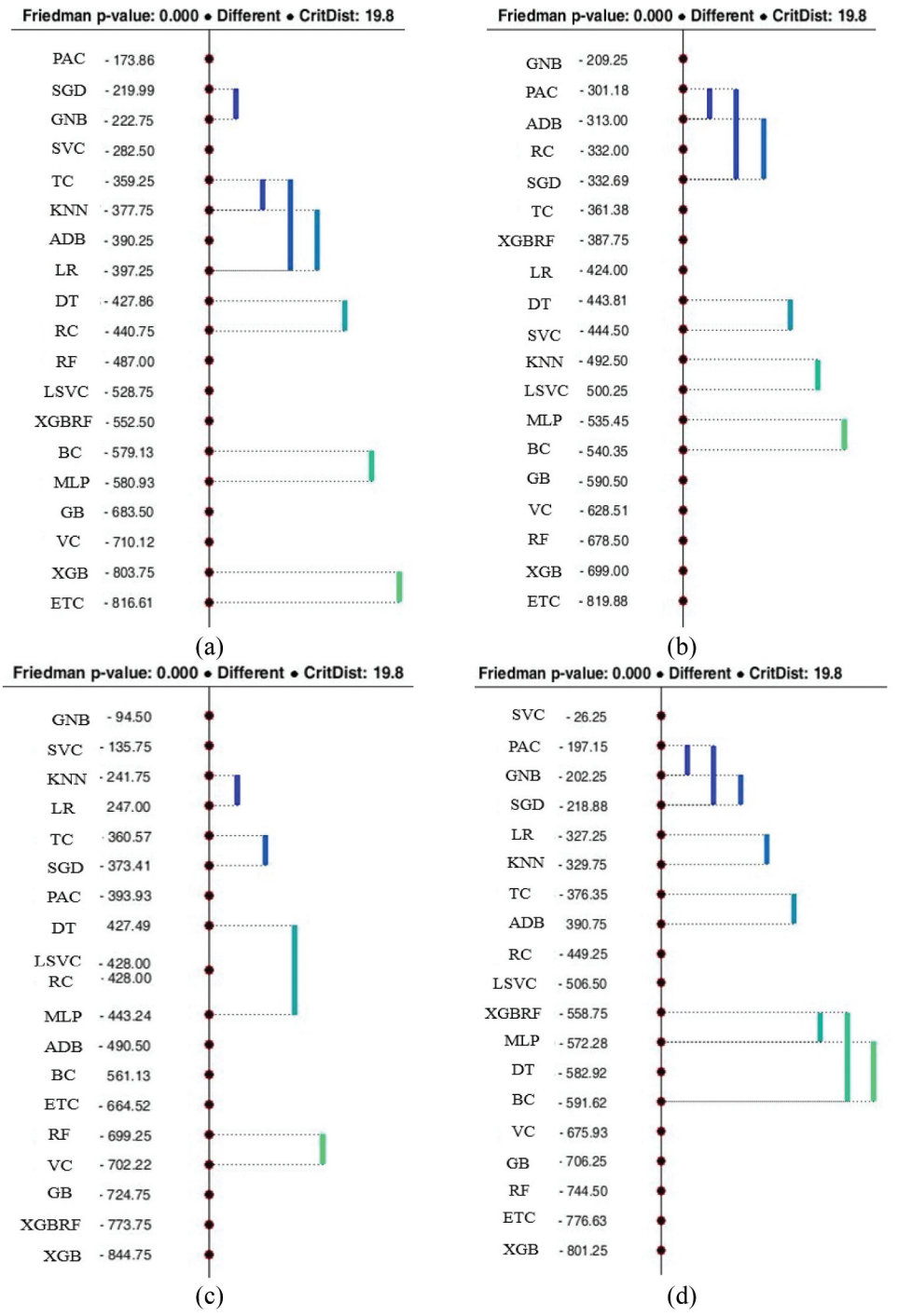

Figure 6. Mean ranks of ML models using Friedman and Nemenyi hypothesis test on accuracy, precision, recall, F1, and Cohen's kappa measure studying both datasets.

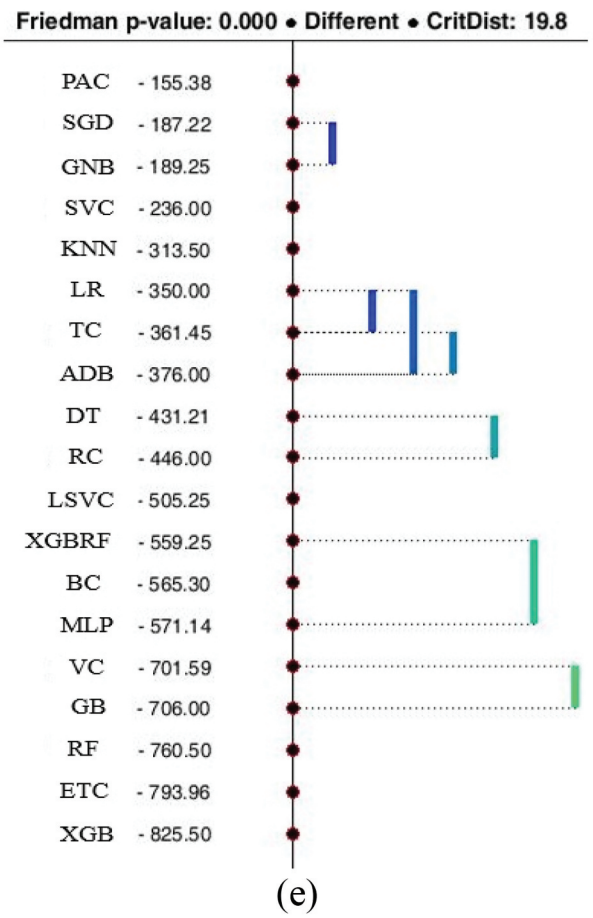

Figure 6. (Continued).

6. Conclusions and future work

Cardiovascular disease is a significant public health issue, especially among elderly people. The prevalence of CVD can be decreased by early detection of the disease. Hence, this can aid in excluding the continuation of the disease. A graphical analysis of both datasets found that mortality in men is greater than that in women due to various risk factors. For this reason, during the past decade, a massive number of ML models have been applied for the prediction of CVD. Motivated by this, in this paper, 19 promising ML models are statistically compared using two benchmark datasets to rank the models using five performance measures for CVD diagnosis. A graphical analysis of the results revealed that male patients had a greater mortality rate in comparison to females. It was presuming that, this can be a reason for the rise in global deaths due to several major risk factors that were more common in males. A comparison in Figure 5, the models Decision Tree (DT) and Extra Trees (ETC) achieved an accuracy performance of 100% for Heart Failure dataset. The Friedman and Nemenyi hypothesis tests are applied to study the two datasets together. In this test, XGB achieves the first rank in terms of the recall, F1 score, and Kappa performance measures, whereas ETC achieves the first rank in terms of both Accuracy and Precision measure, respectively. Hence, irrespective of the CVD dataset, if one wants to apply an ML model to diagnose CVD, it is suggested by using XGB and ETC. Out of the 19 ML models used, there are few, namely, RC, SVC, and LR, that do not perform as expected in terms of the f1 score and kappa score. Additionally, this study can be improved by considering a greater number of dataset features by

employing feature selection (filter, wrapper, or embedded methods), feature engineering (early, late, and intermediate fusion), feature transformation (Principal Component Analysis, t-distributed stochastic neighborhood embedding, etc.) techniques and ensembling the resulting models to boost the performance of the ML classifiers. ML algorithms can lead to biased data due to a lack of high-quality, relevant data. The models can perform well during training but may not generate excellent outcomes for unseen, new, unknown problems. In addition, one can hybridize the computer vision techniques of CT-Scan and Chest X-ray with the methods employed in this study to boost the performance of CVD diagnosis.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Financial support from ICMR, Govt. of India, Sanction No. ISRM/12(129)/2020, Project Id: 2020-5341 is highly acknowledged. ICMR – Indian Council of Medical Research [Grant No. ISRM/12(129)/2020].

ORCID

Tapan Kumar Behera http://orcid.org/0009-0001-9601-3898
Siddhartha Sathia http://orcid.org/0000-0001-7929-6200
Pradeep Kumar Naik http://orcid.org/0000-0001-7044-2427

Data sources

Dataset_1: https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records.
Dataset_2: https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final/data.

References

Ali, M. M., B. K. Paul, K. Ahmed, F. M. Bui, J. M. Quinn, and M. A. Moni. 2021. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine* 136:104672. doi: 10.1016/j.compbiomed.2021.104672.

Alizadehsani, R., A. Khosravi, M. Roshanzamir, M. Abdar, N. Sarrafzadegan, D. Shafie, and U. R. Acharya. 2021. Coronary artery disease detection using artificial intelligence techniques: A survey of trends, geographical differences and diagnostic features 1991–2020. *Computers in Biology and Medicine* 128:104095. doi: 10.1016/j.compbiomed.2020.104095.

Alizadehsani, R., M. Roshanzamir, M. Abdar, A. Beykikhoshk, A. Khosravi, M. Panahiazar, and A. Koohestani, F. Khozeimeh, S. Nahavandi, N. Sarrafzadegan. 2019. A database for using machine learning and data mining techniques for coronary artery disease diagnosis. *Scientific Data* 6 (1):227. doi: 10.1038/s41597-019-0206-3.

Ammar, A., O. Bouattane, and M. Youssfi. 2021. Automatic cardiac cine MRI segmentation and heart disease classification. *Computerized Medical Imaging and Graphics* 88:101864. doi: 10.1016/j.compmedimag.2021.101864.

Arena, S., E. Florian, I. Zennaro, P. F. Orrù, and F. Sgarbossa. 2022. A novel decision support system for managing predictive maintenance strategies based on machine learning approaches. *Safety Science* 146:105529. doi: 10.1016/j.ssci.2021.105529.

Bertsimas, D., L. Mingardi, and B. Stellato. 2021. Machine learning for real-time heart disease prediction. *IEEE Journal of Biomedical and Health Informatics*. [Online Conference] 25(9): 3627–37. doi: 10.1109/JBHI.2021.3066347.

Chen, R., A. Lu, J. Wang, X. Ma, L. Zhao, W. Wu, and H. Liu. 2019. Using machine learning to predict one-year cardiovascular events in patients with severe dilated cardiomyopathy. *European Journal of Radiology* 117:178–183. doi: 10.1016/j.ejrad.2019.06.004.

Chicco, D., and G. Jurman. 2020. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics & Decision Making* 20 (1):1–16. doi: 10.1186/s12911-020-1023-5.

Ciumărnean, L., M. V. Milaciu, V. Negrean, O. H. Orăşan, S. C. Vesa, O. Sălăgean, and S. I. Vlaicu. 2022. Cardiovascular risk factors and physical activity for the prevention of cardiovascular diseases in the elderly. *International Journal of Environmental Research and Public Health* 19 (1):207. doi: 10.3390/ijerph19010207.

Hagan, R., C. J. Gillan, and F. Mallett. 2021. Comparison of machine learning methods for the classification of cardiovascular disease. *Informatics in Medicine Unlocked* 24:100606. doi: 10.1016/j.imu.2021.100606.

Hambali, M., Y. Saheed, T. Oladele, and M. Gbolagade. 2019. ADABOOST ensemble algorithms for breast cancer classification. *Journal of Advances in Computer Research* 10 (2):31–52.

Hassan, M. R., S. Huda, M. M. Hassan, J. Abawajy, A. Alsanad, and G. Fortino. 2022. Early detection of cardiovascular autonomic neuropathy: A multiclass classification model based on feature selection and deep learning feature fusion. *Information Fusion* 77:70–80. doi: 10.1016/j.inffus.2021.07.010.

Jamthikar, A., D. Gupta, N. N. Khanna, L. Saba, J. R. Laird, and J. S. Suri. 2020. Cardiovascular/Stroke risk prevention: A new machine learning framework integrating carotid ultrasound image-based phenotypes and its harmonics with conventional risk factors. *Indian Heart Journal* 72 (4):258–264. doi: 10.1016/j.ihj.2020.06.004.

Janiesch, C., P. Zschech, and K. Heinrich. 2021. Machine learning and deep learning. *Electronic Markets* 31 (3):685–695. doi: 10.1007/s12525-021-00475-2.

Kumari, S., P. Muthulakshmi, and D. Agarwal. 2022. Deployment of machine learning based internet of things networks for tele-medical and remote healthcare. *Evolutionary Computing And Mobile Sustainable Networks: Proceedings Of ICECMSN 2021*, 305–317, Singapore: Springer Singapore.

Lip, G. Y., A. Genaidy, G. Tran, P. Marroquin, and C. Estes. 2021. Incident atrial fibrillation and its risk prediction in patients developing COVID-19: A machine learning based algorithm approach. *European Journal of Internal Medicine* 91:53–58. doi: 10.1016/j.ejim.2021.04.023.

Madabhushi, A., and G. Lee. 2016. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis* 33:170–175. doi: 10.1016/j.media.2016.06.037.

Maini, E., B. Venkateswarlu, B. Maini, and D. Marwaha. 2021. Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India. *Medical Journal Armed Forces India* 77 (3):302–311. doi: 10.1016/j.mjafi.2020.10.013.

Mohan, S., C. Thirumalai, and G. Srivastava. 2019. Effective heart disease prediction using hybrid machine learning techniques. *Institute of Electrical and Electronics Engineers Access* 7:81542–81554. doi: 10.1109/ACCESS.2019.2923707.

Nagarhalli, T. P., V. Vaze, and N. K. Rana. 2021. Impact of machine learning in natural language processing: A review. *2021 third international conference on intelligent communication technologies and virtual mobile networks (ICICV)*, Tirunelveli, India, 1529–1534, IEEE, February.

Nathala, P., V. Salunkhe, H. Samanapally, Q. Xu, S. Furmanek, O. F. Fahmy, J. Huang, A. Glynn, T. McGuffin, and D. C. Goldsmith. 2022. Electrocardiographic features and outcome correlations in 124 hospitalized COVID-19 patients with cardiovascular events. *Journal of Cardiothoracic and Vascular Anesthesia* 36 (8):2927–2934. doi: 10.1053/j.jvca.2022.01.011.

Oh, T. R., S. H. Song, H. S. Choi, S. H. Suh, C. S. Kim, J. Y. Jung, and S. W. Kim. 2021. Predictive Model for high coronary artery calcium score in young patients with non-dialysis chronic kidney disease. *Journal of Personalized Medicine* 11 (12):1372. doi: 10.3390/jpm11121372.

Pavithra, V., and V. Jayalakshmi. 2021. Hybrid feature selection technique for prediction of cardiovascular diseases. *Materials Today: Proceedings*. [Online Conference], 2021.

Pham, V., D. Laghlam, O. Varenne, F. Dumas, A. Cariou, and F. Picard. 2021. Performance of OHCA, NULL-PLEASE and CAHP scores to predict survival in out-of-hospital cardiac arrest due to acute coronary syndrome. *Resuscitation* 166:31–37. doi: 10.1016/j.resuscitation.2021.07.011.

Pires, I. M., G. Marques, N. M. Garcia, and V. Ponciano. 2020. Machine learning for the evaluation of the presence of heart disease. *Procedia Computer Science* 177:432–437. doi: 10.1016/j.procs.2020.10.058.

Ramesh, T. R., U. K. Lilhore, M. Poongodi, S. Simaiya, A. Kaur, and M. Hamdi. 2022. Predictive analysis of heart diseases with machine learning approaches. *Malaysian Journal of Computer Science* 132–148. doi: 10.22452/mjcs.sp2022no1.10.

Razeghi, O., J. A. Solís-Lemus, A. W. Lee, R. Karim, C. Corrado, C. H. Roney, and S. A. Niederer. 2020. CemrgApp: An interactive medical imaging application with image processing, computer vision, and machine learning toolkits for cardiovascular research. *SoftwareX* 12:100570. doi: 10.1016/j.softx.2020.100570.

Reddy, M. K., P. Helkkula, Y. M. Keerthana, K. Kaitue, M. Minkkinen, H. Tolppanen, T. Nieminen, and P. Alku. 2021. The automatic detection of heart failure using speech signals. *Computer Speech & Language* 69:101205. doi: 10.1016/j.csl.2021.101205.

Saboor, A., M. Usman, S. Ali, A. Samad, M. F. Abrar, N. Ullah, and H. T. Rauf. 2022. A method for improving prediction of human heart disease using machine learning algorithms. *Mobile Information Systems* 2022:1–9. doi: 10.1155/2022/1410169.

Sharma, P., K. Choudhary, K. Gupta, R. Chawla, D. Gupta, and A. Sharma. 2020. Artificial plant optimization algorithm to detect heart rate & presence of heart disease using machine learning. *Artificial Intelligence in Medicine* 102:101752. doi: 10.1016/j.artmed.2019.101752.

Sibarama Panigrahi, H. S. B., and H. S. Behera. 2020. A study on leading machine learning techniques for high order fuzzy time series forecasting. *Engineering Applications of Artificial Intelligence* 87:103245. doi: 10.1016/j.engappai.2019.103245.

Skandha, S. S., S. K. Gupta, L. Saba, V. K. Koppula, A. M. Johri, N. N. Khanna, and J. S. Mavrogeni, J. R. Laird, G. Pareek, M. Miner. 2020. 3-D optimized classification and characterization artificial intelligence paradigm for cardiovascular/stroke risk stratification using carotid ultrasound-based delineated plaque: Atheromatic™ 2.0. *Computers in Biology and Medicine* 125:103958. doi: 10.1016/j.compbiomed.2020.103958.

Swathy, M., and K. Saruladha. 2021. A comparative study of classification and prediction of cardio-vascular diseases (CVD) using machine learning and deep learning techniques. *ICT Express* 8 (1):109–116. doi: 10.1016/j.icte.2021.08.021.

Tuli, S., N. Basumatary, S. S. Gill, M. Kahani, R. C. Arya, G. S. Wander, and R. Buyya. 2020. HealthFog: An ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments. *Future Generation Computer Systems* 104:187–200. doi: 10.1016/j.future.2019.10.043.