Research Article
Journal of Intelligent Systems: Theory and Applications 7(2) (2024) 81-94
DOI: 10.38016/jista.1439504

Conventional Machine Learning and Ensemble Learning Techniques in Cardiovascular Disease Prediction and Analysis

Buse Yaren Kazangirler[1*], Emrah Özkaynak[2]

[1] Department of Computer Engineering, Karabük University, Karabük, Türkiye

[2] Department of Software Engineering, Karabük University, Karabük, Türkiye

tekinbuseyaren@gmail.com, eozkaynak@karabuk.edu.tr

Abstract
Cardiovascular diseases, which significantly affect the heart and blood vessels, are one of the leading causes of death worldwide. Early diagnosis and treatment of these diseases, which cause approximately 19.1 million deaths, are essential. Many problems, such as coronary artery disease, blood vessel disease, irregular heartbeat, heart muscle disease, heart valve problems, and congenital heart defects, are included in this disease definition. Today, researchers in the field of cardiovascular disease are using approaches based on diagnosis-oriented machine learning. In this study, feature extraction is performed for the detection of cardiovascular disease, and classification processes are performed with a Support Vector Machine, Naive Bayes, Decision Tree, K-Nearest Neighbor, Bagging Classifier, Random Forest, Gradient Boosting, Logistic Regression, AdaBoost, Linear Discriminant Analysis and Artificial Neural Networks methods. A total of 918 observations from Cleveland, Hungarian Institute of Cardiology, University Hospitals of Switzerland, and Zurich, VA Medical Center were included in the study. Principal Component Analysis, a dimensionality reduction method, was used to reduce the number of features in the dataset. In the experimental findings, feature increase with artificial variables was also performed and used in the classifiers in addition to feature reduction. Support Vector Machines, Decision Trees, Grid Search Cross Validation, and existing various Bagging and Boosting techniques have been used to improve algorithm performance in disease classification. Gaussian Naïve Bayes was the highest-performing algorithm among the compared methods, with 91.0% accuracy on a weighted average basis as a result of a 3.0% improvement.

Keywords: Ensemble learning, classification, conventional techniques, cardiovascular disease, hyperparameter optimization.

Kardiyovasküler Hastalık Tahmini ve Analizinde Geleneksel Makine Öğrenmesi ve Topluluk Öğrenme Teknikleri

Öz
Kalp ve kan damarlarını önemli ölçüde etkileyen kardiyovasküler hastalıklar, dünya çapında önde gelen ölüm nedenlerinden biridir. Yaklaşık 19,1 milyon kişinin ölümüne neden olan bu hastalıkların erken teşhis ve tedavisi büyük önem taşıyor. Koroner arter hastalığı, kan damarı hastalığı, düzensiz kalp atışı, kalp kası hastalığı, kalp kapağı sorunları ve doğumsal kalp kusurları gibi birçok sorun bu hastalık tanımına girmektedir. Günümüzde kardiyovasküler hastalık alanındaki araştırmacılar tanı odaklı makine öğrenmesine dayalı yaklaşımlar kullanmaktadır. Bu çalışmada kardiyovasküler hastalık tespiti için özellik çıkarma işlemi gerçekleştirilmiş ve Destek Vektör Makinesi, Naive Bayes, Karar Ağacı, K-En Yakın Komşu, Torbalı Sınıflandırıcı, Rastgele Orman, Gradyan Artırım, Lojistik Regresyon, AdaBoost, Doğrusal Diskriminant Analizi ve Yapay Sinir Ağları yöntemleri ile sınıflandırma işlemleri yapılmıştır. Cleveland, Macaristan Kardiyoloji Enstitüsü, İsviçre Üniversite Hastaneleri ve Zürih VA Tıp Merkezi'nden toplam 918 gözlem çalışmaya dahil edilmiştir. Veri kümesindeki özellik sayısını azaltmak için bir boyut azaltma yöntemi olan Temel Bileşen Analizi kullanılmıştır. Deneysel bulgularda, özellik azaltmanın yanı sıra yapay değişkenlerle özellik artırımı da gerçekleştirilmiş ve sınıflandırıcılarda kullanılmıştır. Hastalık sınıflandırmasında algoritma performansını artırmak için Destek Vektör Makineleri, Karar Ağaçları, Izgara Arama Çapraz Doğrulama, var olan çeşitli Torbalama ve Artırma teknikleri kullanılmıştır. Gauss Naïve Bayes, %3,0'lık bir iyileştirme sonucunda ağırlıklı ortalama bazında %91,0 doğrulukla karşılaştırılan yöntemler arasında en yüksek performans gösteren algoritma olmuştur.

Anahtar Kelimeler: Topluluk öğrenme, sınıflandırma, geleneksel yöntemler, kardiyovasküler hastalık, hiperparametre optimizasyonu.

[*] Corresponding Author.
E-mail: tekinbuseyaren@gmail.com

Recieved : 19 Feb 2024
Revision : 15 Jun 2024
Accepted : 30 Jul 2024

1. Introduction

In recent years, Machine Learning (ML) studies in many sectors have continued sustainably without slowing down. The studies with sub-branches of Artificial Intelligence (AI), such as ML, pattern recognition, data science, and Deep Learning (DL), are vital in medicine. During the period when ML systematics were not used in medicine and health sciences, physicians and healthcare professionals were developing a manual approach while preparing diagnosis and treatment planning for patients. Therefore, with ML gaining a critical place today, it is concluded that it helps first-level physicians in health sciences to identify better patients who require additional attention and provide personalized tasks for each individual (Malik et al., 2019; Veranyurt et al., 2020). In various kinds of research, ML reveals an automated system to perform the desired task by extracting data-dependent statistical patterns (Chollet, 2021). Thus, computerized solutions become essential to treatment monitoring and planning, helping specialists reduce the adverse effects of time loss, stress, and fatigue in daily practice (Tekin et al., 2022).

Cardiovascular systems in the body of individuals consist of heart and blood vessels. Many various problems can occur in the cardiovascular system. Endocarditis, rheumatic heart disease, and abnormalities in the conduction system are shown as a few of the types of cardiovascular disease. Cardiovascular diseases are the leading cause of mortality in individuals worldwide (Lopez et al., 2022; Vatansever et al., 2021). When the causes of cardiovascular diseases in individuals are analyzed, modifiable and non-modifiable, i.e., congenital risk factors, stand out. These risk factors include adverse factors such as physical inactivity, long work hours, and family history. Regarding risk factors, non-modifiable factors such as age, gender, hypertension, and diabetes have different effects (Gregg and Hedayati, 2018). Family history, early atherosclerotic disease, or a first-degree relative after 55 years of age in men and after 65 years of age in women is recognized as a risk factor. In addition, in terms of gender, another non-modifiable factor, male individuals are more likely to have the disease than female individuals (Lopez et al., 2022). However, cardiovascular diseases, which are caused by many different causes, can also lead to other diseases. For this reason, disease monitoring is vital for diagnosing and treating high-risk patients in the early stages of the disease (Akman and Civek, 2022).

Many academic studies on cardiovascular diseases have been put forward when similar studies are examined in recent years. As a result of the research, while there are academic studies on the disease's risk factors, analysis, and examination determinations, ML needs to be adequately addressed. In 2016, Bektaş et al. (Bektaş and Babur, 2016) conducted a similar study in the health field and analyzed the performance of ML algorithms through feature selection methods on microarray datasets and prominent genes in breast cancer.

In 2018, Cihan (Cihan, 2018) performed a classification model with Random Forest (RF), 86.13% accuracy rate was obtained on the Cleveland dataset and an 86.13% accuracy rate was obtained on the dataset consisting of 596 patient records obtained by combining the Hungarian and Cleveland datasets. Badem (Badem, 2019) brought a different dimension to AI studies in health in 2019 by detecting Parkinson's disease using ML algorithms in audio signals. In addition to the algorithms used in the study, additional analysis was performed with Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) dimensionality reduction techniques. Veranyurt et al. (Veranyurt et al., 2020) used a dataset of 390 patients with 15 attributes to classify different types of diseases. As a result of the study, he compared the classification success of RF and K-Nearest. Neighbor (KNN) algorithms and achieved the highest success result.

In 2020, Taşçı and Şamlı (Taşçı and Şamlı, 2020) performed disease classification with WEKA on a cardiovascular disease dataset. Considering the studies in the literature, the use of 9 different algorithms and 13 attributes in Taşcı and Şamlı's study represents a significant contribution. In addition, the high number of features and relatively low number of cases can sometimes be considered limitations in studies. Although the accuracy rate with ZeroR, a data mining algorithm in their study, was relatively low at 49.18%, the other algorithms mentioned were able to achieve much higher scores with an average performance of 70-87%.

In 2021, another study was carried out to predict cardiovascular disease with Genetic Algorithm (GA) and other different algorithms. In this study, Vatansever et al. (Vatansever et al., 2021) put forward a research paper to analyze the risk factors that cause the disease. The open-source cardiovascular dataset was selected for the dataset used, and feature selection was performed on 14 features in this dataset. In the experimental results, the difference in performance before and after selection is noticeable. A high success rate was obtained with GA to contribute to the literature.

In 2022, Çil and Güneş (Çil and Güneş, 2022) performed a classification of heart diseases using Support Vector Machine (SVM), RF, Artificial Neural Network (ANN), Naive Bayes, and KNN algorithms. Dimensional reduction techniques and feature extraction were performed in the study. The backward elimination method removed insignificant features from the dataset and classified them. During learning, Cihan (Cihan, 2018) used all 11 attributes in the disease dataset. This may lead to unnecessary learning and need improvement in achieving the targeted performance. In addition, it was also concluded that no dimensionality reduction technique was used. Çil and Güneş ü,(Çil and Güneş, 2022) when the classification results of the algorithms they used in their study were analyzed, it was seen that while the precision success rate was very high, other metrics that should be considered in terms of performance were shallow. Accuracy value is sometimes not sufficient for a model to be considered successful.

The dominant aspects of the proposed model are clearly visible when compared with the models used in other studies. For example, while the highest accuracy rate in the Bektaş and Babur (2016) study was 90.7%, the proposed model surpasses this study with an accuracy rate of 95.00%. Additionally, Veranyurt et al. (2020) study, a maximum accuracy rate of 92.3% was achieved with RF, KNN and AdaBoost models, while the 95.00% accuracy rate achieved by the proposed model with various algorithms is beyond this study. Vatansever et al. (Vatansever et al., 2021) study, while an accuracy rate of 93.44% was achieved with various models, even the lowest accuracy rate of the proposed model was 81.00% and showed higher performance in general.

The research projects carried out between 2016 and 2024, the data sets, the machine learning models, and the experimental results are shown in Table 1. The identification of cardiovascular diseases using diverse datasets and ML algorithms has been the subject of numerous studies. These results show that the proposed model works with a wide range of algorithms, using a mixed data set consisting of a combination of various data sets, allowing to obtain higher accuracy scores in the detection of cardiovascular diseases. This reveals that the overall performance and reliability of the model are superior compared to other studies.

The main contribution of this work is to supplement the many algorithms used in the literature for the classification of cardiovascular disease with different conventional ML, ensemble methods, and ANN. PCA achieves dimensionality reduction using the correlation relation for each attribute used in disease detection (Abdi and Williams, 2010). In addition to the performance results obtained in the test runs after the training of the models, an optimization technique, Grid Search Cross-validation (CV) (Liashchynskyi and Liashchynskyi, 2019), was used to determine the best parameters for improvement. Another contribution of the study is the use of Boosting methods, alternative powerful ensemble learning techniques, in addition to the classification algorithms, and specially built ANN models classifier. Unlike other studies, optimized performance results have been achieved with more than one preprocessing technique, which will contribute to the literature.

2. Material and Methods

2.1. Preparation of the Dataset

The data considered in this study combines different datasets that exist independently but have yet to be connected before. The difference from the datasets in the literature is that four other dataset producers use the same variables to replicate the data and store them in a publicly available data store.

Table 1. Detailed review of studies on the detection and classification of cardiovascular disease.

Year	Author	Dataset	Model	Results (Accuracy)
2016	Bektaş and Babur (Bektaş and Babur, 2016)	Breast cancer, Kent Ridge 2 dataset	K-Star, Perceptron ANN, LibSVM, RF,	80.4%, 81.4%, 84.5%, 90.7%
2018	Cihan (Cihan, 2018)	Cleveland, Hungary, Switzerland,VA Long Beach dataset	RF	86.1%
2019	Badem (Badem, 2019)	Parkinson's disease classification dataset	DT, NB, SVM, RF, KNN	79.2%, 79.6%, 86.9%, 87.6%, 91.8%,
2020	Veranyurt et al. (Veranyurt et al., 2020)	Vanderbilt University Dept. of Biostatistics Diabetes dataset	AdaBoost, RF, KNN	90.5%, 92.3%, 92.3%,
2020	Taşcı et al. (Taşçı and Şamlı, 2020)	Cardiovascular disease dataset	ZeroR, OneR, DT, RF, LR, SVM, NB, KNN, Perceptron	49.1%, 73.7%, 78.6%, 83.6%, 85.2%, 86.8%, 86.8%, 88.5%
2021	Vatansever et al. (Vatansever et al., 2021)	USA Cleveland heart dataset	KNN, DT, RF, NB, SVM, GA, LR,	81.9%, 81.9%, 83.6%, 83.6%, 85.2%, 93.4%, 90.1%
2022	Çil and Güneş (Çil and Güneş, 2022)	USA CDC heart dataset	KNN, DT, ANN, RF, SVM, NB, LR	86.2%, 87.2%, 87.2%, 89.2, 90.5%, 90.5%, 90.7%
2024	**Our proposed model**	**Mixed heart disease dataset (combination of four dataset)**	**GB, XGBoost, DT, LR, LDA, KNN, RF, SVM, AdaBoost, GNBC, ANN**	**81.0%, 82.0%, 83.0%, 84.0%, 85.0%, 86.5%, 87.0%, 88.0%, 88.0%, 90.0%, 91.0%, 95.0%**

The original dataset includes 303 observations from the Cleveland Clinic Foundation, 293 observations from the Hungarian Institute of Cardiology, 123 observations from the Swiss University Hospitals, and 199 from the Long Beach VA Medical Centre (Zein Elabedin Mohammed et al., 2020). As a result of analyzing the information provided by individuals with cardiovascular diseases, 11 attributes created in the dataset are given in Table 2. When the dataset is analyzed, modifiable and innate attributes are housed together. The attribute “Cardiovascular Disease” as the target class is a numeric variable that produces the result 0 or 1.

2.2. Exploratory Data Analysis

The main modifiable risk factors affecting coronary cardiovascular diseases are overweight, diabetes, tobacco use, blood pressure, and cholesterol (Çil and Güneş, 2022). Therefore, the 6th attribute in the table, “FastingBS” is directly related to diabetes. Thus, as control problems increase daily in diabetic patients, blood pressure and total cholesterol levels also increase (Kara and Çınar, 2011). Another attribute, “Cholesterol” is a blood lubricant that forms a circulation found in all body cells. It was observed that FastingBS and cholesterol-derived risk factors indirectly matched with criteria such as gender, low physical activity, and family history (Çil and Güneş, 2022).

Table 2. Descriptions of the attribute’s cardiovascular disease dataset.

Feature	Feature Type	Details of attributes
Age	Numerical	[28, 32, 42, ..., 77]
Sex	Nominal	[M: Male, F: Female]
ChestPainType	Nominal	[TA, ATA, NAP, ASY]
RestingBP	Numerical	[0, 80, 100, ..., 200]
Cholesterol	Numerical	[0, 120, 180, ..., 603]
FastingBS	Numerical	[0: False, 1: True]
RestingECG	Nominal	[Normal, ST-T, LVH]
MaxHR	Numerical	[60, 74, 88, ..., 202]
ExerciseAngina	Nominal	[Y: Yes, N: No]
Oldpeak	Numerical	[-2.6, 0.04, ..., 6.2]
ST-Slope	Nominal	[Y: Yes, N: No]
HeartDisease	Numerical	[0: Disease, 1: Normal]

As seen in Table 2, the variables are nominal, i.e., categorical, and numerical, i.e., numerical. For example, for FastingBS, if the value is more excellent than 120 mg, it represents 1, i.e., true, and if the value is less than 120 mg, it means 0, i.e., false. The risk factor “Sex’” nominally represents male for M (Male) and female for F (Female). For another attribute, “ChestPainType”, TA represents typical angina, ATA represents atypical angina, NAP represents non-anginal pain, and ASY represents asymptomatic angina. Angina is a feeling of chest pain caused by spasms and pain in coronary cardiovascular disease. It is concluded that the existing attributes for angina measurements for “ExerciseAngina” and “ChestPainType” should be given to the algorithms for learning purposes. For “RestingECG”, electrocardiogram measuring wave abnormality (T-wave inversions and ST elevation or depression of 0.05 mV), LVH indicates possible or definite left ventricular hypertrophy. Heart rate adjustment of ST-segment depression during exercise, performed by calculating the “Oldpeak” index, offers measurement of upsloping ST segments that may improve sensitivity with preservation of specificity from improved classification of patients with heart rate adjustment.

For the target category of the study, “HeartDisease” attribute, the total observations include 508 normal and 410 patient observations. The fact that these observations are chosen to be close to each other in terms of classification means that the algorithms are not prone to bias. Looking at the existing correlations with the target class for the attributes in the cardiovascular disease dataset, the results in Table 2 are obtained. However, we also set up a second dataset with 410 normal and 410 patient classes to check whether there was a problem with the fully balanced dataset in the experimental results. In order to avoid confusion in the study, 2 different datasets are denoted as Balanced: B, Unbalanced: UB to avoid confusion. Dataset B represents 410 normal 410 patient, while dataset UB represents 508 normal 410 patient.

Table 3. Feature correlation measurements for class of cardiovascular disease in the UB dataset after preprocessing.

Feature	Feature Type	Correlation Result
ST-Slope-Up	Numerical	-0.622164
ChestPainType-ATA	Numerical	-0.401924
MaxHR	Numerical	-0.400421
Cholesterol	Numerical	-0.232741
ChestPainType-NAP	Numerical	-0.212964
RestingECG-Normal	Numerical	-0.091580
ChestPainType-TA	Numerical	-0.054790
RestingECG-ST	Numerical	0.102527
RestingBP	Numerical	0.107589
FastingBS	Numerical	0.267291
Age	Numerical	0.282039
Sex-M	Numerical	0.305445
Oldpeak	Numerical	0.403951
ExerciseAngina-Y	Numerical	0.494282
ST-Slope-Flat	Numerical	0.554134

Regression analysis is a statistical technique for accommodating a cause-and-effect relationship. It is used for prediction (no prediction beyond the data used in the analysis), while correlation is used to determine the degree of the relationship (Asuero et al., 2006). In this study, assuming the number and dependency of the features, it is concluded that multiple regression analysis should be performed.

2.3. Preprocessing of Data

Preprocessing steps for data cleaning during data analysis are considered one of the essential steps in data-dependent studies in the literature. The dataset examined in the study is a mixed data source consisting of nominal and numerical values with 11 attributes. While 80% of a total of 701 observations were reserved for training, 20% were determined to be used in the testing phase. The correlation coefficient r revealed negative and positive correlation relationships for the target class, provided there were non-normalized features in the first step (Mintemur, 2021). The Label Encoder technique was used to digitize the nominal data. The components were standardized by removing the mean and scaling with the Standard Scaler, the next preprocessing step (Imad et al., 2022). The Standard Scaler technique is used to standardize the features. The correlation measurements between the components in the formed cluster and the target variable were calculated. Table 2 presents the new correlation values obtained. In this study, outlier data analysis and identification, which is another preprocessing step, was performed.

$$\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1} \qquad (1)$$

Q1 in Equation 1 is the first quartile of the data, 25% of the data lies between the minimum and Q1. Q3 is the third quarter of the data, meaning 75% of the data falls between the minimum and Q3. The outliers to be reduced after the calculated Q3 and Q1 values are obtained by applying the observations that are less than or equal to Q3+1.5*IQR for the upper limit and greater than or equal to Q1-1.5*IQR for the lower limit (Perez and Tah, 2020). While outlier data were in the observations, observations were 918, and with the removal of outliers, observations were 701. Figure 1 belongs to the correlation matrix between the features after removing outliers with the interquartile range technique. Negative measurements between values in the matrix indicate that it has the opposite relationship with the target variable.

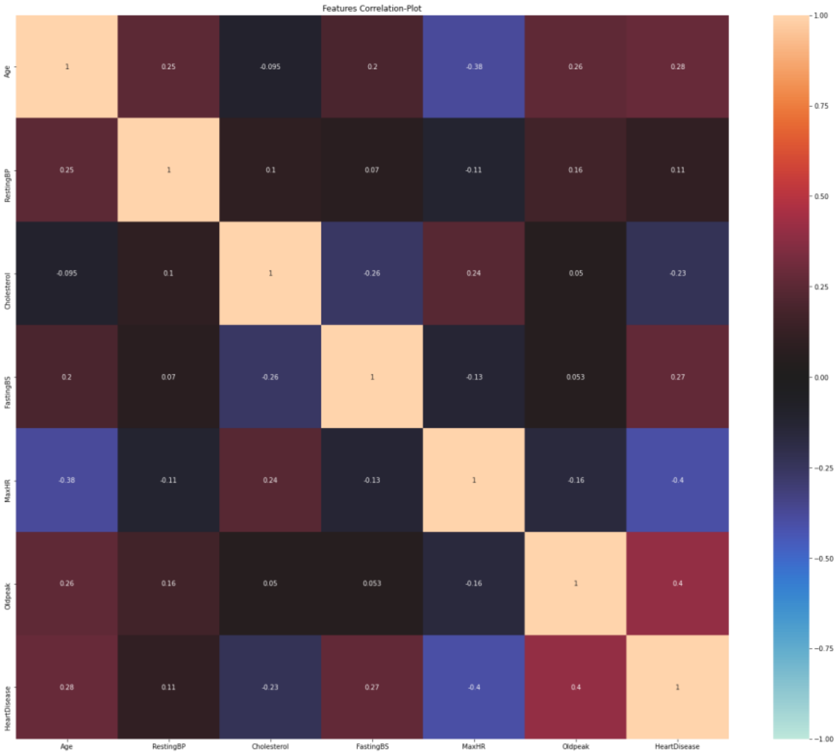

Figure 1. Identification of outliers in the data set with the interquartile range technique and the inter-feature correlation matrix in the UB dataset.

Accordingly, the diagonal is colored with the lightest color corresponding to +1, as there is ideal correspondence between the features. Measures with negative values in the matrix indicate they have the opposite relationship with the target variable. For example, a negative correlation exists between “Cholesterol” and the class “HeartDisease'”. After feature extraction according to the standardized observation data in the dataset, which was divided into training and test sets, the next preprocessing step was the dimensional reduction technique.

2.4. Dimensional Reduction Technique

The use of datasets with too many attributes for algorithms determined in ML projects leads to poor performance. The number of observations in the dataset should be high with the discovery of a certain amount of selection of features. Dimensional reduction techniques mean reducing unnecessary and redundant features in datasets. Reducing feature space with necessary feature selection and extraction ways is a proper statistical technique and a familiar method for discovering designs in high-dimensional data (Karamizadeh et al., 2013; Meng and Yang, 2012). PCA is one of the most famous techniques for reduction. To study a more down-dimensional space, the data is directed toward linear dimensionality reduction. The input data is centered. In the new variable space created by minimizing the cardiovascular dataset size, it is ensured that the most relevant features are in that space (Çil and Güneş, 2022). When Figure 2 is examined, it is concluded that maximum heart rate decreases with age and cardiovascular disease increases as maximum heart rate decreases. The "Age" and "MaxHR" attributes refer to the graph before and after pre-processing. As seen in the figure, correlation measurements were performed for all features. In this way, the connections of the features in the dataset with each other were also controlled formally. The data to be removed were determined by ranking the variance inflation factor and attribute values according to the principal component method. Instead of working with multiple original numerical features, linear combinations of them are obtained, paying attention to those that describe as many variations as possible from the original observations. Choosing linear combinations of predictors based on the maximum variance of the observations for the target variable "HeartDisease" was beneficial for prediction. Thus, the PCA transformation was carried out by providing dimension reduction. In PCA analysis, the error term is neglected in the calculation of the common factor variances of the features (Alkan, 2008).

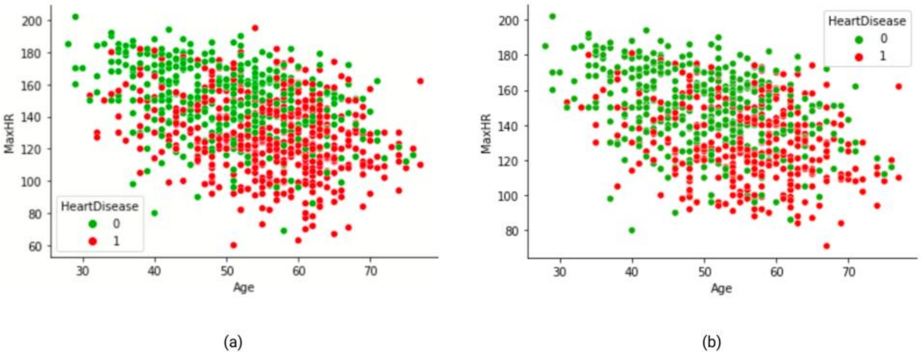

Figure 2. Correlation measures of age and maximum heart rate variables for cardiovascular disease. (a) correlation graph of age and maximum heart rate variables without pre-processing, (b) correlation graph of age and maximum heart rate variables as a result of pre-processing.

Figure 3 reveals the cumulative variance value by calculating the variance explained by the sum of the eigenvalues. In this step, 15 principal components were selected, and the variances explained by the components and the cumulated variance values were graphed. As can be seen, the variance of the first component is more meaningful than the other principal components. Therefore, the first 6 components may be sufficient to make sense of an average dataset. Components are calculated by capturing the variance in the data in the best way for dimension reduction with the PCA method. As seen in Figure 3, the plot shows the variance explained by each component against the number of components. According to these values, 6 principal components were selected as it is unnecessary to add additional components from the point where the curve flattens (Umargono et al., 2019). The curve breakpoint principle aims to select components that explain a large proportion of the total variance. Here, the point at which the plot bars and the curve become significantly flatter is designated as the break point. Therefore, component selection was performed where it did not provide a significant increase. Since the cumulative variance ratio reached sufficient saturation on this graph, 6 features were selected. The selection of these components is based on PCA analysis and the sum of the component loadings. The 6 most important features selected by PCA are Sex-F, Sex-M, RestingECG-ST, ST-Slope-Flat, RestingECG-Normal and RestingECG-LVH. Their values are 1.457506, 1.457506, 1.454983, 1.425109, 1.375829 and 1.322278 respectively.

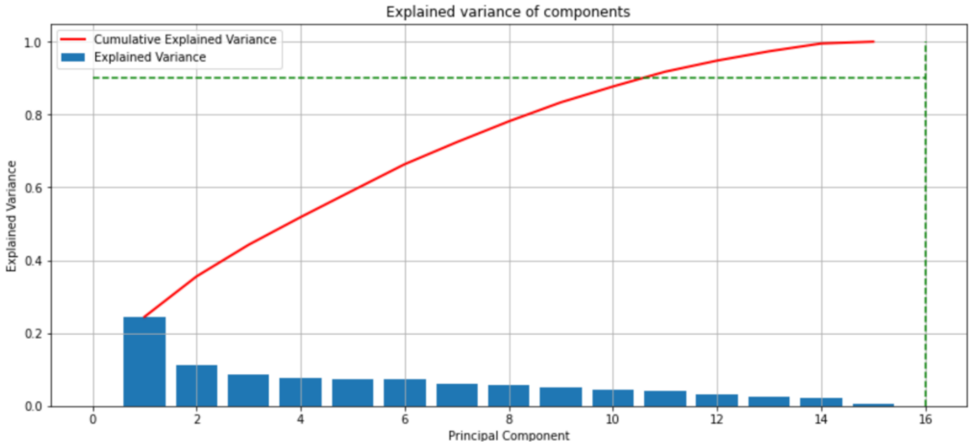

Figure 3. PCA analysis results total principal component count graph of the UB dataset.

2.5. Conventional Classification Techniques

After the preliminary preparation of the data, appropriate ML algorithms should be selected for the patterns to be found on the observations in the data sets. Classification is one of the supervised ML algorithms and is a frequently used task in studies with dependent features (Kaba and Kalkan, 2022).

Support Vector Classifier (SVC) is understanding from labeled training data to create estimations learning technique embedded in Structural Risk Minimization (SRM); it is among the well-known methods in machine learning (Cervantes et al., 2020; Moosaei et al., 2023). The support vectors are also recollection influential since they use a subset of the training topics. SVC's extraordinary generalization ability, optimal solution, and discriminating power have recently attracted attention. An infinite number of hyperplanes for linear separation of data are called optimal separation hyperplanes (Cristianini and Shawe-Taylor, 2000).

Naive Bayes Classifier (NBC) is a supervised learning algorithm that functions with the belief of "naive" dependent sovereignty between each couple with attributes and the class attribute. The training process of NBC is to predict the class preliminary probability based on the training set Zhang, 2004). GaussianNBC implements the Gaussian NBC algorithm for classification. The GaussianNBC classifier can be operated when the likelihoods of the features give the exact consequences (Pushpakumar et al., 2022). The classification problem in the study is to predict whether heart disease is present or absent.

Decision Trees are generally more rapid than artificial neural networks but do not have the suppleness to parameters Like SVCs, Decision Tree Classifiers (DTC) are practical techniques for appropriately challenging datasets (Singh et al., 2022). The aim is to make a technique that foresees a target variable. Additionally, the deeper the tree, the more complicated the rules and the more suitable the approach (Géron, 2022). The study handled this problem, and community learning techniques were used.

K-nearest Neighbor (KNN) techniques are an approach that is easy to implement but often runs quite slowly when the input dataset is huge. It is susceptible to extrinsic parameters. This classification algorithm, which has low efficiency due to lazy learning, is effective despite being a simple method (Guo et al., 2003). In this case, selecting the k parameter well is crucial to perform successfully. These are the resemblance measure between two data topics and the k's choice. The typical consequence of the foremost question is that various applications require various length sizes (Zhang, 2010; Zhang et al., 2017). Therefore, the choice of the k value merely uses the Euclidean length to compute the resemblance (Qin et al., 2007).

Logistic Regression (LR) is a particular point of approach with Binomial or Bernoulli distribution. The numerical result of the LR, which is the estimated likelihood, is used as a model. It is believed that target y_i accepts values in the set 0-1 for data point i. Once deployed, LR's prediction method predicts the probability of the positive class. LR is usually utilized to indicate the likelihood that a sample belongs to a specific class. If the estimated likelihood is greater than 50%, the model estimates that the sample belongs to that class (Géron, 2022).

Discriminant Analysis (DA) is one of the prevalent techniques for extracting the best features. It is developed as a problem to find an optimal value. It is also helpful but must be developed for nonlinear cases for more complicated ones (Kurita et al., 2009). Linear Discriminant Analysis (LDA) and Normal Discriminant Analysis (NDA) generalize Fisher's linear discriminant. Also, the algorithm supplies a Gaussian density to all types (Tharwat et al., 2017).

2.6. Ensemble Learning Techniques

Ensemble learning techniques are divided into two: bagging and boosting. In the bagging technique, new trees are created by repeatedly pulling samples from the dataset to be replaced. Then, a community emerges with the created trees. The boosting technique makes inferences from the ensemble by giving different weights to the dataset. One way to obtain various approaches is to utilize diverse techniques. Another technique is using the exact technique for each estimator. When sampling with replacement, this approach is called bagging. (Zhang et al., 2017). The Bagging Classifier (BC) is presented via Leo Breiman in 1994. This technique can use classification and regression methods. It is developed to enhance the strength and precision of ML approaches used. BC has received much attention for its simple implementation and increased accuracy. Therefore, it can be considered a "smoothing operation", which is advantageous when improving the forecast performance of trees (Breiman, 2001; Géron, 2022).

A RF Classifier (RFC) is a group DTCs commonly trained by the bagging technique and generally with a maximum sample set. Rather than creating a GC and giving a DTC to it, it will likely utilize the RFC. The RFC algorithm provides an additional lacking pattern when growing trees; it explores the most helpful attribute. This source of randomness aims to reduce the variance of the forest predictor (Breiman, 2001). The prevailing opinion of most boosting strategies is to train estimators, each attempting to repair the earlier one. Many boosting methods are known, but the most famous are Adaptive Boosting (AdaBoostC) and Gradient Boosting (GBClassifier, GBC). The GBC algorithm makes a progressively forward extra model. At each stage, the n class number regression trees are provided for the adverse gradient of the loss function. The model adds estimators sequentially to an ensemble, each updating the previous one (Géron, 2022). Extreme Gradient Boosting Classifier (XGBC), the optimized version of the GBC, is highly enhanced and adaptable. Also, the XGBC is frequently considered crucial. This algorithm, which has a place in the literature as an ensemble learning algorithm, is considered excellent. A genetic algorithm has optimized the hyperparameter vector of the XGBC approach to enhance the forecast exactness and trustworthiness of the XGBoost model (Gu et al., 2022). An AdaBoostC is introduced and utilized to estimate the training set (Hastie et al., 2009). AdaBoostC has been demonstrated to be a thriving learning approach; it iteratively produces different vulnerable trainees and includes their results using the weighted plurality voting rule (Sun et al., 2016).

2.7. Artificial Neural Networks

DL is a branch of ML and, thus, pattern recognition and emanates from Artificial Neural Networks (ANNs) that affect the design of moving and processing data between neurons. For the sequential ANNs to be created, the model consisting of a single-layer stack connected sequentially is built. Since the first layer in the model will give an input vector, after the input size has been determined, the batch size should be chosen depending on the samples for the dataset. Then, a model suitable for the problem should be constructed. In this step, dense hidden layers with a certain number of neurons are added. It will use the Rectified Linear Unit (ReLU). The basic unit of deep neural networks are layers, which are data processing modules to be considered as filters for data. The data is taken as raw data to the layers for neural networks and reaches a level that will be more useful. Relevant layers have been added for the neural network to be built, and the selection of the activation function and loss function has been carried out (Chollet, 2021; Géron, 2022). The neural network in Figure 4 is obtained as a result of adding the relevant Dense layers by choosing a Binary Cross Entropy (BCE) loss function. This loss function performs the calculation of the cross-entropy loss between the real labels and the predicted labels.

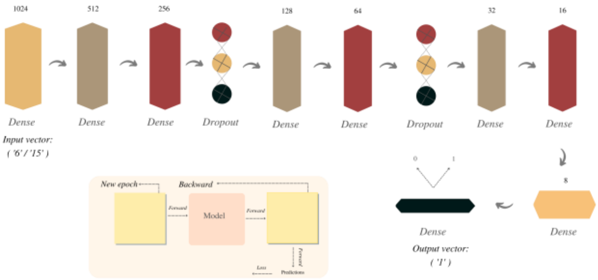

Figure 4. Neural network architecture suitable for cardiovascular disease prediction.

Since there is no categorical classification problem, it can be considered appropriate as a loss function since the cardiovascular disease result is 0 / 1. As activation functions, ReLU and Sigmoid functions were used respectively. Also, Mean Squared Error (MSE), which measures the mean of squares of errors is used. Thus, the mean of the sum of the squares of each difference between the predicted value and the true value was obtained. The network was trained for batch size: 2, optimizer: Adam, kernel initializer: Glorot uniform for a total of 500 epochs. While training the ANN, validation loss was continuously checked using Early Stopping techniques and training was terminated when the network stopped learning.

2.8. Hyperparameter Optimization with Grid Search Cross Validation and Randomized Search Cross Validation Techniques

In an ML study, hyperparameter optimization is the last step before experimental findings. Grid Search Cross Validation (GridSearchCV) is one of the various methods to discover a thriving and robust parameter for an algorithm. Grid search is a parameter-tuning approach to build and evaluate the selected model parameters (Ranjan et al., 2019). The n estimator parameters used in the approach were chosen at the level [10, 50, 100, 250, 500] (number of trees) to be transmitted to the classifier to be trained. In the evaluation procedure for the hyperparameter improvement part of the study, the model selection was provided by the RepeatedStratifiedKFold technique (Kramer, 2016). The parameter n is 3, and the number of folds is 10. For the values determined as the best parameters found in the AdaBoostC model as a result of GridSearchCV, the learning rate was 0.1, n estimators were 250, and the model result reached 87% accuracy. For the values determined as the best parameters found in the AdaBoostC model as a result of GridSearchCV, the learning rate was 0.1, n estimators were 250, the model result reached 87% accuracy. For the RF classifier, 64 candidates are selected for 10 folds in the same way and the algorithm is run. The maximize feature was 3, the minimum sample separation was 10, and the total number of trees was 200, and the best result was achieved with 90% accuracy for the classifier. Randomized Search Cross Validation (RandomizedSearchCV) is another method used for hyperparameter optimization. This method is similar to GridSearchCV, but requires less computational cost because it performs parameter searches over random samples rather than trying all possible combinations. The parameters of the RandomizedSearchCV model are optimized by a cross-validated search across many options, and unlike GridSearchCV, where all possible parameter values are tested, this method only tries a small subset of them from the selected distributions (Sharma et al., 2023). Accordingly, the method was applied for RFC and AdaboostC algorithms respectively. For the RFC algorithm, as in GridSearchCV, n estimators were trained to be 100, min samples split 20 and max features 3. As a result of testing the test set, an accuracy of 84.78% was obtained. In addition, AdaBoostC algorithm has set its best parameters according to RandomizedSearchCV technique with n estimators 100, learning rate 0.1. In this direction, the necessary training was performed and tested on the test set and the accuracy result was obtained as 84.42%.

2.9. Performance Evaluation Metrics

In ML studies, the confusion matrix reveals the connection between the class's ground truth classes and the model's estimated classes. Assessment of algorithm implementation is according to precision, recall, f1-score, and accuracy values in the equations in Equation 2, Equation 3, Equation 4, and Equation 5. Precision and recall metrics are often inversely proportional, as seen in Equation 2. F1-score is obtained from the harmonic average of the consequences in the equations to validate the optimization methods (Keser and Keskin, 2022; Tekin et al., 2022).

$$Precision\ = \frac{TP}{TP + FP} \tag{1}$$

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

$$F1 - score\ =\ 2x\frac{Precision * Recall}{Precision + Recall} \tag{3}$$

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

In the research, the Receiver Operator Characteristic (ROC) curve is often employed to demonstrate the efficiency of an algorithm. The ROC curve gives detailed knowledge about algorithm implementation and can be outlined as a single number area under the ROC Curve (AUC) (Meseci et al., 2022). AUC in Figure 5, revealed as an approach to calculate the performance, determines the accuracy of prediction in various techniques (Muschelli, 2020).

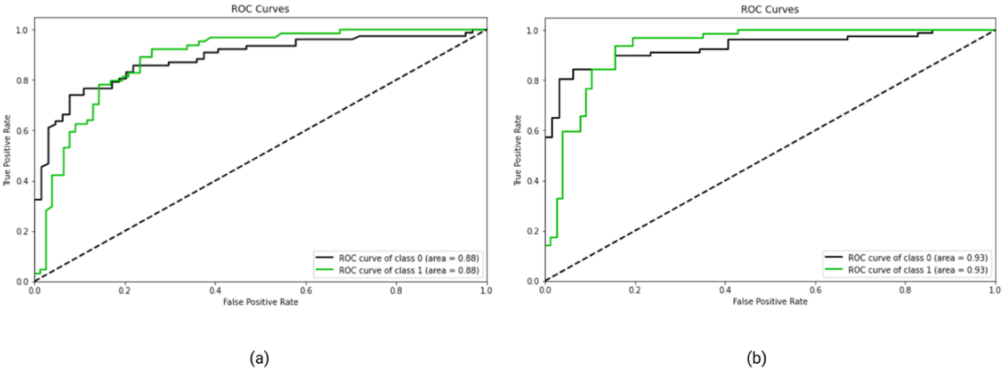

Figure 5. AUC graphs under the ROC curve in line with true positive and false positive rates for the worst and best classifier from the prediction scores in the UB dataset. (a) AUC-ROC graph for GBC, (b) AUC-ROC graph for the KNN classifier.

3. Experimental Results

Experimental results are the quantitative values obtained as a result of the studies performed during the evaluation of different types of ML models. This section includes the experimental findings before and after the pre-processing, as well as the performance results depending on the change of the attribute value. In addition to conventional classifiers in the literature such as SVC, NBC, DTC, KNN, LR, LDA, BaggingClassifier, RFC, GBC, AdaBoostC, etc. tree-based ensemble methods and ANN models such as were used. Thus, experimental findings that will contribute to the academic literature were obtained. As a result, while the weighted average accuracy was 71.0% for these two classes, the macro average accuracy was 70.0%. On top of that, when the classifier model was applied for the "linear, radial basis function" kernels with the GridSearchCV technique, the best score was obtained as 84.88% as a result of parameter selection. As a result of the cross-validation technique, the precision value for the "Normal" label was 85.0%, the recall value was 78.0% and the f1-score was 81.0%. As a result, the weighted average and macro accuracy for these two classes was 83.0%. Table 3 represents the experimental results obtained according to the features in the correlation matrix in Figure 1 after PCA analysis.

Table 4. Performance comparison table of preprocessing conventional and ensemble learning algorithms for the UB dataset.

Algorithm	Precision	Recall	F1-score	Accuracy
GBC	81.0%	80.9%	80.9%	81.0%
XGBC	82.0%	82.0%	82.0%	82.3%
DTC	84.0%	83.7%	83.7%	83.7%
RFC	**86.0%**	**86.5%**	**86.5%**	**86.5%**
KNN	**86.5%**	**86.0%**	**86.0%**	**87.0%**

In the next stage of the study, in addition to the previously mentioned features, a classification process was carried out with artificial indicators included in the dataset. In this case, instead 6 attributes, the categorical data in the data set was transformed into 15 artificial variables. Table 4 is based on the performance comparison of the classifier algorithms through 6 features. Table 5 shows the classification task results of 15 features included in the dataset as a result of the required pre-processing technique. In this case, performance improvement was observed for many classifiers and the algorithm with the best score was updated to GaussianNBC. Accuracy, ROC-AUC values are observed to show an improvement of 3.0%. In Table 6, the learning process is completed using BCE loss for 100 iterations. When the findings were analyzed, it was followed that the "Normal" class learned better, as expected. In Table 6, using the MAE loss metric, it is trained under the same conditions as the neural network used in BCE loss.

Table 5. Performance comparison table of conventional and ensemble learning algorithms by feature reduction for the UB dataset.

Algorithm	Accuracy	ROC	AUC
DTC	83.0%	82.0%	82.0%
LDA	84.0%	85.0%	85.0%
AdaBoostC	85.0%	85.0%	85.0%
KNN	85.0%	86.0%	86.0%
LR	85.0%	86.0%	86.0%
LinearSVC	85.0%	86.0%	86.0%
GaussianNBC	86.0%	86.0%	86.0%
RFC	**88.0%**	**87.0%**	**87.0%**
SVC	**88.0%**	**87.0%**	**87.0%**

When figure is carefully observed, it is concluded that the correlation connections increase with the variables "ChestPainType", "RestingECG" and "ST-Slope", which are not included in the 6-attribute classification problem. The features in the correlation matrix were used in classification and new values were added to the experimental findings.

Table 6. Performance comparison table of conventional and ensemble learning algorithms by feature increase for the UB dataset.

Algorithm	Accuracy	ROC	AUC
DTC	85.0%	85.0%	85.0%
AdaBoostC	87.0%	87.0%	87.0%
KNN	89.0%	89.0%	89.0%
SVC	89.0%	89.0%	89.0%
LDA	89.0%	89.0%	89.0%
LR	89.0%	89.0%	89.0%
LinearSVC	89.0%	89.0%	89.0%
RFC	90.0%	90.0%	90.0%
GaussianNBC	**91.0%**	**91.0%**	**91.0%**

Accordingly, evaluating an ANN algorithm is more suitable than many classifier approaches. When the results in the table are examined carefully, the precision value for the "HeartDisease" class is low, but the recall value is quite high. Therefore, it is concluded that there are too many false positive values. Contrary to Figure 1, a correlation matrix with more features is created and given in Figure 5.

Table 7. Performance comparison table of conventional and ensemble learning algorithms by feature reduction for the UB dataset.

Target	Precision	Recall	F1-score	Accuracy
HeartDisease-BCE	85.0%	84.0%	84.0%	87.0%
Normal-BCE	**90.0%**	**91.0%**	**91.0%**	**88.0%**
HeartDisease-MAE	82.0%	93.0%	87.0%	88.0%
Normal-MAE	**95.0%**	**85.0%**	**90.0%**	**89.0%**

In addition to the UB dataset, the algorithms used were also applied to the B dataset. Table 8 presents the performance comparison table of the conventional and ensemble learning algorithms for dataset B. According to the table, AdaBoostC, RF Classifier and SVC algorithms show the highest performance with a slight difference. In particular, RF Classifier and AdaBoostC algorithms outperform the other algorithms with 89.0% accuracy, ROC, and AUC values.

Table 8. Performance comparison table of conventional and ensemble learning algorithms for the B dataset.

Algorithm	Accuracy	ROC	AUC
DTC	82.0%	82.0%	81.5%
LDA	85.3%	85.0%	85.0%
LinearSVC	85.3%	85.4%	85.3%
LR	85.3%	85.4%	85.4%
XGBC	87.2%	87.3%	87.2%
KNN	87.2%	87.3%	87.2%
GaussianNBC	87.8%	87.7%	87.9%
GBC	88.4%	87.3%	87.3%
SVC	88.4%	88.5%	88.4%
AdaBoostC	89.0%	89.0%	89.0%
RFC	**89.0%**	**89.1%**	**89.0%**

These results show that ensemble methods and SVC algorithm perform better than other conventional algorithms and their performance improves.

In the feature increase process, 6 features obtained using PCA were transformed into artificial variables. This was done to better represent the data and improve the performance of the classification algorithms. The artificial variables were created using linear combinations of the original attributes, thus adding additional information to the dataset. As can be seen in Figure 5, new variables were selected for the main selected principal components taken from their internal categories. These attributes include interaction terms and higher-order polynomials of the original features. For example, the “Normal” and “ST” categories of the RestingECG attribute were taken as additional features, while the ATA, TA, and NAP attributes were added for ChestPainType, resulting in a total of 15 artificial variables.

Experimental findings show that feature reduction and increase techniques and hyperparameter optimization significantly improve the performance of the algorithms. The performance of the classifiers was significantly improved by using feature reduction and increase techniques. In particular, the best results were obtained when feature increase was applied by adding artificial variables. After these procedures, the Naive Bayes algorithm showed the highest performance with 91% accuracy. The best results were obtained with Naive Bayes, AdaBoostC and Random Forest algorithms. This study demonstrates the effectiveness of machine learning techniques in cardiovascular disease detection.

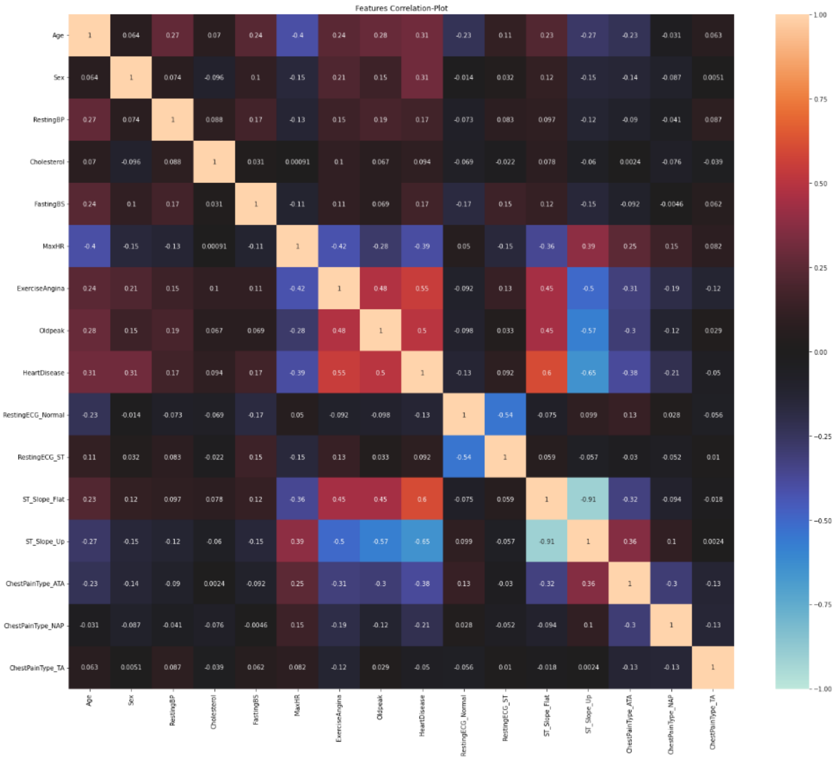

Figure 5. Correlation matrix of increasing features as a result of adding artificial variables for the UB dataset.

4. Conclusions

In this study, detection, and classification of cardiovascular disease with many algorithms was performed on a mixed dataset. Common algorithms in the literature were selected for classification and their success was increased according to the performance results obtained in similar studies. For this, both feature reduction and feature enhancement were applied by performing more than one pre-processing. In addition, statistically outlier data were cleaned with the IQR technique, and then improved by hyperparameter optimization with the GridSearchCV technique, a successful originality was demonstrated with a different approach compared to similar studies. With many ensemble learning techniques, algorithmically diverse results have been achieved. During the study, the correlation matrices were evaluated during each step, and the steps that gave the best performance were progressed in the process.

In the experimental findings section, all experiments carried out were meticulously supported by tables and figures. As a result of the study, the classifiers that gave the best results were GaussianNBC with 91.0%, RF Classifier with 88.0%, SVC with 88.0% and ANN model with 89.0% in the UB dataset. In addition, by providing hyperparameter optimization with the GridSearchCV technique, an improvement of approximately 3.0% was achieved in the results obtained in the experimental findings. Besides, RF Classifier was the algorithm that gave the highest score to the comparison table for dataset B. When the RF Classifier algorithm applied for the B dataset was compared with the result obtained for the UB dataset, it was concluded that there was a 1% performance increase.

This study successfully classified cardiovascular disease as a laborious and time-taking situation in the health field. Future studies and research aim to obtain more successful performances by minimizing the current error margin for detecting health problems, which is a difficult task.

5. Discussion

This study provides various machine learning and ensemble learning techniques are used for the detection and analysis of cardiovascular diseases. The results obtained are significant when compared to existing work in the literature. In this section, we will discuss the place and contributions of our work in the literature from a broad perspective. Research on the detection and analysis of cardiovascular diseases has made significant progress in recent years with the use of machine learning techniques. In their study, Bektaş and Babur (Bektaş and Babur, 2016) evaluated the performance of various machine learning algorithms for breast cancer diagnosis and obtained the highest accuracy rate of 90.7% with the RFC algorithm. Cihan (Cihan, 2018) demonstrated the effectiveness of the RFC algorithm with an accuracy rate of 86.1% using Cleveland and Hungary datasets.

In contrast to these studies, in our study, different reduction and augmentation techniques were applied for the features in the dataset for the detection of cardiovascular diseases and more algorithms were used. Feature reduction and enhancement techniques are frequently used to improve the performance of machine learning models. In our study, feature reduction was performed using PCA and then feature increase was applied by adding artificial variables. In particular, the Naive Bayes algorithm showed the highest performance with an accuracy of 91.0%. This result shows that the Naive Bayes algorithm can be effectively used in such classification problems.

In our study, bagging and boosting techniques and various ensemble learning algorithms were used. RFC and AdaBoostC algorithms are frequently used in the literature and have shown high performance (Breiman, 2001; Hastie et al., 2009). In this study, the RFC algorithm showed high performance with an accuracy of 88.0%. This result is consistent with the findings in the literature and confirms that the RFC algorithm is an effective method for cardiovascular disease detection. Moreover, ANN and deep learning techniques have achieved significant success in the medical field in recent years. In our study, the ANN model showed a high performance with an accuracy of 89.0%. This result shows that deep learning techniques are a powerful tool for the detection of cardiovascular diseases.

One of the most important contributions of this study is the comprehensive evaluation of the effectiveness of different machine learning and ensemble learning techniques in cardiovascular disease detection. In particular, the high accuracy rates achieved using feature augmentation with artificial variables and hyperparameter optimization are an important contribution to the literature. Our recommendation for future work is to improve the generalizability of the models using larger and more diverse datasets and to test different attribute reduction and augmentation techniques. Furthermore, evaluating the performance of deep learning models on more complex and larger datasets may contribute to better results in the detection of cardiovascular diseases.

References

Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. 2, 433–459.

Akman, M., Civek, S., 2022. Dünyada ve Türkiye'de kardiyovasküler hastalıkların sıklığı ve riskin değerlendirilmesi. J. Turk. Fam. Physician 13, 21–28.

Alkan, Ö., 2008. Temel bileşenler analizi ve bir uygulama örneği. Atatürk Üniversitesi Sos. Bilim. Enstitüsü İşletme Anabilimdalı Üksek Lisans Tezi Erzurum 125s.

Asuero, A.G., Sayago, A., González, A.G., 2006. The correlation coefficient: An overview. Crit. Rev. Anal. Chem. 36, 41–59.

Badem, H., 2019. Parkinson Hastaliğinin Ses Sinyalleri Üzerinden Makine Öğrenmesi Teknikleri ile Tanimlanmasi. Niğde Ömer Halisdemir Üniversitesi Mühendis. Bilim. Derg. 8, 630–637.

Bektaş, B., Babur, S., 2016. Makine Öğrenmesi Teknikleri Kullanılarak Meme Kanseri Teşhisinin Performans Değerlendirmesi.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., Lopez, A., 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing 408, 189–215.

Chollet, F., 2021. Deep learning with Python. Simon and Schuster.

Cihan, Ş., 2018. Koroner arter hastalığı riskinin makine öğrenmesi ile analiz edilmesi (PhD Thesis). Yüksek Lisans Tezi. Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü, Kırıkkale.

Çil, E., Güneş, A., 2022. Makine öğrenmesi algoritmalarıyla kalp hastalıklarının tespit edilmesine yönelik performans analizi. İstanbul Aydin Üniversitesi Dergisi Anadolu Bil Meslek Yüksekokulu.

Cristianini, N., Shawe-Taylor, J., 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

Géron, A., 2022. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc.

Gregg, L.P., Hedayati, S.S., 2018. Management of traditional cardiovascular risk factors in CKD: what are the data? Am. J. Kidney Dis. 72, 728–744.

Gu, Z., Cao, M., Wang, C., Yu, N., Qing, H., 2022. Research on Mining Maximum Subsidence Prediction Based on Genetic Algorithm Combined with XGBoost Model. Sustainability 14, 10421.

Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K., 2003. KNN model-based approach in classification. In: On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings. Springer, pp. 986–996.

Hastie, T., Rosset, S., Zhu, J., Zou, H., 2009. Multi-class adaboost. Stat. Interface 2, 349–360.

Imad, M., Abul Hassan, M., Hussain Bangash, S., Naimullah, 2022. A Comparative Analysis of Intrusion Detection in IoT Network Using Machine Learning. In: Big Data Analytics and Computational Intelligence for Cybersecurity. Springer, pp. 149–163.

Kaba, G., Kalkan, S.B., 2022. Kardiyovasküler Hastalık Tahmininde Makine Öğrenmesi Sınıflandırma Algoritmalarının Karşılaştırılması. İstanbul Ticaret Üniversitesi Fen Bilim. Derg. 21, 183–193.

Kara, K., Çınar, S., 2011. Diyabet bakım profili ile metabolik kontrol değişkenleri arasındaki ilişki. Kafkas J Med Sci 1, 57–63.

Karamizadeh, S., Abdullah, S.M., Manaf, A.A., Zamani, M., Hooman, A., 2013. An overview of principal component analysis. J. Signal Inf. Process. 4, 173.

Keser, S.B., Keskin, K., 2022. Ağırlıklı Oy Tabanlı Topluluk Sınıflandırma Algoritması ile Göğüs Kanseri Teşhisi. Mühendis. Bilim. Ve Araştırmaları Derg. 4, 112–120.

Kramer, O., 2016. Scikit-Learn. In: Kramer, O. (Ed.), Machine Learning for Evolution Strategies, Studies in Big Data. Springer International Publishing, Cham, pp. 45–53.

Kurita, T., Watanabe, K., Otsu, N., 2009. Logistic discriminant analysis. IEEE International Conference on Systems, Man and Cybernetics. Presented at the 2009 IEEE International Conference on Systems, Man and Cybernetics - SMC, IEEE, San Antonio, TX, USA, pp. 2167–2172.

Li, L., Zhou, Z., Bai, N., Wang, T., Xue, K.-H., Sun, H., He, Q., Cheng, W., Miao, X., 2022. Naive Bayes classifier based on memristor nonlinear conductance. Microelectron. J. 129, 105574.

Liashchynskyi, Petro, Liashchynskyi, Pavlo, 2019. Grid search, random search, genetic algorithm: a big comparison for NAS. ArXiv Prepr. ArXiv191206059.

Lopez, E.O., Ballard, B.D., Jan, A., 2022. Cardiovascular disease. In: StatPearls [Internet]. StatPearls Publishing.

Malik, P., Pathania, M., Rathaur, V.K., 2019. Overview of artificial intelligence in medicine. J. Fam. Med. Prim. Care 8, 2328.

Meng, J., Yang, Y., 2012. Symmetrical two-dimensional PCA with image measures in face recognition. Int. J. Adv. Robot. Syst. 9, 238.

Meseci, E., Ozkaynak, E., Dilmac, M., Ozdemir, D., 2022. PDC Dünya Dart Şampiyonası Karmaşık Ağlarında Komşuluk Tabanlı Bağlantı Tahmini. 5th Int. Conf. Data Sci. Appl. ICONDATA'22.

Mintemur, Ö., 2021. Doğrusal regresyonla vücut yağ tahmininde korelasyon türlerinin etkisi. EurasianSciEnTech 2021.

Moosaei, H., Ganaie, M.A., Hladík, M., Tanveer, M., 2023. Inverse free reduced universum twin support vector machine for imbalanced data classification. Neural Netw. 157, 125–135.

Muschelli, J., 2020. ROC and AUC with a Binary Predictor: a Potentially Misleading Metric. J. Classif. 37, 696–708.

Perez, H., Tah, J.H., 2020. Improving the accuracy of convolutional neural networks by identifying and removing outlier images in datasets using t-SNE. Mathematics 8, 662.

Platt, J., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv. Large Margin Classif. 10, 61–74.

Pushpakumar, R., Prabu, R., Priscilla, M., Renisha, P.S., Prabu, R.T., Muthuraman, U., 2022. A Novel Approach to Identify Dynamic Deficiency in Cell using Gaussian NB Classifier. In: 2022 7th International Conference on Communication and Electronics Systems (ICCES). IEEE, pp. 31–37.

Qin, Y., Zhang, S., Zhu, X., Zhang, J., Zhang, C., 2007. Semi-parametric optimization for missing data imputation. Appl. Intell. 27, 79–88.

Ranjan, G.S.K., Verma, A.K., Radhika, S., 2019. K-nearest neighbors and grid search cv based real time fault monitoring system for industries. In: 2019 IEEE 5th International Conference for Convergence in Technology (I2CT). IEEE, pp. 1–5.

Sharma, N., Malviya, L., Jadhav, A., Lalwani, P., 2023. A hybrid deep neural net learning model for predicting Coronary Heart Disease using Randomized Search Cross-Validation Optimization. Decis. Anal. J. 9, 100331.

Singh, N., Jena, S., Panigrahi, C.K., 2022. A novel application of Decision Tree classifier in solar irradiance prediction. Mater. Today Proc. 58, 316–323.

Sun, B., Chen, S., Wang, J., Chen, H., 2016. A robust multi-class AdaBoost algorithm for mislabeled noisy data. Knowl.-Based Syst. 102, 87–102.

Tekin, B.Y., Ozcan, C., Pekince, A., Yasa, Y., 2022. An enhanced tooth segmentation and numbering according to FDI notation in bitewing radiographs. Comput. Biol. Med. 146, 105547.

Tharwat, A., Gaber, T., Ibrahim, A., Hassanien, A.E., 2017. Linear discriminant analysis: A detailed tutorial. AI Commun. 30, 169–190.

Umargono, E., Suseno, J.E., S. K., V.G., 2019. K-Means Clustering Optimization using the Elbow Method and Early Centroid Determination Based-on Mean and Median: In: Proceedings of the International Conferences on Information System and Technology. Presented at the International Conferences on Information System and Technology, Scitepress-Science and Technology Publications, Yogyakarta, Indonesia, pp. 234–240.

Vatansever, B., Aydın, H., Çetinkaya, A., 2021. Genetik algoritma yaklaşımıyla Öznitelik seçimi kullanılarak makine Öğrenmesi algoritmaları ile kalp hastalığı tahmini. J. Sci. Technol. Eng. Res. 2, 67–80.

Veranyurt, Ü., Deveci, A., Esen, M.F., Veranyurt, O., 2020. Makine Öğrenmesi Teknikleriyle Hastalık Sınıflandırması: Random Forest, K-nearest Neighbour ve Adaboost Algoritmaları Uygulaması. Uluslar. Sağlık Önetimi Ve Strat. Araşt. Derg. 6, 275–286.

Zein Elabedin Mohammed, A., Osama Fathy Kayed, M., Samy Abd El-Samee, M., 2020. Heart rate recovery time after excercise stress test in diabetic patients with suspected coronary artery disease. Al-Azhar Med. J. 49, 1845–1852.

Zhang, H., 2004. The optimality of naive Bayes. Aa 1, 3.

Zhang, S., 2010. KNN-CF approach: Incorporating certainty factor to knn classification. IEEE Intell Inform. Bull 11, 24–33.

Zhang, S., Li, X., Zong, M., Zhu, X., Cheng, D., 2017. Learning k for knn classification. ACM Trans. Intell. Syst. Technol. TIST 8, 1–19.