World Academy of Science, Engineering and Technology
International Journal of Health and Medical Engineering
Vol:18, No:3, 2024

An Approach for the Prediction of Cardiovascular Diseases

Nebi Gedik

Open Science Index, Health and Medical Engineering Vol:18, No:3, 2024 publications.waset.org/10013540.pdf

***Abstract*—**Regardless of age or gender, cardiovascular illnesses are a serious health concern because of things like poor eating habits, stress, a sedentary lifestyle, hard work schedules, alcohol use, and weight. It tends to happen suddenly and has a high rate of recurrence. Machine learning models can be implemented to assist healthcare systems in the accurate detection and diagnosis of cardiovascular disease (CVD) in patients. Improved heart failure prediction is one of the primary goals of researchers using the heart disease dataset. The purpose of this study is to identify the feature or features that offer the best classification prediction for CVD detection. The support vector machine classifier is used to compare each feature's performance. It has been determined which feature produces the best results.

***Keywords*—**Cardiovascular disease, feature extraction, supervised learning, support vector machine.

I. Introduction

A variety of conditions affecting the heart or blood vessels, including coronary heart disease, are grouped together under the phrase "cardiovascular disease." Symptoms of CVD disease might vary depending on the situation, but they can include pain or tightness in the chest, discomfort, weariness, disorientation, and swelling in the legs or abdomen [1]. While chest and arm discomfort are the most common sign of CVD, some people may not experience any symptoms at all (particularly in the early stages of the disease) [2]. Healthcare systems can benefit from the application of machine learning and artificial intelligence models to help with patient detection and diagnosis of CVD. This could allow for early and more specialized therapy for those who are expected to develop CVD, as well as reduce strain on healthcare systems [3]-[5]. Several algorithms are presented by researchers to predict CVDs utilizing various datasets (demographic data, medical history, clinical examination results, laboratory test results, images, etc.) and methodologies [6], [7]. Based on the patient's clinical characteristics, a hybrid decision support system is developed [8] that may help in the early diagnosis of cardiac disease. To deal with the missing values, the authors employ the multivariate imputation by chained equations approach. The selection of appropriate features from the supplied dataset is done using a hybridized feature selection technique that combines recursive feature elimination with the Genetic Algorithm (GA). SMOTE (Synthetic Minority Oversampling Technique) and conventional scalar approaches are also employed for data pre-processing. AdaBoost, logistic regression, Naïve Bayes, random forests, and support vector machines are used as classifiers to complete the development of the proposed hybrid system. A machine learning framework is proposed in [9] to use many algorithms to estimate the likelihood of getting heart disease. Five algorithms are used to run the framework: Hoeffding decision tree, Naïve Bayes, support vector machine, random forest, and Logistic Model Tree (LMT). The model is trained and tested using data from the Cleveland dataset. An approach called orthogonal local preserving projection (OLPP) is presented in [10] to lower the function dimension of the high-dimensional input data. The dimension reduction boosts the prediction rate by merging the Group Search Optimization (GSO) technique with the Levenberg-Marquardt (LM) training method for the neural network. The LM training technique finds the optimal network parameters, such as weights and bias, that minimize error when applied to the optimization issue. The performance measurements of accuracy, sensitivity, and specificity are coupled with the optimization technique's ultimate output.

The authors [10] identify and forecast human heart disease using the Cleveland heart disease dataset, which comprised six samples with incomplete data and 303 records overall. Only 13 of the 76 features that are initially included in the data are likely to be mentioned in any papers that are published. The influence of the condition is explained in the remaining feature. Another general dataset that researchers use for the prediction procedure is Z-Alizadeh Sani. It has 55 input components, 303 patient data points, and a class label variable for every patient. For the final dataset, it is employed nine machine-learning classifiers, both with and without hyper-parameter adjustments. The authors also standardize the dataset, perform the necessary pre-processing, and adjust hyper-parameters to guarantee precision on the standard dataset for cardiac disease. The authors also use the k-fold cross-validation method for machine learning algorithm training and validation. In the end, the experiment findings show that improving the hyper-parameters increases the prediction classifiers' accuracy. The standardization and hyper-parameter tuning of the machine learning classifiers yield notable outcomes. Numerous criteria are utilized, including F-measure, specificity, sensitivity, and classification accuracy, to assess the effectiveness of algorithms. A prediction model for heart disease is provided by [11]. The specific goals are to save lives and reduce the incidence of heart attacks. The National Cardiovascular Disease Surveillance (NCDS) System and the Cleveland database, which also contains data on heart illnesses, are the two databases that deal with cardiac conditions. The four pooled

Nebi Gedik is with the University of Health Sciences, Institute of Hamidiye Health Sciences, Turkey (e-mail: nebi.gedik@sbu.edu.tr).

Open Science Index, Health and Medical Engineering Vol:18, No:3, 2024 publications.waset.org/10013540.pdf

databases that make up the Cleveland heart disease dataset for the study are VA Long Beach, Switzerland, Hungary, and Switzerland. There are fourteen attributes in the dataset, and each one has a value. 1025 patient records across a range of ages are included in a subset of this dataset; 713 of the records are male and 312 are female. 25% of the test data and 75% of the training data are used by each classifier. Standardized datasets are used to assess classification. Several techniques including ensemble learning, KNN, SVM, AdaBoost, LDA, and K-fold cross-validation are applied for the classification. To improve accuracy, [12] developed a machine-learning method for CVD. The Python programming language and packages Matplotlib, Numpy, and Keras are used by the authors. The Cleveland dataset is the source of the dataset used for their investigation. The data are pre-processed using a variety of approved methods to increase the detection accuracy of the employed ML models. With the Cleveland and CHSLB datasets, the KNN model performs better than the other models. To identify potential risk variables linked to a patient's heart disease state, [13] presents a predictive analysis of patient data. Analysis is conducted using two separate publicly available datasets related to cardiac disease: the Cleveland data, which are utilized to construct classification models, and the Statlog data, which are used to validate the results. Following the training of ten standard classification models for class prediction, a thorough exploratory examination of the Cleveland data is carried out utilizing the Chi-square test of independence. The Cleveland data are randomly divided into 208 (70%) training samples and 89 (30%) test samples to create the classification model. In [14], a CVD detection method that performs feature extraction using a Convolutional Neural Network (CNN) is presented. The aim is to produce a solution to provide sufficient features for machine learning models to achieve good performance, because the feature sets used in previous studies are thought to be small. An ensemble model is designed comprising CNN, stochastic gradient descent classifier, logistic regression, and support vector machine. Then, the experiments are conducted to analyze the performance using different proportions of feature sets. Performance analysis is performed using four different data sets.

In this work, using the SVM technique on two feature data sets, a machine learning model is developed to determine the risk of CVD. The study's feature data with the greatest correct classification are to find by comparing the accuracy of classification of each feature set.

II. Dataset and Method

A. Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique that can be used to orthogonally transform a set of observations of possibly linked variables into a set of values of linearly uncorrelated variables. PCA is a technology that reduces multidimensional data to lower dimensions while maintaining most of the information. It consists of standard deviation, covariance, and eigenvectors (Fig. 1) [15]. In the figure, a K-dimensional space, a swarm with N points, represents a data matrix X. A one-component PC model is depicted in the picture as a 3-space with a straight line fitted to the points. An object's orthogonal projection on the PC line is its PC score (t_i). The loading vector p_k's direction coefficients are for the line [16].

PCA steps are as follows.

Step1. The definition of the data set mean (μ), which is the feature data set ($X_1, X_2, \dots X_n$) is as follows.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

Step2. The mean value is subtracted from each value of the data set.

$$\tilde{X} = X - \mu \tag{2}$$

Step3. To do a principal components analysis, the covariance matrix C must be obtained.

$$C = \tilde{X}\tilde{X}^T \tag{3}$$

Step4. The covariance matrix is used to find the eigenvalues and eigenvectors.

$$Cv = \lambda v \tag{4}$$

where λ stands for eigenvalues and v for eigenvectors.

Step5. The first p number of eigenvectors, w, yields the principal components.

$$y = w^T X \tag{5}$$

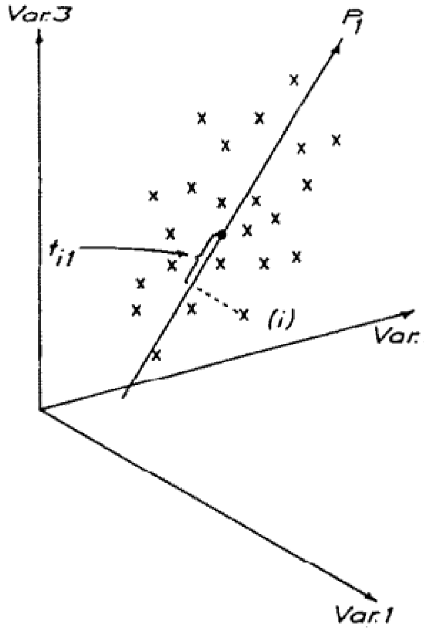


Fig. 1 Illustration of a one-component PC model [16]

B. Support Vector Machine

One of the supervised learning techniques for regression

Open Science Index, Health and Medical Engineering Vol:18, No:3, 2024 publications.waset.org/10013540.pdf

and classification is the support vector machine (SVM) [17]. In order to separate the input vector, SVM seeks to construct a separating hyperplane with the highest margin made up of parallel hyperplanes on both sides. The hyperplane that maximizes the distance between two parallel hyperplanes is known as the separating hyperplane. It is expected that the classifier's discriminative feature will be better the greater the margin or distance between these parallel hyperplanes.

To separate the training data with two classes, the hyperplane is expressed as fallow.

$$w.X + b = 0 \tag{6}$$

Here b is a scalar and w is a p-dimensional (p represents the number of features) vector perpendicular to the separating hyperplane where $y = \{1, -1\}$ is the label data. Parallel hyperplanes for the maximum margin requirement are defined by:

$$w.X + b = 1 \tag{7}$$

$$w.X + b = -1 \tag{8}$$

This can be written as

$$y_i(w.X_i + b) \geq 1 \tag{9}$$

If the training data can be separated linearly, the separating plane and classifier to be created are called linear SVM. Instances where the data set samples touch the hyperplane are called support vectors. The hyperplane with the intended maximum margin is defined by $M = 2 / |w|$ and is solved by minimizing $|w|$. SVMs are visually expressed in Fig. 2.

C. Entropy

The scientific concept of entropy is most associated with a state of chaos, randomness, or uncertainty. The idea and term are used in many different contexts, such as information theory, the microscopic description of nature in statistical physics, and classical thermodynamics, the field in which it is first used [19]. Here is an example of how to define entropy mathematically:

$$S = -\sum_i^n P_i log_2 P_i \tag{10}$$

where n is the class number and P_i is the probability of choosing an example at random from class i.

D. Variance

In probability theory and statistics, variance is the expected value of a random variable's squared deviation from its mean. The standard deviation is the variance squared. Variance, as a measure of dispersion, indicates how far a set of data deviates from the mean [20]. The mathematical definition of variance is:

$$\sigma^2 = \frac{\Sigma(X_i - \mu)^2}{n} \tag{11}$$

E. Dataset

The dataset used in this work is the Cleveland CVD dataset, which is housed in the UCI machine learning repository [21]. The collection consists of 303 total records and 14 characteristics, which are 13 independent variables and 1 dependent variable (also known as an output variable, target variable or label values). The results of the invasive coronary angiography, which indicate whether or not the patient has coronary artery disease, make up the output variable. Labels 0 and $1 - 4$ indicate the existence and absence of cardiac disease, respectively. Most of the studies [22], [23] that have used this dataset have concentrated on only trying to distinguish between presence (values 1, 2, 3, 4) and absence (value 0). Therefore, a classification process based on the presence and absence of CVD was preferred in this study. Table I provides an explanation of the definition and types of characteristics for Cleveland CVD dataset.

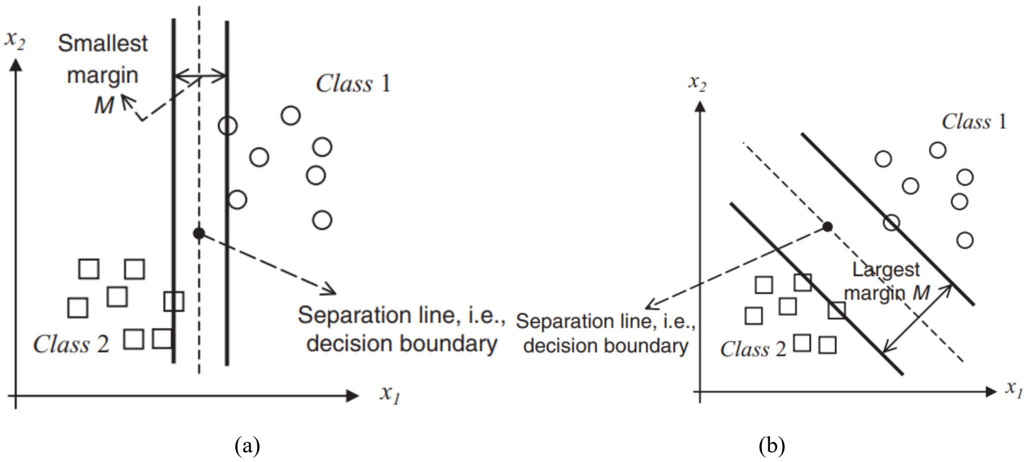


Fig. 2 A decent separating line with a large margin (b) and a less acceptable separating line with a tiny margin (a) are two of numerous separating lines [18]

Open Science Index, Health and Medical Engineering Vol:18, No:3, 2024 publications.waset.org/10013540.pdf

TABLE I
DESCRIPTION OF CLEVELAND DATASET FEATURES

Symbol	Quantity	Conversion from Gaussian and CGS EMU to SI [a]
1	AGE	Age
2	SEX	Sex
3	CPT	Chest Plain Type
4	RBP	Resting Blood Sugar
5	SCH	Serum Cholesterol
6	FBS	Fasting Blood Sugar
7	RES	Resting Electrocardiographic Results
8	MHR	Maximum Heart Rate Achieved
9	EIA	Exercise Induced Angina
10	OPK	ST depression induced by exercise relative to rest
11	PES	Peak Exercise Slope
12	VCA	Number of Major Vessels Colored by Fluoroscopy
13	THA	Thallium Scan
14	Target	Class label

F. Method

The process consists of several phases for feature extraction, classification, and normalization. The dataset is first subjected to the normalizing procedure. The use of min-max normalization is demonstrated in (12). The feature data set is then created using the PCA components and the values of entropy (Enp) and log variance (Lvar). Two stages are involved in classifying the data set: first, the classifier is fed with the normalized raw data, and then the classification is performed again using the feature data created by applying Lvar, Ent, and PCA to the normalized raw data. The linear SVM method is used for the classification process. The six gaps in the dataset are assigned a value of zero during the preprocessing stage. The flow chart for the procedure is shown in Fig. 3. Metrics for accuracy (10), precision (11), recall (12), f1-score (13), and specificity (14) are used to assess the performance of the classification:

$$x_{norm} = \frac{x - x_{min}}{x_{max} + x_{min}} \qquad (12)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (13)$$

$$Precision = \frac{TP}{TP+FP} \qquad (14)$$

$$Recall\ (Sensitivity) = \frac{TP}{TP+FN} \qquad (15)$$

$$F1 = 2x\,\frac{PrecisionxRecall}{PrecisionxRecall} \qquad (16)$$

$$Specificity = \frac{TN}{TN+FP} \qquad (17)$$

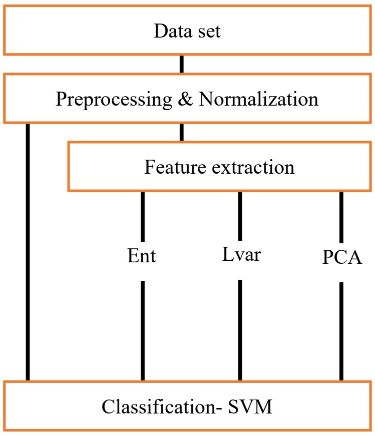


Fig. 3 The method's flow chart

III. RESULT

To carry out the two distinctive processes that form the basis of the classification, the data set is divided into 70% training data and 30% test data. The performance results obtained with the SVM classifier are shown in Table II. The table demonstrates that when the Lvar, Ent, and PCA feature datasets are used, the classification results get worse. The feature set makes use of the first five PCA components. Fig. 4 displays the performance values in a graphical format. Fig. 5 displays the study's ROC curves, which provide an analytical graphical assessment of the effectiveness of the categorization technique.

TABLE II
CLASSIFICATION SUCCESS RATES

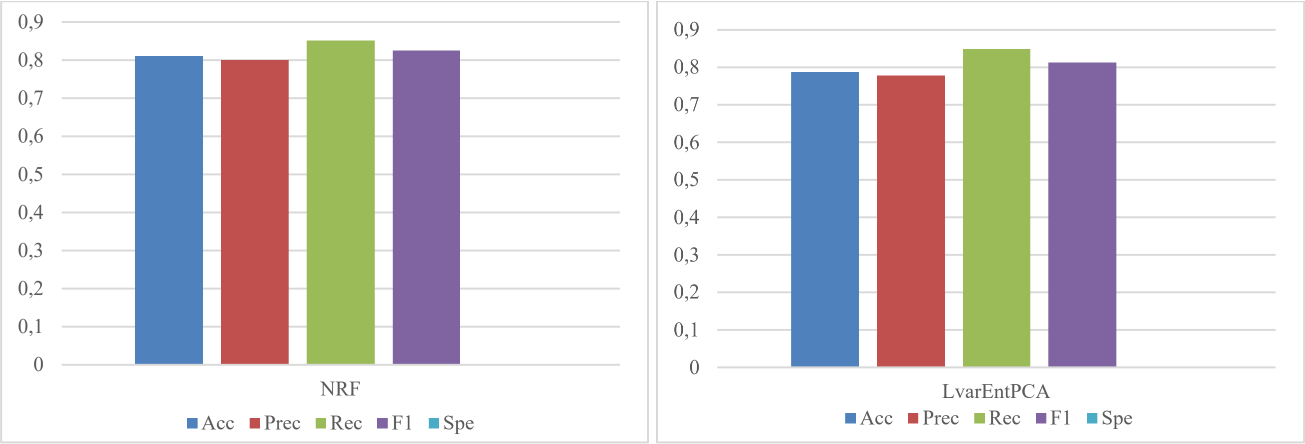
Features	Acc	Prec	Rec	F1	Spe
NRF	0,811	0,80	0,8511	0,8247	0.7674
LvarEntPCA	0,7869	0,7778	0,8485	0,8116	0.7142

Acc: Accuracy, Prec: Precision, Rec: Recall, F1: f1 score, Spe: Specificity, NRF: Normalized raw features.

IV.CONCLUSION

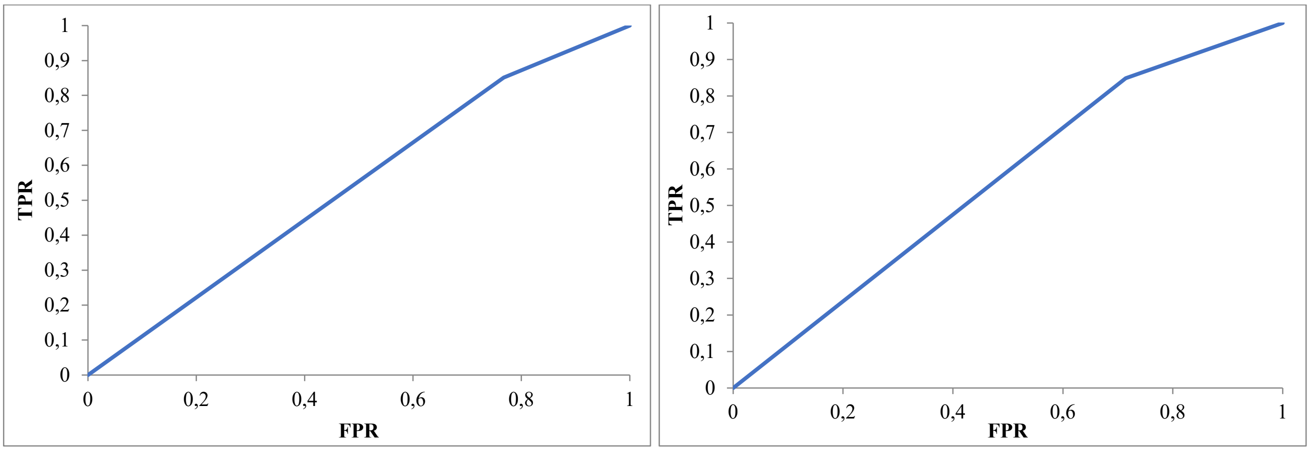
A major cause of death globally, according to the world health organization, is heart disease. Despite the fact that doctors are typically the ones who make medical diagnoses because of their training and expertise, computer-aided decision support systems are extremely important in the medical industry. As a result, forecasting systems that offer readers data in several categories must be developed. In this study it is reported a classification system for identifying CVD risk based on two feature sets. While SVM shows good performance using raw feature set, it is worsened with the feature set formed by feature extraction using PCA, entropy and log-variance. Future studies aim to improve performance by conducting research on the classification model and feature extraction.

Open Science Index, Health and Medical Engineering Vol:18, No:3, 2024 publications.waset.org/10013540.pdf



(a) (b)

Fig. 4 Classification success rates for normalized raw features (a) and extracted features using Lvar, Ent and PCA (b)



(a) (b)

Fig. 5 the classification results' ROC curves for both normalized raw features (a) and LvarEntPCA (b); TPR: True Positive Rate (Sensitivity) FPR: False Positive Rate (1-Specificity)

REFERENCES

[1] Heart Disease Symptoms and Causes - Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118.

[2] D.E. Jonas, S. Reddy, J.C. Middleton, et al. Screening for Cardiovascular Disease Risk with Electrocardiography: An Evidence Review for the U.S. Preventive Services Task Force (Internet). Rockville (MD): Agency for Healthcare Research and Quality (US); 2018 Jun.

[3] WHO World Health Organization, Cardiovascular Diseases. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[4] K. Subhadra, B. Vikas. "Neural network based intelligent system for predicting heart disease", International Journal of Innovative Technology and Exploring Engineering, vol. 8(5), 2019, pp. 484–487.

[5] A. Jain, S. Tiwari, V. Sapra, "Two-phase heart disease diagnosis system using deep learning", Int J Control Autom, vol. 12(5), 2019, pp. 558–573.

[6] S. K. Jain and B. Bhaumik, ''An ultra low power ECG signal processor design for cardiovascular disease detection,'' in Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Aug. 2015, pp. 857–860.

[7] C. Martin-Isla, V. M. Campello, C. Izquierdo, Z. Raisi-Estabragh, B. Baeßler, S. E. Petersen, and K. Lekadir, ''Image-based cardiac diagnosis with machine learning: A review,'' Frontiers Cardiovascular Med., vol. 7, 2020, pp. 1.

[8] P. Rani, R. Kumar, N. M. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," Journal of Reliable Intelligent Environments, vol. 7, 2021.

[9] P. Motarwar, A. Duraphe, G. Suganya, and M. Premalatha, "Cognitive approach for heart disease prediction using machine learning," in Proceedings of the 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), IEEE, India, February 2020, pp. 1–5.

[10] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, "A method for improving prediction of human heart disease using machine learning algorithms," Mob. Inf.Syst., 2022, pp. 1–9.

[11] G. N. Ahmad et al., "Mixed machine learning approach for efficient prediction of human heart disease by identifying the numerical and categorical features," Appl. Sci. (Basel), vol. 12, no. 15, 2022, p. 7449.

[12] N. Absar et al., "The efficacy of machine-learning-supported smart system for heart disease prediction," Healthcare (Basel), vol. 10, no. 6, 2022, p. 1137.

[13] Ogundepo, Ezekiel Adebayo, and Waheed Babatunde Yahya. "Performance analysis of supervised classification models on heart disease prediction." *Innovations in Systems and Software Engineering,* vol. 19.1, 2023, pp. 129-144.

[14] Rustam, Furqan, et al. "Incorporating CNN Features for Optimizing Performance of Ensemble Classifier for Cardiovascular Disease Prediction." *Diagnostics*, vol. 12.6, 2022, pp. 1474.

[15] Karamizadeh, Sasan, et al. "An overview of principal component analysis." Journal of Signal and Information Processing, vol. 4.3B, 2013, p. 173.

[16] S. Wold, K. Esbensen, P. Geladi, "Principal component analysis", *Chemometrics and Intelligent Laboratory Systems*, vol.2, 1987, pp.37-

Open Science Index, Health and Medical Engineering Vol:18, No:3, 2024 publications.waset.org/10013540.pdf

52.

[17] D. K. Srivastava, and B. Lekha. "Data classification using support vector machine." *Journal of theoretical and applied information technology*, vol. 12.1, 2010, pp. 1-7.

[18] Kecman, Vojislav. "Support vector machines–an introduction." *Support vector machines: theory and applications*. Berlin, Heidelberg: Springer, 2005, pp. 1-47.

[19] Rényi, Alfréd. "On measures of entropy and information." Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1: Contributions to the Theory of Statistics. Vol. 4. University of California Press, 1961.

[20] John, Peter WM. "The analysis of variance." Modern Statistics, Methods and Applications, vol. 23, 1980, p. 19.

[21] CVD dataset https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.

Nebi Gedik received his B.S. degree in Electrical and Electronics Engineering from Fırat University in 2001, his PhD degrees in Electrical and Electronics Engineering from Karadeniz Technical University in 2013, and his MSc degree in 2005 from Atatürk University. He is now an Associate Professor at the University of Health Science. His research interests include medical image and signal processing, pattern recognition and machine learning.