World Academy of Science, Engineering and Technology
International Journal of Biomedical and Biological Engineering
Vol:18, No:4, 2024

Prediction of Cardiovascular Disease by Applying Feature Extraction

Nebi Gedik

***Abstract*—**Heart disease threatens the lives of a great number of people every year around the world. Heart issues lead to many of all deaths; therefore, early diagnosis and treatment are critical. The diagnosis of heart disease is complicated due to several factors affecting health such as high blood pressure, raised cholesterol, an irregular pulse rhythm, and more. Artificial intelligence has the potential to assist in the early detection and treatment of diseases. Improving heart failure prediction is one of the primary goals of research on heart disease risk assessment. This study aims to determine the features that provide the most successful classification prediction in detecting cardiovascular disease. The performances of each feature are compared using the K-Nearest Neighbor machine learning method. The feature that gives the most successful performance has been identified.

***Keywords*—**Cardiovascular disease, feature extraction, supervised learning, k-NN.

I. Introduction

THE heart is one of the most important organs of the body and pumps blood to every part of the body through arteries, veins, and capillaries [1]. According to World Health Organization (WHO) data, the majority of people who lost their lives are caused by sudden diseases such as heart attacks or strokes. The heart's pumping systems are impacted by heart disease and become dysfunctional [2], [3]. The most typical sign of cardiovascular illness is pain in the arms and chest. Other symptoms include physical weakness, edema in the feet, exhaustion, and shortness of breath [4]. The use of Machine Learning (ML) to determine the likelihood of a disease occurring is rapidly increasing and is being used to examine Cardiovascular Disease (CVD) data sets for both diagnostic and operational purposes [5], [6]. Based on the patient's clinical characteristics, Rani et al. [7] suggested a hybrid decision support system that can help with the early diagnosis of heart disease. To deal with the missing values, the multivariate imputation by chained equations approach is employed. 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 have also been employed for data pre-processing. Support Vector Machines, Naive Bayes, logistic regression, random forests, and AdaBoost classifiers are employed for the classification task in the final stage of developing the suggested hybrid system. It is concluded that the method produces the most precise outcomes. The Particle Swarm Optimization (PSO) technique is proposed in [8] as a potential means of improving a NN's accuracy even more. A dataset comprising 303 cases of both healthy and unwell individuals serves as the basis for the study. Out of the 72 features in the dataset, only 13 are used in the PSO feature selection procedure. The dataset is preprocessed, and then the PSO is used for feature selection and ranking. Ranking results show that eight of the thirteen alternatives are the most beneficial for increasing NN training accuracy using feed-forward-back propagation. Additionally, it is demonstrated that an improved version of the DNN employing PCA and the grey wolf optimization (GWO) algorithm might be able to detect diabetic retinopathy. The hybrid random forest with linear model (HRFLM) approach is presented in [9]. The suggested hybrid HRFLM approach combines the benefits of the linear method (LM) and random forest (RF) techniques. Improving the accuracy of heart disease prediction is the primary goal of the research. A technique for developing a more useful and accurate risk prediction system to offer supplemental medical services is presented in [10]. The system is composed of four components: feature selection, data preparation, data interface, and classification. This allows it to be practically utilized for patient data collection and analysis in the healthcare industry. Data preprocessing steps, including data integration, data cleaning, and rating mapping, are included in the data interface response used to acquire raw data from hospitals. After the dataset is created, essential features are obtained by reducing dimensionality using the best-first-search approach. To establish the baseline classifier for assessing CVD risk, a RF is implemented.

The authors [11] use the Cleveland heart disease dataset, which contained 303 records and 6 samples with missing values, to identify and forecast human heart disease. Out of the 76 features that are initially present in the data, only 13 are probably going to be cited in any published studies. The remaining feature describes the condition's influence. The Z-Alizadeh Sani dataset is another generic dataset that researchers utilize for the prediction process. It comprises the data of 303 patients with 55 input components and a class label variable for each patient. For the final dataset, nine machine-learning classifiers were employed, both before and after hyper-parameter adjustment. In order to ensure precision on the standard dataset for cardiac disease, pre-processing and standardizing the dataset and tune hyper-parameters are carried out. To train and validate the machine-learning algorithms, the author also develops the K-fold cross-validation technique. Ultimately, the results of the experiment

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

demonstrate that modifying the hyper-parameters improve the prediction classifiers' accuracy while data standardization and hyper-parameter tuning of the ML classifiers produce noteworthy results. To evaluate the algorithms' performance, a variety of metrics is used, such as F-measure, specificity, sensitivity, and classification accuracy. The author [12] offers a model for heart disease prediction. The specific goals are to identify new patients quickly, expedite diagnoses, reduce the incidence of heart attacks, and save lives. There are two databases dedicated to heart diseases: the National Cardiovascular Disease Surveillance (NCDS) system and the Cleveland database, which also includes information on heart disorders. The Cleveland heart disease dataset for this study comprises four aggregated databases: Hungary, Switzerland, and VA Long Beach. The dataset has 14 attributes, each of which has a value. A subset of this dataset includes 1025 patient records representing a range of ages, with 713 male and 312 female records. Each classifier uses 75% of the training data and 25% of the test data. Classifier accuracy is also tested both before and after using standardized datasets. Most of the algorithms described are not based on neural networks and have superior accuracy. Examples of these include k-fold cross-validation, LR, KNN, SVM, Nu SVC, DT, RFC, AdaBoost, GBC, NB, LDA, Q DA, NN, and ensemble approaches. The author [13] developed a machine-learning assistance system in this work that can improve accuracy. The author employs the Python programming language together with libraries like Matplotlib, Numpy, and Keras. The dataset used in the study is collected from the Cleveland dataset. The Cleveland and Cleveland, Hungary, Switzerland, and Long Beach (CHSLB) datasets are utilized to predict coronary heart disease using four ML models. In order to improve the detection accuracy of the used ML models, the data is pre-processed using a range of authorized techniques. Using the Cleveland and CHSLB datasets, the KNN model performs better than the other models.

In this study, to identify the risk of CVD, a ML model is created using relief feature selection method and the k-NN classification algorithm. Different feature sets are composed using relief method and classifier feed them to performance evaluation. By comparing the accuracy of each feature's categorization, the study's feature data with the highest correct classification are identified.

II. Dataset and Method

A. Relief Method

Kira and Rendell [14], [15] initially present the Relief algorithm as an easy, quick, and efficient method for attribute weighing inspired by instance-based learning. A weight between -1 and 1 is the Relief algorithm's output for each feature; higher weights correspond to more selective features. The closest neighbor example from the same class (nearest hit) and the closest neighbor example from the opposite class (nearest miss) are found by selecting an example from the data. When a class change coincides with an attribute value change, the attribute is weighted according to the hunch that the attribute change might be the cause of the class change. Conversely, if an attribute's value changes but the class remains same, the attribute's weight will also fall because the attribute change has no bearing on the class. This process is carried out either for every sample in the data or for a random subset of the samples to update the weight of the attribute. After that, the weight updates are averaged to get a final weight that falls within [−1, 1]. Relief's estimation of the attribute weight has a probabilistic meaning. It is proportionate to the difference between two conditional probabilities, or the likelihood that the attribute value will vary based on the nearest hit and miss, respectively, that are provided [18].

B. k-Nearest Neighbors

When there is little to no prior knowledge about the distribution of the data, one of the most fundamental and straightforward classification techniques - K-nearest neighbor, or kNN - should be the primary option for classification work [19].

By using the observation values in a sample set with certain classes, it is determined which class a new observation to be included in the sample belongs to. The algorithm calculates nearest neighbors for the new observation (unknown) that needs to be classified. The unknown is then allocated to the greatest number of the class by examining the classes to which these neighbors belong. The "Euclidean Distance" is the standard distance measurement technique.

There are two key phases to the approach.

Step1. Determine Euclidean distance (by default) between each piece of data in the dataset.

If the data set is an nxp dimensional (n total number of observations, p total number of features) matrix, the Euclidean distance between sample x_i and x_j $(j = 1,2, \dots, n)$ is defined as follows:

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2} \quad (1)$$

Step2. Selecting the k parameter, which controls the number of neighbors for the algorithm, is carried out. An appropriate choice of k has a major effect on how well the k-NN algorithm performs as a diagnostic. Although a large k involves the danger of ignoring a minor but significant pattern, it lessens the effect of variance caused by random error. To choose the right k value, one must balance the risks of overfitting and underfitting. Fig. 1 shows the representation of the stages of the algorithm.

C. Entropy

A state of chaos, randomness, or uncertainty is most frequently connected to the scientific notion of entropy. The notion and word are applied in a variety of contexts, including the concepts of information theory, statistical physics' microscopic description of nature, and classical thermodynamics, the subject in which it is initially identified [20]. A sample mathematical definition of entropy is as follows:

$$S = -\sum_i^n P_i log_2 P_i \quad (7)$$

where P_i is probability of randomly selecting an example in class 1, n is class number.

D. Variance

Variance is the expected value of a random variable's squared deviation from its mean in probability theory and statistics. The variance's square root yields the standard deviation. As a measure of dispersion, variance expresses the degree to which a group of numbers deviates from its mean [22]. The mathematical definition is:

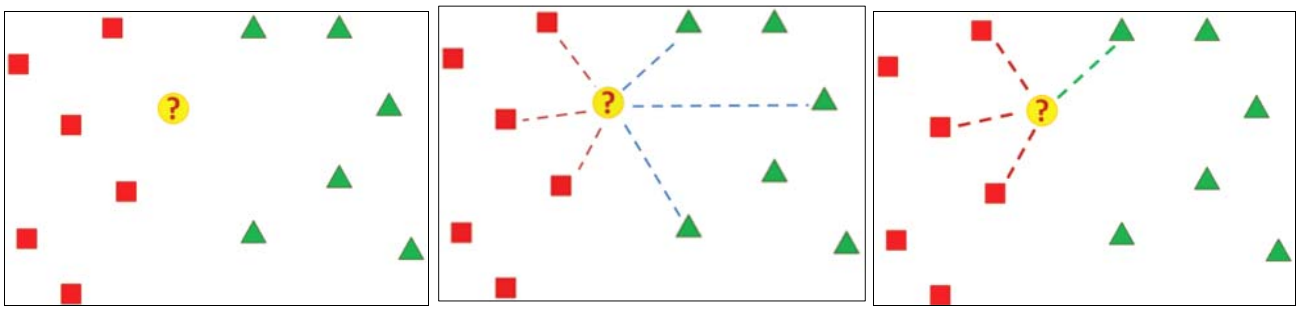
$$\sigma^2 = \frac{\sum(X_i - \mu)^2}{n} \quad (8)$$

E. Dataset

The Cleveland CVD dataset [21], which is part of the UCI ML repository, is the dataset used in this work. There are 303 records in the collection, and 14 attribute values are present. 13 independent variables and one dependent variable (sometimes referred to as an output variable, goal variable, or label values) make up the 14 features. The output variable is the invasive coronary angiography results, which show whether the patient has coronary artery disease. In the data set, labels 0 and 1-4 are used to indicate those with and without heart disease. Most of the research using this dataset has focused solely on attempting to differentiate between presence (values 1, 2, 3, 4) and absence (value 0) of the disease. As a result, in this investigation, a categorization scheme predicated on the existence or absence of CVD is favored. The definition and categories of qualities are explained in Table I.

TABLE I
DESCRIPTION OF CLEVELAND DATASET FEATURES

Symbol	Quantity	Description
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



(a) (b) (c)

Fig. 1 k-NN process steps: (a) two-class data and the data to be categorized; (b) determining the distance between the samples; (c) selecting the k closest values for the decision; k = 4 for this example [21]

F. Method

The method includes normalization, feature selection, and classification stages. The six gaps in the dataset are given a value of zero before normalization process is applied to the dataset. Equation (9) illustrates the application of min-max normalization. Then, feature datasets are created by selecting the 5 most effective features among 13 features using the relief method. MATLAB function is used for the relief method and four feature selection processes are performed by assigning values of 3, 5, 7 and 10 for the k variable. Therefore, four feature sets including five first five effective features according to the relief with k value are created and classified with k-NN. Fig. 2 depicts the method's flow chart. Classification performance is evaluated with accuracy (10), precision (11), recall (12), and f1-score (13) metrics.

$$x_{norm} = \frac{x - x_{min}}{x_{max} + x_{min}} \quad (9)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

$$Recall\ (Sensitivity) = \frac{TP}{TP+FN} \quad (12)$$

$$F1 = 2x \frac{PrecisionxRecall}{PrecisionxRecall} \quad (13)$$

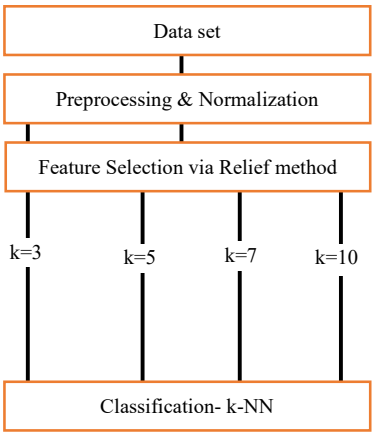


Fig. 2 The method's flow chart

III. Result

The data set is split into 70% training data and 30% test data in order to carry out the two different ways that the classification procedure is planned and implemented. Using the training set of data, the ideal k value is determined as 100. Table II displays the performance results that are acquired using the k-NN classifier. The table shows that as the k value increases, an improvement in the classification result is observed. The graphical representation of the performance values for each feature set composed of changing k values is shown in Fig. 3. The comparative display of the results is shown in Fig. 4.

TABLE II
Classification Success Rates

k for Relief	Acc	k for k-NN	Prec	Rec	F1
3	0.78	7	0.8	0.8	0.8
5	0.78	3	0.8	0.8	0.8
7	0.8	3	0.83	0.8	0.81
10	0.82	65	0.82	0.85	0.83

Acc: Accuracy, Prec: Precision, Rec: Recall, F1: f1 score

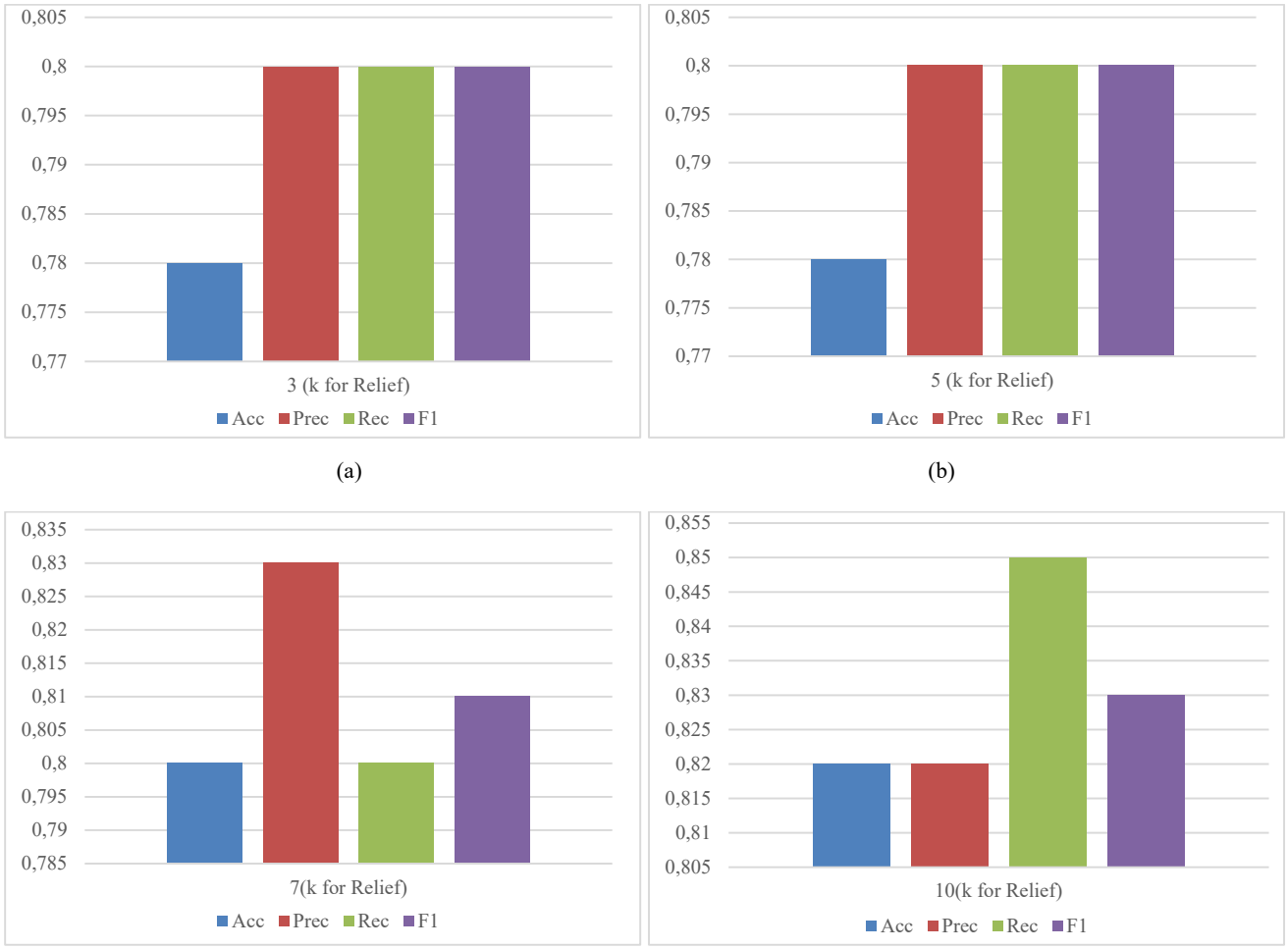


Fig. 3 Classification success rates according to the changing k value for the relief method

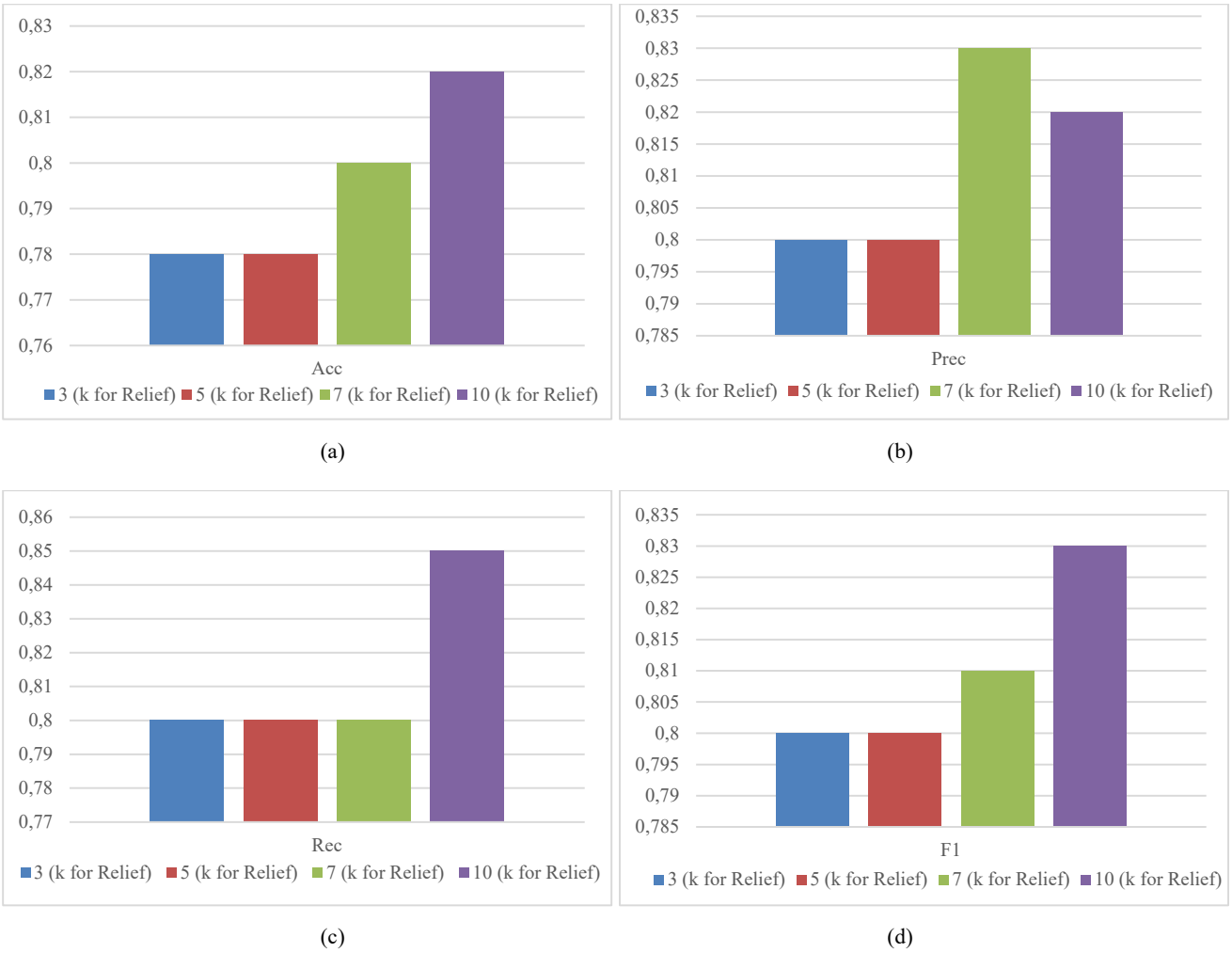


Fig. 4 The comparative display of the results for accuracy (a), precision (b), recall (c) and f1 score (d)

IV. Conclusion

According to the WHO, heart disease is one of the leading causes of death worldwide [22]. 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 field of medicine. Therefore, it is necessary to develop forecasting systems that provide readers with information in different categories. A classification algorithm for detecting the risk of cardiovascular disease based on four feature sets is reported in this study. The k-NN classifier adopting the relief feature selection strategy yields the best accuracy value of 0.82.

References

[1] National Institutes of Health, Practical Guide: Identification, *Evaluation and Treatment of Overweight and Obesity in Adults*, National Institutes of Health0, New York, NY, U.S.A, 2000.

[2] Who Health Organization (WHO), cardiovascular diseases. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[3] D. Deng, P. Jiao, X. Ye, and L. Xia, "An image-based model of the whole human heart with detailed anatomical structure and fiber orientation," *Computational and Mathematical Methods in Medicine*, vol. 2012, 2012, p.16.

[4] A. Ishaq, S. Sadiq, M. Umer et al., "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques," *IEEE Access*, vol. 9, 2021, pp. 39707–39716.

[5] C. J. Harrison and C. J. S. Gibbons, "Machine learning in medicine: a practical introduction to natural language processing," *BMC Medical Research Methodology*, vol. 21(1), 2021, p. 158.

[6] İ. Ozcan, B. Tasar, A. B. Tatar, and O. Yakut, "Destek Vektör Makinesi Algoritması ile Kalp Hastalıklarının Tahmini", *Computer Science*, 4(2), 2019, pp. 74- 79.

[7] 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.

[8] M. G. Feshki and O. S. Shijani, "Improving the heart disease diagnosis by evolutionary algorithm of PSO and feed forward neural network," *in Proceedings of the 2016 Artificial Intelligence and Robotics (IRANOPEN)*, IEEE, Qazvin, Iran, April 2016, pp. 48–53.

[9] S. Mohan, C. 'irumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, 2019, pp. 81542–81554.

[10] Xu, Shan, et al. "Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework." *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*. IEEE, 2017.

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

[12] 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.

[13] 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.
[14] K. Kira, L.A. Rendell, "The feature selection problem: traditional methods and a new algorithm", *in: AAAI*, vol. 2, 1992a, pp. 129–134.
[15] K. Kira, L.A. Rendell, "A practical approach to feature selection", *in: Proceedings of the Ninth International Workshop on Machine Learning*, 1992b, pp. 249–256.
[16] J.P. Callan, T. Fawcett, E.L. Rissland, "Cabot: an adaptive approach to case-based search", *in: IJCAI*, vol. 12, 1991, pp. 803–808.
[17] D.W. Aha, D. Kibler, M.K. Albert, "Instance-based learning algorithms", *Mach. Learn*. Vol. 6 (1), 1991, pp. 37–66.
[18] S.F. Rosario, and K. Thangadurai. "RELIEF: feature selection approach." *International journal of innovative research and development*, vol. 4(11), 2015, pp. 2018-224.
[19] L.E. Peterson. "K-nearest neighbor." *Scholarpedia,* vol. 4.2, 2009. p. 1883.
[20] 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.
[21] Mucherino, Antonio, et al. "K-nearest neighbor classification." Data mining in agriculture (2009): 83-106.
[22] John, Peter WM. "The analysis of variance." *Modern Statistics, Methods and Applications*, vol. 23, 1980, p. 19.
[23] 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.