Advances in Artificial Intelligence Research (AAIR)
ISSN 2757-7422
www.dergipark.com/aair/

Vol. 2 (No. 2), pp. 59-64, 2022
doi: 10.54569/aair.1141465
Published online: Sept 15, 2022

Hybrid Artificial Intelligence-Based Algorithm Design For Cardiovascular Disease Detection

Buse Nur Karaman [1,*], Zeynep Bağdatlı [1], Nilay Taçyildiz [1], Sude Çiğnitaş [1], Derya Kandaz [1], Muhammed Kürşad Uçar [1]

[1] Sakarya University, Faculty of Engineering, Department of Electrical and Electronics Engineering. Sakarya, Turkey

Abstract

Objective: Cardiovascular Disease (CVD) is a disease that negatively affects the blood vessel system due to plaque formation as a result of accumulation on the inner wall of the vessels. In the diagnostic phase, angiography results are evaluated by physicians. New diagnostic algorithms based on artificial intelligence, including new technologies, are needed for diagnosing CVD due to the time-consuming and high cost of diagnostic methods.
Materials and Methods: The heart disease dataset available on the open-source sharing site Kaggle was used in the study. The dataset includes 14 clinical findings. In the study, after the features were selected with the Fischer feature selection algorithm, they were classified with Ensemble Decision Trees (EDT), k-Nearest Neighborhood Algorithm (kNN), and Neural Networks (NN). A hybrid artificial intelligence algorithm was also created using the three methods.
Results: According to the classification results, EDT %96.19, kNN %100, NN %86.17, and hybrid artificial intelligence determined CVD with a %99.3 success rate.
Conclusion: According to the obtained results, it is evaluated that the proposed CVD diagnosis hybrid artificial intelligence algorithms can be used in practice.

Keywords: *Cardiovascular disease, angiography, hybrid, artificial intelligence, neural networks.*

1. Introduction

Heart disease can be generalized as disorders in the structure or functioning of the heart. Heart diseases are one of the leading causes of death worldwide. However, while CVD deaths decrease in high-income countries, more deaths occur in many low- and middle-income countries [1]. Death due to coronary heart disease ranks first among the causes of death in Turkey. According to the follow-up results of the TEKHARF study covering the years 1990-2008, deaths from coronary heart disease in the 45-74 age group are 7.64 per 1000 person-years in men and 3.84 in women, and it is one of the countries with the highest rate in Europe [2]. The leading causes of cardiovascular diseases are physical inactivity, unhealthy diet, hypertension, smoking, and alcohol consumption [3]. Early diagnosis of the disease and appropriate treatment are required to reduce the number of patients who continue due to heart disease each year and minimize the risk of death [4]. Cardiovascular disease detection is a complicated process that can be diagnosed by examining and evaluating the results obtained using various clinical diagnostic methods such as electrocardiography (ECG), echocardiography (heart ultrasound), exercise test, and angiography.

Even though physicians and radiologists mostly make a correct diagnosis, new diagnostic methods with less error rate and more sensitivity are sought and researched for the diagnosis stage of the disease, which is created by today's technology. Artificial intelligence has an important place in today's technology. It is used in almost every field, including health, based on artificial intelligence and data mining and provides considerable convenience to human life [5]. As a basic definition, "Data Mining is the trivial extraction of implicit, previously unknown and potentially useful information about data" [6]. Artificial intelligence-based computer-aided systems can make much faster, more precise, and more accurate detections. The literature shows that early disease diagnosis can be made by using many machine learning, artificial intelligence, and classification methods. Artificial intelligence uses many different algorithms and methods to detect the disease. When a literature review is done, it is seen that the hybrid method algorithm, which results by combining multiple different methods and algorithms, is also used. The hybrid method can be defined as an algorithm type that can be classified separately with more than one algorithm, the value is taken, and results are obtained according to the standard value [7]. Thanks to the hybrid method, it is aimed to increase the accuracy rate obtained from the study.

Three different algorithms were used as hybrid method algorithms during the classification process. These algorithms are Neural Networks (NN), Ensemble Decision Trees (EDT), and k-Nearest Neighborhood (CNN) algorithms. Since the highest accuracy rate was achieved among the different classification algorithms, it was

* Author
E-mail address: buse.karaman@ogr.sakarya.edu.tr

Received: 07/Jul/2022; Received in revised form: 29/Aug/2022; Accepted: 01/Sept/2022.

deemed appropriate to use these three algorithms by aiming to increase the accuracy rate. Today, there is an increase in the use of artificial intelligence-based computer-aided systems in disease diagnosis. [8]. In such cases, it is peculiar to people; Although concepts such as talent, experience, education, daily mood, and distraction are not in question for artificial intelligence, they will significantly benefit physicians at the diagnosis stage. After processing the data set in a MATLAB environment, this study aims to develop a system that will help physicians diagnose CVD disease with an artificial intelligence-based hybrid method algorithm and to use the developed system in practice.

2. Literature Review

Classification systems make it easier for doctors to diagnose diseases. The hybrid model used for this study is one of the most common methods encountered in the literature. Studies in the literature have revealed that the features in the data set can have adverse effects, called noise features, during the diagnosis and diagnosis phase. In a study conducted in 2016, %92.59 classification accuracy was achieved in diagnosing heart disease using the hybrid classification system.

The most important contribution of this study has been the observation that feature selection methods can improve the performance of classification algorithms. Using Least Squares Support Vector Machines and the F-score feature selection method in the study, an accuracy rate of %85.59 was obtained, while an accuracy rate of %83.37 was obtained in the SVM result [9]. Three machine learning algorithms are used in the study of professors M Kavitha, G Gnaneswar, R Dinesh, YR Sai, RS Suraj: Random Tree, Decision Trees, and Hybrid Model. In the study, an accuracy rate of %88.7 was achieved with the hybrid model, and it was shown that the hybrid model was the method that would give the best results in the diagnosis of heart disease [10]. In another study on the diagnosis of heart disease, the focus was on improving the classification methods used to reduce the number of features with feature selection. The k-nearest neighbor algorithm was used [11]. Increasing the accuracy rate by reducing the effects of unnecessary features is prominent in the literature. In a study in which Probabilistic Principal Component Analysis was used for the features to be extracted to increase the classification success, radial basis function-based Support Vector Machines (SVM) were used for classification. Thus, %82.18, %85.82, and %91.30 success rates were obtained from the three data sets used in the study [12].

In the study in which another data set was used in which the hybrid model was recommended in the diagnosis of coronary heart disease, it was observed that an accuracy rate of %86.3 was obtained. Obtaining the AUC parameter of %92.1 in this study revealed that the proposed classification system can be used to diagnose heart disease [13].

3. Material and Methods

Various machine learning methods have been used to diagnose heart disease in the literature. The study suggests that successful results can be obtained with the hybrid model among many classification methods in the literature, such as Random Forest, Logistic Regression, and C4.5 decision tree. Fischer Feature Selection algorithm was used to reduce unnecessary features to increase the success in diagnosis and diagnosis. The feature values calculated with this algorithm are ordered from the largest to the smallest, and it means taking the number of samples that will provide the classification success in the best way. According to the Fischer Algorithm, the features' mean and standard deviation values are calculated, and the ranking is made according to the obtained scores. Neural networks, nearest neighbor algorithm (k-NN), and ensemble decision trees (EDT) used in the study are the most widely used algorithms in machine learning and classification studies. The steps followed in the study are modeled in the flow chart shown in **Figure 1**.

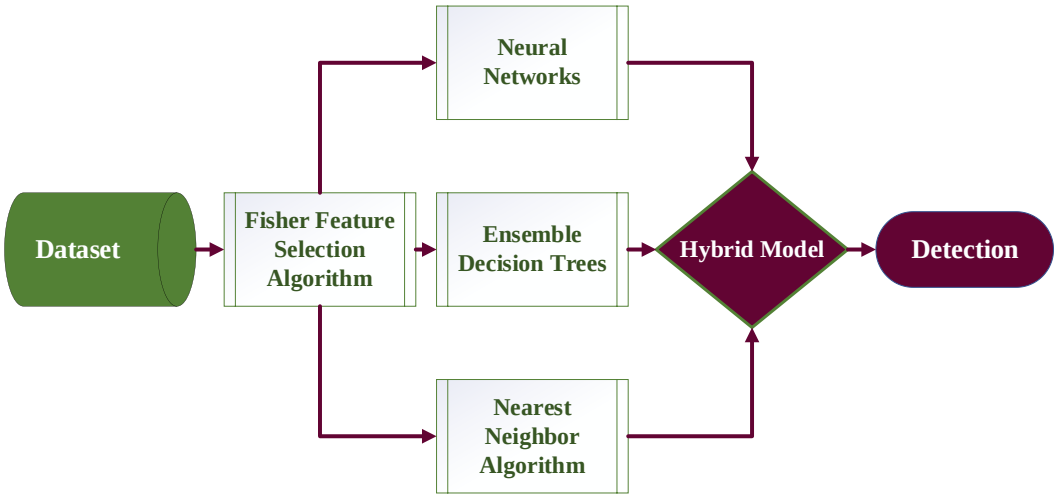

Figure 1. *Flowchart*

3.1. Dataset

In the study carried out for the diagnosis of heart disease, a ready data set consisting of 4-databases from Kaggle belonging to Cleveland, Switzerland, Long Beach V., and Hungary was used. The most important feature to be considered in selecting the data set is whether it has enough extensive data. For machine learning, the information used with data sets with sufficiently large data provides a more successful prediction for future studies. The dataset contains 14 features, age, gender, type of chest pain, resting blood pressure, cholesterol information, fasting blood glucose, resting electrocardiography results, highest heart rate achieved, exercise-induced angina, exercise-induced ST depression, peak exercise ST segment slope, the number of colored veins, the disease and health status of the patients as labels. The characteristics of the data set are shown in **Table 1**.

Table 1. *Representations of Features*

Feature Number	Features	Values	Explanation
1	age	Numerical value	Patient's Age
2	gender	1:Male, 0:Female	Patient's Gender
3	cp	1: Typical Angina, 2: Atypical Angina, 3: Non-anginal Pain, 4: Asymptomatic	Chest Pain Type
4	trestbps	Numerical Value (140mm/Hg)	Resting Blood Pressure (Blood Pressure) (mm/Hg)
5	chol	Numerical Value (289mg/dl)	Serum Cholesterol Amount in the Blood (mg/dl)
6	fbs	1: True, 0: False	Fasting Blood Sugar>120 mg/dl
7	restecg	0: Normal, 1: Has ST-T, 2: Hypertrophy	Resting Electrocardiographic Results
8	th	140,173	Max Heart Rate Reached
9	exangial	1: yes, 0: no	Exercise-Induced Angina
10	old peak	Numerical value	Exercise-Induced ST Depression
11	slope	1: Curved Up, 2: Straight, 3: Curved Down	The slope of the Peak Exercise ST Segment
12	ca	0-3	Number of Vessels Colored by Fluoroscopy (0-3)
13	thal	0: Normal, 1: Fixed Fault 2: Reversible Defect	Thalassemia
14	target	0: <50% Diameter Reduction 1:> 50% Diameter Reduction	Diagnosis of Heart Disease (Angiographic Disease Status)

3.2. Feature selection

Feature selection is a process that directly affects the accuracy and efficiency of classification. It provides the most valuable features for the problem being studied. Selecting the most valuable features also reduces the data size, reducing redundant results of the analysis. In the study, the Fisher algorithm, a feature selection algorithm, was used to increase the success rate of artificial intelligence. Ideally, feature selection methods decompose features into subsets and try to find the best among candidate subsets. This process can be costly and practically impossible, especially for high-volume feature vectors. The Fisher Score criterion assigns an indicative value for each sample. This value should be similar for instances in the same class and distinctive for instances in different classes. The Fisher Score equation providing this expression is given below;

$$F(i) = (\overline{x}\, i\,(+) - \overline{x}i)\, 2 + (\overline{x}\, i\,(-) - \overline{x}i)\, 2\; 1\; n + -1 \sum (xk,i\,(+) - \overline{x}\, i\,(+))\, 2 + 1\; n — 1 \sum (xk,i\,(-) - \overline{x}\, i\,(-) \quad (1)$$

The number of features selected is an attribute at the researcher's discretion. In the feature selection phase, the aim is to rank the features that give the best results among the features from best to worst and to increase the accuracy rate by reducing the number of features. This algorithm involves ordering the features whose values are calculated from the largest to the smallest and taking the desired number of top samples. The best feature ranking obtained as a result of the algorithm was obtained. The table of classification results for the best nine features selected from all features is shown below.

Table 2. *Feature Selection Table*

	Accuracy	Sensitivity	Specificity	F-Measurement	Kappa	AUC
Model 1	97,99	0,9679	0,9920	0,9798	0,9599	0,9799
Model 2	96.09	0.9479	0,97247	0,9600	0,9215	0,9601
Model 3	95.60	0,9791	0,9357	0.9569	0,9121	0.9574
Model 4	95,6	0,9174	1	0,9569	0,9123	0,9587
Model 5	94,63	0,9583	0,9357	0,9469	0,8924	0,9470
Model 6	94,15	0,9791	0,9032	0,9423	0,8830	0,9437
Model 7	93,17	0,9375	0,9266	0,9320	0,8630	0,9320
Model 8	92,68	0,9375	0,9174	0,9273	0,8533	0,9274
Model 9	90,73	0,9270	0,8899	0,9081	0,8146	0,9084

3.3. Neural Networks

Neural networks, which are encountered in many scientific and technological studies, emerged as a result of taking the human brain as an example; It is used in areas such as diagnosis, classification, and control. Artificial neural networks have been created, considering the structure of biological neural networks and the brain's learning process. Biological neural networks mainly consist of the nucleus, dendrite, and axon. The axon transmits the information collected by the dendritic ends. The basic structure of artificial neural networks consists of 3 parts. These sections are named the input layer, hidden layer, and output layer. The input layer has the same function as the dendrites in biological neural networks and is defined as the part where data entry is made. The hidden layer corresponds to the core of biological neural networks. Here, the relationship between the input and output values is learned and processed in line with the algorithm. On the other hand, the output layer acts as an axon and provides the output of the results. After adding the skewness of the input values with the weights, a bias is added and transferred to the output after the activation function. The weight value is the most critical parameter of the learning process of artificial neural networks. The weight value, which takes a random value at the beginning of the training process of artificial neural networks, is updated according to the input data throughout the process.

3.4. Nearest Neighbor Algorithm (k-NN)

The K Nearest Neighbor method is among the supervised learning methods that solve the classification problem. The important thing is that the characteristics of each class are predetermined. The method's performance is affected by the number of k nearest neighbors, threshold value, similarity measurement, and the sufficient number of expected behaviors in the learning set [14]. KNN is based on estimating the class of the vector formed by the independent variables of the predictable value, based on the information in which class the nearest neighbors are dense. The K-NN algorithm is one of the most fundamental algorithms among machine learning algorithms. This algorithm makes predictions on two different components: distance and neighborhood.

The number of distance neighbors is several; how many neighbors closest to that value are used to determine which class the value to be included in a class will be included. The value whose class is the closest to which

class is in this number is included in this class.

3.5. Ensemble Decision Trees (EDT)

The decision tree algorithm is one of the data mining classification algorithms. Decision tree-based methods can quickly calculate performance value criteria such as stability, specificity, and high accuracy. Unlike linear models, it has the advantage of being able to map nonlinear complex models well. The decision tree algorithm, one of the machine learning algorithms, is a classification algorithm that helps us analyze the learned model by dividing it into subsets in a fast, simple, and interpretable way. A decision tree is a structure that decomposes a data set containing a large amount of data into smaller subsets by subjecting it to decision rules. Decision trees stand out among other algorithms with their visual intelligibility. The basic principle of decision trees is the root, node, and leaf-based model. The leaf is the last stage in the model where we reach the desired result.

4. Findings and Discussion

The study aims to analyze the relationship between blood values, chest diseases, chest pain type, minimum heart rate, breathing rate, resting blood pressure, age, cholesterol, fasting blood sugar, maximum heart rate, gender, and exercise with the diagnosis of cardiovascular diseases and cardiovascular disease. In addition, the study aims to diagnose the disease with an artificial intelligence-based hybrid model and to measure its usability in practice. The dataset, which contains 1025 data, includes 526 data with cardiovascular disease and 499 data without cardiovascular disease. The model includes 14 features. The feature selection algorithm is used to improve the performance of the machine learning algorithms. By optimizing the size of the data set with feature selection, the workload is removed, and performance is increased. Given this situation, the Fisher algorithm was used with the hybrid method to improve the study's success. After the processes, the model was evaluated with the hybrid model algorithm. In the study, three different classification algorithms, k-NN, NN, and Ensemble Decision Trees, were included in the Hybrid Model. Training and test percentages were calculated according to these three different classification methods, and a 'Hybrid' result was obtained. Another critical point after the feature selection to increase the classification success is the creation of training and test classes. With these correctly determined class percentages, the algorithm successfully performs the test set with the method it learned from the training set. The ratio of the test dataset was determined as %25 to increase the accuracy rate. In the study, a dataset with a specific label value was trained in the training step by simply using a two-class (with/without heart disease) dataset. Then the testing phase started. Accordingly, the accuracy values of the model were obtained. The success of the diagnosis was evaluated according to sensitivity, specificity, F-measurement, kappa parameter, and AUC values. According to the classification results, EDT has an accuracy rate of %96.192, %kNN 100, NN %86.17, and in line with these values, an accuracy rate of %97.99 was obtained with the hybrid model. The performance evaluation results of the hybrid algorithm are shown in **Table 3**. The performance evaluation results of the other three algorithms within the hybrid algorithm are given in **Table 4**, **Table 5**, and **Table 6**.

As a result, CVD was detected with a high success rate.

Table 3. *Performance evaluation results for the hybrid model*

Accuracy	Sensitivity	Specificity	F-Measurement	Kappa	AUC
97,99	0,9679	0,9920	0,9798	0,9599	0,9799

Table 4. *Performance evaluation results for the NN model*

Accuracy	Sensitivity	Specificity	F-Measurement	Kappa	AUC
86,17	10	0,8840	0,8611	0,7234	0,8617

Table 5. *Performance evaluation results for the k-NN model*

Accuracy	Sensitivity	Specificity	F-Measurement	Kappa	AUC
100	1	1	1	1	1

Table 6. *Performance evaluation results for the EDT model*

Accuracy	Sensitivity	Specificity	F-Measurement	Kappa	AUC
96,1924	0,9598	0,9640	0,9619	0,9238	0,9619

5. Conclusion and Recommendations

Diagnosis of diseases that negatively affect human life, such as cardiovascular disease, is crucial in terms of quality of life and health. It is a promising development that computer-aided systems and artificial intelligence have recently played an active role in the health field and are conducive to positive innovations. This study aims to provide ease of diagnosis to physicians and contribute positively to the literature and human life by accelerating the process. According to the results obtained, the high accuracy rate obtained as a result of the study and the method used in the study contributed to the literature, and it was proven that the proposed methods diagnose CVD at a high rate. Based on this, it is evaluated that hybrid artificial intelligence algorithms can be used in practice.

Thanks

We want to thank the person(s) who transferred the "Heart Disease UCI" dataset to the open-source website (kaggle.com) in the study.

Author(s) Contributions

BK and NT are responsible for the activities carried out in the computer environment. SÇ and ZB did artificial intelligence analysis studies. BK, NT, ZB, and SÇ wrote the article. All four authors read and approved the final version of the article.

Declaration of interest

The authors declare that there is no conflict of interest. It was presented as a summary at the ICAIAME 2022 conference.

References

[1] Onat A, Uğur M, Tuncer M, Ayhan E, Kaya Z, Küçükdurmaz Z, et al. "Age at death in the Turkish Adult Risk Factor Study: temporal trend and regional distribution at 56,700 person-years follow-up", *Türk Kardiyol Dern Arş* 37(2009), 155-60.

[2] Üner, S, Balcılar, M ve Ergüder, T. "Türkiye hanehalkı sağlık araştırması: bulaşıcı olmayan hastalıkların risk faktörleri prevalansı", Ankara: Dünya Sağlık Örgütü, Türkiye Ofisi, 2017.

[3] Liu X., Wang X., Su Qiang. "A hybrid classification system for heart disease diagnosis based on the RFRS method", *Computational and Mathematical Methods in Medicine*, vol. 2017, Article ID 8272091, 11 pages, 2017. https://doi.org/10.1155/2017/8272091.

[4] Bulut F. "Heart attack risk detection using Bagging classifier". 24th Signal Processing and Communication Application Conference (SIU) (pp. 2013-2016).

[5] Priyanka N. and Kumar P. R., "Usage of data mining techniques in predicting the heart diseases — Naïve Bayes & decision tree", 2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT), 2017, pp. 1-7, doi: 10.1109/ICCPCT.2017.8074215.

[6] Taşçı M. E. ve Şamlı R., "Veri Madenciliği İle Kalp Hastalığı Teşhisi", *Avrupa Bilim ve Teknoloji Dergisi*, (2020) 88-95; doi:10.31590/ejosat.araconf12.

[7] Eray, A., Ateş, E., & Set, T. "Yetişkin bireylerde kardiyovasküler hastalık riskinin değerlendirilmesi". *Türkiye aile hekimliği dergisi*, 22 (2018), 12-19.

[8] Atay R., Odabaş D. E., Pehlivanoğlu M.K. (2019). "İki Seviyeli Hibrit Makine Öğrenmesi Yöntemi İle Saldırı Tespiti", *Gazi Mühendislik Bilimleri Dergisi*, 5 (2019), 258-272.

[9] Kavitha M., Gnaneswar G., Dinesh R., Sai Y. R. and Suraj R. S., "Heart Disease Prediction using Hybrid machine Learning Model", *6th International Conference on Inventive Computation Technologies (ICICT)*, 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.

[10] Nourmohammadi-Khiarak, J., Feizi-Derakhshi, MR., Behrouzi, K. et al. New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection. *Health Technol*. 10 (2020), 667–678. https://doi.org/10.1007/s12553-019-00396-3.

[11] Shah, S M, ve diğerleri. Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis. basım yeri bilinmiyor: Physica A: Statistical Mechanics and its Applications, 2017.

[12] Wiharto W., Kusnanto H. & Herianto H. "Hybrid system of tiered multivariate analysis and artificial neural network for coronary heart disease diagnosis", *International Journal of Electrical and Computer Engineering*, 7(2), (2017) http://doi.org/10.11591/ijece.v7i2.pp1023-1031 .

[13] Babur, S., Turhal, U., Akbaş, A., DVM Tabanlı Kalın Bağırsak Kanseri Tanısı için Performans Geliştirme. Elektronik ve Bilgisayar Mühendisliği Sempozyumu (ELECO 2012), Bursa, 2012.

[14] Çalışkan, S. K., & Soğukpınar, İ., "KxKNN: K-means ve k en yakin komşu yöntemleri ile ağlarda nüfuz tespiti", EMO Yayınları, 120-24, 2008.