Sakarya University Journal of Science

ISSN : 2147-835X
Publisher : Sakarya University

Vol. 28, No. 1, 101-107, 2024
DOI: https://doi.org/10.16984/saufenbilder.1312109

Research Article

Heart Disease Prediction with Machine Learning-Based Approaches

Ayhan Küçükmanisa[1*], Zeynep Hilal Kilimci[2]

[1*] Kocaeli University, Faculty of Engineering, Department of Electronics and Communication Engineering, Kocaeli, Türkiye, ayhan.kucukmanisa@kocaeli.edu.tr
[2] Kocaeli University, Faculty of Technology, Department of Information Systems Engineering, Kocaeli, Türkiye, zeynep.kilimci@kocaeli.edu.tr
*Corresponding Author

ARTICLE INFO

Keywords:
Heart Disease Prediction
Machine Learning
Artificial Neural Networks
Gradient Boosting

Article History:
Received: 09.06.2023
Accepted: 12.11.2023
Online Available: 27.02.2024

ABSTRACT

Heart disease, a global ailment with substantial mortality rates, poses a significant health concern. The prevalence of heart disease has escalated due to the demanding nature of contemporary occupations and inherent genetic predispositions. Hence, timely detection of cardiac disorders is paramount to preserving lives. However, the analysis of routine clinical data presents a formidable challenge in identifying cardiovascular ailments. Leveraging machine learning approaches to scrutinize clinical data can furnish effective solutions for informed decision-making and precise prognostications. This research endeavors to predict heart disease by examining the data of 303 individuals encompassing 14 distinct categories. Several machine learning methodologies, namely K-Nearest Neighbor, Gaussian Naive Bayes, Logistic Regression, Random Forest, Gradient Boosting, and Artificial Neural Networks, are proposed as potential remedies to address the problem. The experimental findings unveil that Gradient Boosting attains a remarkable accuracy of 95% and Artificial Neural Networks exhibit a commendable accuracy of 90.1%, establishing them as the most successful models in this study. These results underscore the superior performance of the proposed techniques vis-à-vis the existing literature.

1. Introduction

The early and precise detection of heart diseases plays a vital role in the preservation of patients' lives, like numerous other medical conditions. It is noteworthy that heart disease accounts for approximately 32% of total fatalities, amounting to 17.9 million deaths annually [1]. For this reason, the applications made in the field of health, the diagnoses made, and the results obtained are of great importance for human life.

Today, artificial intelligence applications have affected many areas of life. Artificial intelligence has also made very useful contributions to the field of medicine in evaluating, classifying, and analyzing data. The investigation and development of prediction methods focused on heart diseases represent a crucial area of study within the field of medicine. Deep learning and machine learning algorithms, both categorized under artificial intelligence, have the capability to be employed on carefully curated datasets, incorporating specific information gathered from individuals. This approach allows for the development of highly accurate models, yielding exceptional precision rates. By putting the information from the newly arrived patient into these models, it becomes possible to diagnose many patients with high accuracy in a very short time.

Within the existing body of literature, numerous investigations have been conducted in the field of heart disease prediction. A recent study [2] put forward the utilization of six distinct machine

Cite as: A. Küçükmanisa, Z. H. Kilimci (2024). Heart Disease Prediction with Machine Learning-Based Approaches, *Sakarya University Journal of Science,* 28(1),101-107. https://doi.org/10.16984/saufenbilder.1312109

learning techniques to identify the likelihood of heart disease occurrence, aiming to discern the prevalence of this medical condition. The performance of these methods was evaluated over eight different classification metrics. Among the methods proposed in the study, logistic regression provided the highest performance with 85% accuracy, 89% sensitivity and 81% specificity.

Random Forest, Decision Tree and Hybrid version of them are employed to solve disease prediction problem on Cleveland dataset in [3]. Based on the experimental findings, it is revealed that the hybrid model attains an accuracy rate of 88.7% in the prediction task. [4] introduces the Hybrid Random Forest with Linear Model (HRFLM) approach. Many studies use some feature selection techniques to get better performances. The HRFLM technique, in contrast, employs all features from the Cleveland dataset without any limitations on feature selection. In the HRFLM approach, all 13 clinical characteristics are utilized as input to an ANN with back propagation. The proposed method's accuracy is 87%.

In [5], ensemble learning based methods for heart disease prediction are proposed and tested. A classifier that is derived via randomness analysis of distance sequences is used as the base estimation for a bagging strategy. The medical Spectf dataset has successfully tested the approach. In the case of the UCI dataset, known as Statlog, a classifier based on Graph Lasso and Ledoit-Wolf shrinkage techniques is devised. These two methods yield highly satisfactory accuracy outcomes, with Spectf achieving 88.7% accuracy and Statlog achieving 88.8% accuracy.

A hybrid approach utilizing data mining techniques for heart disease prediction is proposed in the study conducted by [6]. Experimental results demonstrate that higher accuracy is achieved by the hybrid approach compared to individual algorithms. The complex relationships between risk factors and heart disease are effectively captured by combining the outputs of multiple algorithms, resulting in improved prediction performance. By applying Naive Bayes, Support Vector Machines, k-NN, ANN, J4.8, Random Forests, and Genetic Algorithm to the entire dataset, an accuracy rate of 89.2% is achieved, and the feature set is reduced from 14 to 12 without compromising accuracy, as determined through calculations.

In [7], a hybridized approach is introduced to enable early detection of heart disease. The dataset undergoes a feature selection process combining the Genetic Approach (GA) with recursive feature removal, thereby identifying the most relevant features. Pre-processing of the data involves utilizing both SMOTE (Synthetic Minority Oversampling Technique) and traditional scalar methods. The proposed hybrid system incorporates support vector machine, naive Bayes, logistic regression, random forest, and Adaboost classifiers. Experimental results reveal that the random forest classifier achieves the highest accuracy rate of 86.6% within the proposed methodology.

[8] proposes a method based on deep neural networks (DNN). After two convolutional layers, the model has eight dense layers. The 6 layers of the model have 128,128,128,128,64,1 neuron, respectively. The activation function used throughout the network, except for the last layer, is the exponential linear unit (ELU). The proposed DNN model achieved accuracy of 91.7%.

2. Proposed Method

In this work, prediction of heart disease is considered as a classification problem. To solve this classification problem, various machine learning based approaches are proposed.

2.1. Dataset

Heart Disease Dataset [9] is created 1988 with the work of 4 different health institutions, including V.A. Medical Center Long Beach and Cleveland Clinic Foundation, Hungarian Institute of Cardiology Budapest, University Hospital Basel Switzerland, University Hospital Zurich Switzerland. This dataset contains 76 features in 14 different categories. For those who do machine learning studies, especially the Cleveland part has been the only dataset for a long time.

Within the labeled data, the target section contains grading information about the presence of the disease. A scale ranging from 0 to 4 is utilized to denote the absence or presence and severity of the disease. A rating of 0-4 corresponds to the absence of the disease, whereas a rating higher than 4 indicates the presence and increasing severity of the disease. Although the personal information of the patients, such as name and surname, was included in the dataset at first, it was then anonymized. Dataset properties and details are given in Table 1.

Of the 14 features in the dataset, 9 of them are categorical and 5 of them are numerical data. Components of high risk of developing heart disease: weight, high cholesterol, smoking, diabetes, high blood pressure, and family history are stated in the dataset comment sections. The metrics that cannot be changed and that we cannot have an impact on are stated as increasing age, gender, and heredity. Factors that can be found in the data set and changed in a person's life: high blood pressure, smoking, being overweight, high cholesterol, sedentary life, and having diabetes.

2.2. Proposed machine learning methods

In this research investigation, a range of machine learning techniques is employed to address the problem at hand. The utilized methods encompass Artificial Neural Networks (ANN), K-nearest neighbor (kNN), Random Forests, Gaussian Naïve Bayes, Logistic Regression, Gradient Boosting and Support Vector Machine (SVM).

Table 1. Heart disease dataset properties

Feature	Description
Age	Current age of persons
Sex	People's gender
Cp	Chest pain types: Atypical angina, asymptomatic, typical angina, non-anginal pain.
Trestbps	A person's blood pressure is tested at a hospital when they are at rest.
Chol	Measurement of a person's cholesterol

Table 2. Heart disease dataset properties (Continue)

Feature	Description
Fbs	Fasting blood glucose level of a person (if value > 120 mg/dl, 1 means true; 0 means false)
Restecg	Electrocardiographic measurement at rest 0.1,2 normal, wave abnormal and possible hypertrophy, respectively
Thalach	Person's highest heart rate
Exang	Pain formation due to exercise is present or absent, in order of 0 or 1
Oldpeak	ECG position of ST depression at rest
Slope	The slope of the ST segment during peak exercise is categorized into three distinct patterns: upward slope, flat slope, and downward slope, denoted by the values 1, 2, and 3, respectively.
Ca	Main change metric
Thal	Thalassemia is a blood disorder that has three levels: normal (3), fixed (6), and reversible (7).
Target	Heart disease 0 or 1 respectively no, yes

Support Vector Machine (SVM) is a supervised learning approach that can employ various kernel functions depending on the data's characteristics during the algorithm's execution. This flexibility enables it to perform both linear and nonlinear classification tasks effectively. Its main objective is to establish a hyperplane that efficiently separates the data points.

K-nearest neighbor (k-NN) [10] is an example-based learning method without an initial training set of examples based on a full theoretical model. Basically, in this method, the value to be estimated is determined by the value of its neighbors in the sample space. The k-nearest neighbor (kNN) makes its estimations based on two basic measurements (distance and number of neighbors). Distance is expressed as the distance of the state or value from which the estimation will be made to other states or values. Euclidean, Minkowski or Manhattan criteria can be used to calculate the distance. K indicates how many of the nearest neighbors will be selected. The variables in the dataset entering the model are evaluated in space according to the determined K value. Then, input's state or value is determined as the majority of the k neighbors. In this work, Euclidean distance criteria and K=3 are selected.

The utilization of Naive Bayes (NB), a probabilistic machine learning approach grounded in Bayes' theorem, is observed in

various classification tasks. An improvement on naive Bayes, Gaussian Naive Bayes assumes that each class has a Gaussian distribution. A change in the value of one attribute in the algorithm has no direct impact on the value of any other attribute. The Gaussian Naive Bayes method is favored due to its simplicity and effectiveness as an algorithm, as it calculates the mean and standard deviation of the training data. It is built on a probabilistic model with an easy-to-code algorithm that makes predictions in real time. As a result, because it can be designed to reply rapidly to user inquiries, this algorithm is a common choice for solving real-world issues.

Logistic Regression (LR) serves as a predictive analytic method in the domain of statistical analysis used in classification issues that is based on the idea of likelihood. Similar to a linear regression model, logistic regression utilizes a more advanced cost function known as the "sigmoid function" instead of a linear or logistical function. The logistic regression hypothesis limits the cost function to values between 0 and 1.

Decision trees are utilized as models capable of acquiring fundamental decision rules for predicting the class or value of a target variable based on historical information derived from training data. The initiation of prediction takes place at the root node, which serves as the highest decision node. The last nodes of the tree are the leaf nodes when predictions of a category or a numerical value are made. There are intermediate nodes between the root node and the leaf node. Feature-based comparisons are made until the leaf node is reached.

The splits of these nodes are determined according to the entropy value. Random Forest (RF) employs the concept of bagging. A number of models are trained using different dataset subsets in bagging, and the final output is created by integrating the findings of all the models. Gradient Boosting (GB) benefits from the boosting strategy. Boosting is a sequential construction procedure that focuses on lowering prior model faults while increasing the effect of high-performance. The base model for random forests and Gradient boosting is decision trees [11].

Artificial Neural Networks (ANNs) [12] are models inspired by the human brain, designed to simulate learning processes. They mimic the structure and learning capabilities of biological neural networks found in the brain, allowing them to learn, store information, and generalize. In ANNs, artificial neurons are interconnected to form a network of connections. Artificial neural networks employ three primary layers, namely the input layer, the middleware (hidden) layer, and the output layer, which collectively govern the network's operations.

Through the input layer, data is sent to the network. In the hidden layers, it is processed before being transmitted to the output layer. Data processing is the process of transforming data input into output using the network's weight values. In order to ensure accurate output results for the given inputs, the network necessitates the establishment of suitable weights. The ANN model employed in this study is depicted in Figure 1.

3. Experiment Results

The proposed method is trained and evaluated Cleveland Heart Disease Dataset [9]. Parameters of proposed methods are given in Table 2. Proposed methods in this work are evaluated using Precision, Recall, F1-score and Accuracy metrics which of formulas are given in (1), (2), (3) and (4), respectively. These metrics are obtained from the contingency table shown in Table 3. An effective heart disease prediction system will have greater TP and TN while having lower FP and FN.

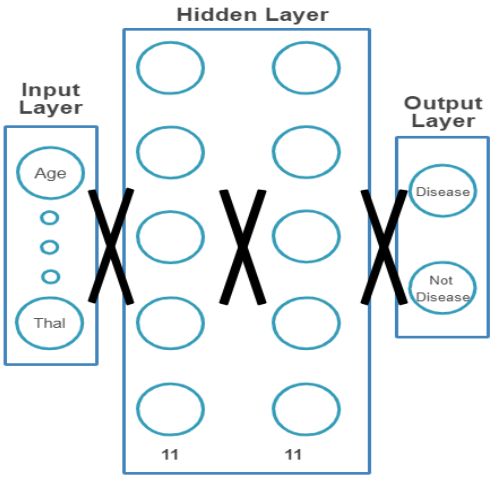

Figure 1. Proposed ANN model

Cleveland dataset contains data of 303 patients in total. These patients are randomly divided into 80% training and 20% test. The performance outcomes of the suggested methodologies on the Cleveland dataset are displayed in Table 4. For general comparison, accuracy is preferred. Table 4 reveals that the proposed Gradient Boosting model exhibits the highest accuracy value. This outcome serves as evidence that machine learning methods generally outperform neural network-based methods (ANN, Deep Neural Network) on smaller datasets. To further explore the capabilities of the proposed methods, an analysis of confusion matrices is conducted. When Figure 2 is examined, it becomes apparent that the Gradient Boosting method displays the least amount of confusion between classes.

Table 2. Experimental parameters of proposed methods

SVM	penalty=2, loss='squared_hinge', C=1.0
k-NN	n_neighbors=3, distance_metric = 'Minkowski'
GNB	var_smooting = 0.1
LR	max_iter=1000, random_state=1, solver='liblinear', penalty='l1'
RF	n_estimators=1000, random_state=1, max_leaf_nodes=20, min_samples_split=15
GB	random_state=1, n_estimators=100, max_leaf_nodes=3, loss='exponential', min_samples_leaf=20
ANN	optimizer = 'adam', loss = 'binary_crossentropy'

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + \ Recall} \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Table 4. Evaluation of proposed methods for heart disease prediction

Model	Precision	Recall	F1-Score	Accuracy
SVM	91.0	92.0	92.0	91.8
k-NN	87.0	88.0	87.0	86.8
GNB	88.0	89.5	88.0	88.5
LR	91.0	92.0	92.0	91.8
RF	88.0	89.5	88.5	91.8
GB	95.0	95.0	95.0	95.0
ANN	90.0	89.5	89.0	90.1

In terms of accuracy, Table 5 showcases a comparison between the proposed method and recent approaches in the literature on the Cleveland dataset. The accuracy results of recent approaches are derived from their original papers. As seen from Table 5, the proposed method outperforms the compared methods. Also, the processing speed performance of the proposed method is analyzed. Processing speed is obtained 16 ms on a PC with 2.80 GHz Quad core CPU, 8 GB RAM.

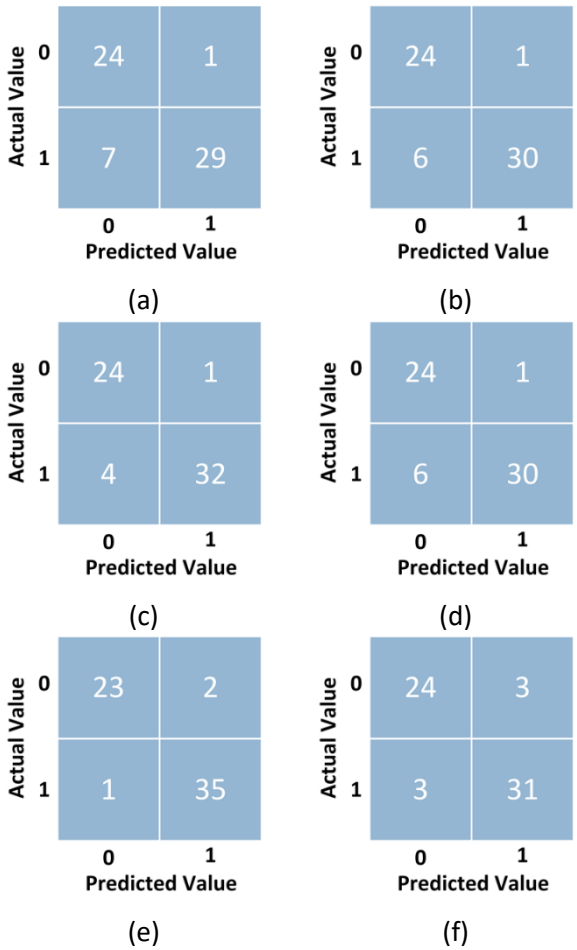

Figure 2. Confusion matrix of proposed methods (a) k-NN (b) GNB (c) LR (d) RF (e) GB (f) ANN

Table 5. Evaluation of proposed method and recent methods

Method	Accuracy
Dwedi et al. [2]	85.0
Kavitha et al. [3]	88.7
Mohan et al. [4]	87.0
Karadeniz et al. [5]	88.7
Tarawneh et al. [6]	89.2
Rani et al. [7]	86.6
Aarof et al. [8]	91.7
Ahamad et al. [13]	87.9
Chandrasekhar et al. [14]	93.4
Proposed Method	**95.0**

4. Conclusion

Heart disease continues to pose a substantial worldwide health challenge, significantly impacting both morbidity and mortality rates. Recent years have witnessed extensive research endeavors dedicated to diagnosing, predicting, and preventing heart disease. In this study, a range of machine learning techniques is employed to forecast heart disease occurrence. The proposed Gradient Boosting approach exhibits exceptional efficacy in predicting heart disease on the Cleveland dataset, surpassing contemporary methods in terms of accuracy. This investigation contributes to the field of heart disease prediction by introducing a dependable and precise methodology for early detection and prevention of cardiovascular disorders.

Article Information Form

Funding
The authors have not received any financial support for the research, authorship, or publication of this study.

Authors' Contribution
The authors contributed equally to the study.

The Declaration of Conflict of Interest/ Common Interest
No conflict of interest or common interest has been declared by the authors.

The Declaration of Ethics Committee Approval
This study does not require ethics committee permission or any special permission.

The Declaration of Research and Publication Ethics
The authors of the paper declare that they comply with the scientific, ethical and quotation rules of SAUJS in all processes of the paper and that they do not make any falsification on the data collected. In addition, they declare that Sakarya University Journal of Science and its editorial board have no responsibility for any ethical violations that may be encountered, and that this study has not been evaluated in any academic publication environment other than Sakarya University Journal of Science.

References

[1] World Health Organization, "Cardiovascular Diseases," World Health Organization, Available: https://www.who.int/health-topics/cardiovascular-diseases. Accessed: May 15, 2023.

[2] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," Neural Computing & Applications, vol. 29, pp. 685-693, 2018.

[3] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, R. S. Suraj, "Heart disease prediction using Hybrid Machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021.

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

[5] T. Karadeniz, G. Tokdemir, H. H. Maraş, "Ensemble methods for heart disease prediction," New Generation Computing, vol. 39, no. 3–4, pp. 569–581, 2021.

[6] M. Tarawneh, O. Embarak, “Hybrid approach for heart disease prediction using data mining techniques,” Advances in Internet, Data and Web Technologies, pp. 447–454, 2019.

[7] P. Rani, R. Kumar, N. M. Ahmed, A. Jain, “A decision support system for heart disease prediction based upon machine learning,” Journal of Reliable Intelligent Environments, vol. 7, no. 3, pp. 263–275, 2021.

[8] S. Arooj, S. ur Rehman, A. Imran, A. Almuhaimeed, A. K. Alzahrani, A. Alzahrani “A deep convolutional neural network for the early detection of heart disease,” Biomedicines, vol. 10, no. 11, p. 2796, 2022.

[9] UCI Machine Learning Repository, “Heart Disease Dataset,” Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease. Accessed: Feb 10, 2023.

[10] D. W. Aha, “Lazy Learning,” Berlin: Kluwer Academic Publishers, 1997.

[11] A. Cutler, D. R. Cutler, J.R. Stevens, “Random Forests,” in Ensemble Machine Learning, C. Zhang and Y. Ma, Eds. New York, NY: Springer, pp. 123-145, 2012.

[12] H. Ergezer, M. Dikmen, E. Özdemir, “Yapay sinir ağları ve tanıma sistemleri,” PiVOLKA, vol. 2, no.6, 11-17

[13] G. N. Ahamad, H. Fatima, S. M. Zakariya, M. Abbas, “Influence of optimal hyperparameters on the performance of machine”, Learning Algorithms for Predicting Heart Disease,” Processes, vol. 11, 734, 2023.

[14] N. Chandrasekhar, S. Peddakrishna, “Enhancing heart disease prediction accuracy through machine learning techniques and optimization,” Processes, vol. 11, no. 4, 1210, 2023.