Coronary Artery Disease Diagnosis Using Optimized Adaptive Ensemble Machine Learning Algorithm

Burak Kolukisa[1]*, Levent Yavuz[2], Ahmet Soran[1], Burcu Bakir-Gungor[1], Dilsad Tuncer[3], Ahmet Onen[2], V. Cagri Gungor[1]

[1] Department of Computer Engineering, Abdullah Gül University, Kayseri, Turkey.
[2] Department of Electrical Engineering, Abdullah Gül University, Kayseri, Turkey.
[3] Keydata Bilgi İşlem Teknoloji Sistemleri A.Ş., Ankara, Turkey.

* Corresponding author. Email: cagri.gungor@agu.edu.tr
Manuscript submitted June 2, 2019; accepted October 29, 2019.
doi: 10.17706/ijbbb.2020.10.1.58-65

Abstract: Cardiovascular diseases (CVD) involving the heart and blood vessels are reported as the leading causes of mortality worldwide. Coronary Artery Disease (CAD) is a major group of CVD in which presence of atherosclerotic plaques in coronary arteries leads to myocardial infarction or sudden cardiac death. In the past decades, several research efforts have been made to better understand the etiology of CAD, which will enable effective CAD diagnosis and treatment strategies. In this study, we have proposed a novel Self Optimized and Adaptive Ensemble Machine Learning Algorithm for the diagnosis of CAD. In our proposed method, the system automatically selects the most appropriate machine learning models. Our main goal is to design an Optimized Adaptive Ensemble Machine Learning Algorithm that works in different CAD datasets with high accuracy even with raw dataset. One of the important aspects of the proposed method is that the solution works on real-time data without using any pre-processing techniques on the datasets. Throughout this research attempt, we obtained 88.38% accuracy using two publicly available CAD diagnosis datasets.

Key words: Coronary artery disease, cardiovascular disease, machine learning, data mining, ensemble methods, classification.

1. Introduction

Cardiovascular Diseases (CVD) are one of the most prevalent diseases all over the world. According to the World Health Organization (WHO), the global prevalence of CVD is increasing and the deaths caused by CVD will reach approximately to 30 million by 2030. Coronary artery disease (CAD), myocardial infarction (MI), stroke, and peripheral vascular disease (PVD) are the major groups of CVD. In CAD, the presence of atherosclerotic plaques in coronary arteries leads to myocardial infarction or sudden cardiac death [1]. CAD is a complex disease such that both genetic architecture and its interaction with several environmental factors contribute to disease development. While the environmental risk factors of cardiovascular diseases include smoking, alcohol intake, physical inactivity, high caloric diets rich in fat, cholesterol, and sugar, infections, environmental chemicals and pollutants, and stress [2]; other risk factors include high blood pressure, high blood cholesterol, diabetes, overweight or obesity. Research studies conducted during the past decades resulted in important discoveries on CVD mechanisms and this lead to the development of highly effective cholesterol-lowering medications, and hence the death toll of CVD decreased. In addition to

these studies, machine learning and data mining approaches made it possible to predict diagnosis of CVD by checking particular values. Many studies have been conducted on the diagnosis of heart diseases, where researches have experimented several classifiers, feature selection and ensemble methods on several CVD datasets to improve classification [3] we focused only on public avaible datasets. Some of these studies are shown in Table 1 [1], [4]-[10]. Examining different performance metrics such as accuracy, sensitivity, specificity, FMeasure, Area Under Curve (AUC) and running time are also important for heart disease diagnosis. While, sensitivity indicates the percentage of prediction sick people as sick, F-Measure shows the balance between sensitivity and specificity.

Table 1. Comparision of Different Classification Methods for CAD Diagnosis

Reference	Method	FS	SN	SP	FM	AUC	ACC	RT	Dataset
Kemal Polat *et al.* [4]	KNN	No	92.30%	92.30%	-	-	87%	-	Cleveland
Resul Das *et al.* [5]	ANN	No	80.95%	95.91%	-	-	89.01%	-	Cleveland
Shouman *et al.* [6]	DT	-	77.90%	85.20%	-	-	84.1%	-	Cleveland
Alizadehsani *et al.* [7]	SMO	Yes	97.22%	79.31%	-	-	92.09%	-	Z-Alizadehsani
Rajalaxmi *et al.* [8]	NB	Yes	-	-	-	-	86.4%	-	Cleveland
Randa El-Biary *et al.* [9]	C4.5	Yes	-	-	-	-	78.54%	Yes	Cleveland
Luxmi Verma *et al.* [1]	MLR	Yes	-	-	-	-	90.28%	-	Cleveland
Frantisek Babic *et al.* [10]	SVM	Yes	-	-	-	-	86.67%	-	Z-Alizadehsani

FS: Feature Selection, SN: Sensitivity, SP: Specificity, FM: F-Measure, AUC: Area Under Curve, ACC: Accuracy, RT: Running Time

It is also important to note that traditional algorithms are not very adaptive and hence, when dataset structure changes, the performance of the existing studies decreases. In general, feature selection methods do not give reliable, applicable and sustainable performance results with different datasets, since they are specialized and valid only for specific datasets. The purpose of this article is to show that without applying any preprocessing or feature selection techniques, satisfactory result in terms of several performance metrics can be achieved on raw datasets of heart diseases. To address this need, in this study, we developed an Optimized Adaptive Ensemble Machine Learning Algorithm for the classification of CAD. We experimented our methodology on two publicly available data sets, i.e., Cleveland and Z-Alizadehsani Dataset. Since our solution is able to work with raw unprocessed data, it also has a potential application on the diagnosis of CVD in intensive care units, where accurate predictions needs to be made via fast processing of real time data.

This study is organized as follows: In section 2, we introduce the proposed Optimized Adaptive Ensemble Machine Learning Algorithm and publicly available CVD datasets. In section 3, we present performance evaluations of Adaptive Ensemble Machine Learning Algorithms. The last section concludes the paper.

2. Datasets and Methods

2.1. Datasets

In this work, we study with 2 different data sets; i.e Cleveland and Z-Alizadehsani. In each dataset, the number of attributes, the number of healthy (NOR) patients and the number of sick (CAD) patients are shown in Table 2.

Table 2. Features of Publicly Available Heart Disease Datasets

	Attribute	CAD	NOR	Total
Alizadehsani	55	216	87	303
Cleveland	13	165	138	303

In UCI repository, there are several heart disease datasets including Cleveland, Hungarian, Switzerland, all

other CVD datasets at UCI have several missing values. Cleveland dataset includes 303 samples and only 6 of these samples have missing and they are removed from the dataset. Z-Alizadehsani dataset includes 303 samples and it is collected at Terhan's Shaeheed Rajaei Cardiovascular, Medical and Research Center.

Although accuracy is a widely used performance metric to assess the performance of a classifier, this measure is not very suitable in medical domain. In cardiovascular disease diagnosis, one of the important success measures is sensitivity which is related with False Positives, in which predicting a sick person as healthy is a very undesired situation, F-Measure is another good measure, in medical domain because predicting healthy person as a sick is another important issue which we want to minimize this ratio. In order to overcome this disadvantage of accuracy, in this manuscript we focused on several performance measures including AUC, FMeasure, sensitivity and specificity.

2.2. Classification Methods

Classification is a data science of labeling a dataset consisting of numerical or nominal value by creating a model as a result of certain operations. Four main groups of classification algorithms shown in Fig. 1. [11] Classification algorithms can make false learning due to noise, to prevent this false learning ensemble methods can reduce the impact of these cause. In optimized adaptive ensemble machine learning algorithms, we select one classification algorithm in each group and ensemble them with soft voting and particle swarm optimization. k-Nearest Neighbor (kNN) algorithm is one of the most widely used methods in pattern recognition and classification problems. kNN algorithm can handle both continuous and discrete attributes. Since our CVD dataset contains both contiuous and discrete attributes, kNN performed well in our CVD dataset. We use minkowski as a distance metric in our kNN implementation. The sharpness between the classes began to be soft with the increase of k. If the number of classes is 2 in a dataset the k value is not recommended to pass the square root of the sample size.

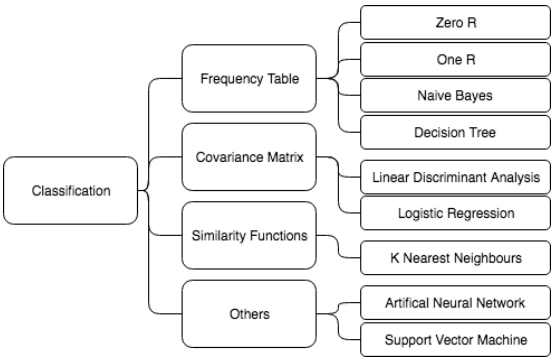

Fig. 1. Main groups of classification algorithms.

2.2.1. k-nearest neighbor (kNN)

k-Nearest Neighbor (kNN) algorithm is one of the most widely used methods in pattern recognition and classification problems. kNN algorithm can handle both continuous and discrete attributes. Since our CVD dataset contains both contiuous and discrete attributes, kNN performed well in our CVD dataset. We use minkowski as a distance metric in our kNN implementation. The sharpness between the classes began to be soft with the increase of k. If the number of classes is 2 in a dataset the k value is not recommended to pass the square root of the sample size.

2.2.2. Logistic regression (LR)

Logistic regression (LR) is a statistical machine learning algorithm that tries to define a logarithmic line that best distinguishes outcome variables on extreme ends. LR is the extended version of linear regression,

where it allows to build more complex decision boundaries by putting higher order polynomials such as stochastic gradient descent. In this way, it is expected to achieve better result on complex datasets.

2.2.3. Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) is also a machine learning algorithm. It is a method that takes attributes of linear combinations to separate classes. LDA clearly tries to find the difference between two classes. The aim of LDA is to prevent overfitting and to reduce cost.

2.2.4. Naïve bayes (NB)

Naïve Bayes (NB) is a classification technique that utilizes both from statistical and probabilistic methods. It is easy to build and gives goods results in large datasets. And also it adapts itself according to the value of the data.

2.2.5. Support vector machine (SVM)

Support Vector Machine (SVM) is a very effective classification algorithm and it is widely used in many different domains. SVM has a simple method in which it tries to separate two groups by pulling two parallel lines between two classes. While bringing lines closer, a common boundary line is obtained. This line is used as a decision boundary to separate the classes.

2.2.6. Ensemble methods

In contrary to basic classification algorithms, ensemble methods yield higher performance measures since they decrease error via learning what can cause noise in the dataset. Ensemble methods construct a model for each classifier and then classify data point by taking a weighted average of each classifier's predictions.

The weighted process is done with soft voting. In this study, we used k-Nearest Neighbor, Support Vector Machine, Naïve Bayes, Logistic Regression, Linear Discriminant Analysis classification algorithms and also combined them using bagging methods using the parameters shown Table 4. The main motivation of bagging algorithms is to find an optimum weight that gives the best result according to the dataset and prevent overfitting. In our proposed model, we use particle swarm optimization to tune the weights on soft voting. Briefly, it creates several particles that generate the initial weights randomly, and then, the particles move around and try to reach the global maxima.

Table 3. Parameters to Calculate the Optimum Values for Selected Machine Learning Algorithms

Machine Learning Algorithm	Parameter	Values
kNN	# of neighbors	[3 to 9, increment by 2]
LDA	tolerance	0.0001
LR	-	-
NB	-	-

3. The Proposed Methodology

In this paper, we experimented several machine learning algorithms for two different CVD datasets and tried to get most efficient, accurate and sensitive results. The proposed methods merge machine learning (ML) algorithms by empowering each algorithm's strength and also cover the flaws and weaknesses. For that purpose, some weights must be given to ML algorithms. To obtain the most appropriate weight value, brute force method can be used and scan all the possibilities. However, this requires high computational power and leads to waste of time. For example, when weights from 0.001 to 1 is tried via brute force method, 1015 possibilities for each dataset should be tried. To this end, Particle Swarm Optimization (PSO) offers a feasible solution to this problem. Swarms act like random flies (uniformly distributed) and when one swarm gets the highest point, others are going to die. In this way, appropriate weights can be calculated using PSO in a very short amount of time.

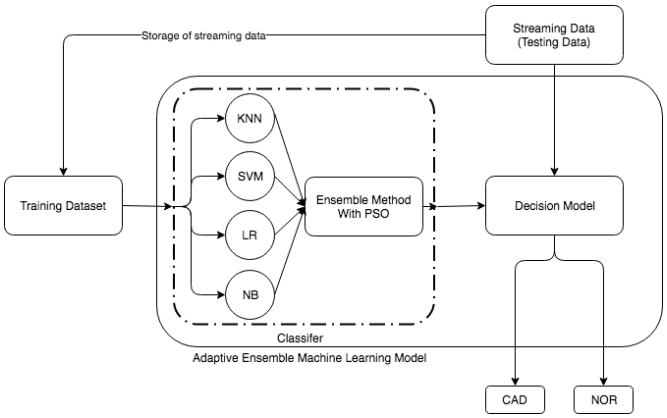

Fig. 2. Architecture of adaptive ensemble machine learning model for CAD classification.

With the proposed adaptive ensemble methodology, we offer an advanced system that is capable of making instant decisions according to the previous data collected by the medical doctors in intensive care units. However, new data can also be considered as supervision because of the adaptive scheme of the proposed method. Hence, further information will feed the system and decisions can be made by the adaptive ensemble method without changing the preprocessing functions. Decision model regenerates itself instead of using additional third-party solution periodically when new dataset is used.

Table 4. Performance Results of Adaptive Ensemble Machine Learning Model

	Methods	SN	FM	AUC	Accuracy	Time
Cleveland	kNN	58.01%	0.602	0.699	65.21%	0.20
	LR	79.4%	0.800	0.906	82.10%	0.09
	LDA	77.06%	0.792	0.905	82.06%	0.09
	NB	80.93%	0.805	0.891	82.42%	0.11
	SVM	82.15%	0.807	0.905	82.43%	0.11
	Ensemble	**78.28%**	**0.811**	**0.824**	**83.43%**	**0.21**
Z-Alizadehsani	kNN	90.03%	0.791	0.477	67.93%	0.13
	LR	90.34%	0.896	0.920	86.08%	0.12
	LDA	89.90%	0.896	0.903	86.06%	0.16
	NB	67.53%	0.773	0.873	74.23%	0.18
	SVM	93.12%	0.915	0.914	88.40%	0.13
	Ensemble	**92.63%**	**0.914**	**0.922**	**88.38%**	**0.38**

4. Performance Results and Discussion

During this study, 2 different publicly available CVD datasets, i.e., Cleveland and Z-Alizadehsani heart disease datasets, are used. In Table 4, the performance of each machine learning algorithm and the proposed approach, which is based on ensemble classification and particle swarm optimization, can be seen. In Cleveland dataset, we obtained the highest accuracy when we use poly kernel method and given parameters (γ = 10−1 , c = 10, and degree = 3). However, In Z-Alizadehsani dataset, radial basis function (rbf) gives better result with γ = 10−2 , c = 10 on SVM shown in Table 4. In this study, 10-fold cross validation is used because in data mining and machine learning community, it is known as a standard rule for performance estimation [12]. Using our adaptive ensemble classification methodology, we obtained 83.43%, 88.38% accuracies and 0.811 and 0.914 F-Measure values in Cleveland and Z-Alizadehsani datasets, respectively. In Z-Alizadehsani dataset, the AUC has also increased to 0.922. Our methods offer a solution to reduce noise, bias and covariance. Using our optimized adaptive ensemble methods, at this point our performance metrics are comparable to single classification methods. However, in insensitive care unit and different datasets our methods are more realistic.

5. Conclusion

The main purpose of this paper is to develop an adaptive ensemble classification machine learning model for several heart disease datasets which can successfully work on raw data. In this study, we have experimented a set of different heart disease datasets, including Cleveland and Z-Alizadehsani. We used several classification algorithms, including kNN, SVM, LR, LDA, NB. The weaknesses of single classification algorithms are their inconsistency against noise and bias. On the other hand, ensemble classifiers and soft voting offers a solution to reduce noise and prevent overfitting. Using our adaptive ensemble classification methodology, we obtained 83.43%, 88.38% accuracies and 0.811, 0.914 F-Measure in Cleveland and Z-Alizadehsani datasets, respectively. With its high accuracy levels, our methodology could be easily run on instant data available in intensive care units. Future work includes the investigation of the performance of the proposed approach in different datasets.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

All authors had conducted the research and approved the final version.

Acknowledgment

This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) TEYDEB program under Project no 3180177.

References

[1] Verma, L., Srivastava, S., & Negi, P. C. (2016). A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. *Journal of medical systems*, *40(7)*, 178.

[2] Heart attack risk. Retrieved from the website: https://seniorcarehomes.com/health-and-wellness-for-seniors-articles/reduce-heart-attack-risk/

[3] Kolukisa, B., Hacilar, H., Goy, G., Kus, M., Bakir-Gungor, B., Aral, A., & Gungor, V. C. (2018). Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease. *Proceedings of 2018 IEEE International Conference on Big Data (Big Data)* (pp. 2232-2238).

[4] Polat, K., Şahan, S., & Güneş, S. (2007). Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing. *Expert Systems with Applications*, *32(2)*, 625-631.

[5] Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert systems with applications*, *36(4)*, 7675-7680.

[6] Shouman, M., Turner, T., & Stocker, R. (2011). Using decision tree for diagnosing heart disease patients. *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121* (pp. 23-30). Australian Computer Society, Inc.

[7] Alizadehsani, R., Hosseini, M. J., Sani, Z. A., Ghandeharioun, A., & Boghrati, R. (2012). Diagnosis of coronary artery disease using cost-sensitive algorithms. *Proceedings of 2012 IEEE 12th International Conference on Data Mining Workshops* (pp. 9-16). IEEE.

[8] Subanya, B., & Rajalaxmi, R. R. (2014). Artificial bee colony based feature selection for effective cardiovascular disease diagnosis. *International Journal of Scientific & Engineering Research*, *5*(5), 606-612.

[9] El-Bialy, R., Salamay, M. A., Karam, O. H., & Khalifa, M. E. (2015). Feature analysis of coronary artery heart disease data sets. *Procedia Computer Science*, *65*, 459-468.

[10] Babič, F., Olejár, J., Vantová, Z., & Paralič, J. (2017). Predictive and descriptive analysis for heart disease diagnosis. *Proceedings of 2017 Federated Conference on Computer Science and Information Systems (FedCSIS)* (pp. 155-163). IEEE.

[11] Saedsayad.com. (2018). *Data Mining Map*. Retrieved from the website: http://www.saedsayad.com/

[12] Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, *1(3)*, 317-328.

Burak Kolukisa received the BS degrees in computer engineering from Erciyes University, Kayseri, Turkey, in 2016. Currently he is studying as a research assistant at the Abdullah Gül University. His research interests include data mining, machine learning, digital ımage processing and computer vision.

Levent Yavuz graduated from the Erciyes University Physics and Electrical Electronic Engineering and currently studying as a research assistant at the Abdullah Gül University. His established a company in ODTU Teknokent as an entrepreneur. He received 3 patents and the Mostar Science Medal. Subject of medal is "inversion nuclear decay as a result of neutralization nuclear weapons."His research interests include virtual power plant, microgrids, solar generation forecasting, machine learning, deep reinforcement learning.

Ahmet Soran received the BS and MS degrees in computer engineering from TOBB ETU, Ankara, Turkey, in 2009 and 2012, respectively, and the Ph.D. degree in computer science and engineering from the University of Nevada, Reno (UNR), in 2017. He joined the Computer Engineering Department at Abdullah Gul University, Kayseri, Turkey in 2018. His research interests include graph theory applications, device-to-device protocols, network and traffic management, network architecture, smart grids, cyber-security, and privacy.

Burcu Bakir-Gungor received her B.Sc. degree in biological sciences and bioengineering from Sabanci University; her M.Sc. degree in bioinformatics from Georgia Institute of Technology; and her PhD degree from Georgia Institute of Technology/Sabanci University. Now, she works as an assistant professor at the Department of Computer Engineering at Abdullah Gul University. Her research interests include bioinformatics, computational genomics, data mining and pattern recognition in bioinformatics.

Dilsad Tuncer received her B.Sc degree in statistics and computer science from Baskent University, Ankara, Turkey, in 2003. She received her M.Sc. degree in medical ınformatics from Middle East Technical University, Ankara, Turkey, in 2009. She has been working in different software companies for health sector in different positions for 16 years. She is currently working in Keydata Bilgi Teknolojileri as software grup manager.

Ahmet Onen received the B.Sc. degree in electrical-electronics engineering from Gaziantep University in 2005. He received the M.S degree in electrical-computer engineering from Clemson University in 2010 and his Ph.D. from Virginia Tech – Electrical and Computer Engineering Department in 2014. He is currently Turkish Transmission Network (TEIAS) Academic Advisor. He is currently working as an assoc. prof. at Abdullah Gul University.

V. Cagri Gungor received the B.S. and M.S. degrees in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 2001 and 2003, respectively. He received his Ph.D. degree in electrical and computer engineering from Georgia Institute of Technology, Atlanta, GA, USA, in 2007. Currently, he is a full professor and the Department Chair of Computer Engineering, Abdullah Gül University, Kayseri, Turkey. His current research interests are in smart grid communications, machine-to-machine communications, He is also the recipient of the TUBITAK Young Scientist Award in 2017.