ACADEMIC PLATFORM

Academic Platform Journal of Engineering and Science

journal homepage: https://dergipark.org.tr/tr/pub/apjes

Feature Selection Based Data Mining Approach for Coronary Artery Disease Diagnosis

*[1]Kemal Akyol

[1]Kastamonu University, Department of Computer Engineering, Kastamonu, 37100, Turkey, kakyol@kastamonu.edu.tr,

Research Paper Arrival Date: 18.03.2021 Accepted Date: 12.07.2021

Abstract

Cardiovascular diseases responsible for many deaths are very common and important health problems. According to World Health Organization, each year 17.7 million people die because of them. Coronary artery disease is the most important type of cardiovascular diseases that cause serious heart problems in patients, affecting the heart's function negatively. Being aware of the important attributes for this disease will help field-specialist in the analysis of routine laboratory test results of a patient coming internal medicine or another medicine unit except for the cardiology unit. In this study, it is aimed to determine the significance of attributes for coronary artery disease by utilizing Stability Selection method. In experiments, the attributes; 'Age', 'Atypical', 'Blood pressure', 'Current smoker', 'Diastolic murmur', 'Dyslipidemia', 'Diabetes mellitus', 'Ejection fraction', 'Erythrocyte sedimentation rate', 'Family history', 'Hypertension', 'Potassium', 'Nonanginal', 'Pulse rate', 'Q wave', 'Regional wall motion abnormality', 'Sex', 'St Depression', 'Triglyceride', 'Tinversion', 'Typical chest pain' and 'Valvular heart disease' were found important for each sub-dataset. Besides, the performances of four traditional machine learning algorithms were evaluated to detection of this disease. Logistic Regression algorithm outperformed others with %90.88 value of accuracy, 95.18% value of sensitivity, and 81.34% value of specificity.

Keywords: Medical data, coronary artery disease, attribute selection, stability selection, machine learning.

1. INTRODUCTION

Cardiovascular diseases responsible for many deaths are very common and important health problems. World Health Organization[1] reports that every year 17.7 million people die and 31% of all deaths occur as a result of cardiovascular diseases. 80% of cardiovascular deaths are heart attacks and strokes. Tobacco use, unhealthy diet and physical inactivity are among the causes that increase the risk of heart attack and stroke. They cause blood pressure, glucose, and lipids in addition to causing overweight and obesity for individuals. Coronary artery disease (CAD) is the most important type of cardiovascular diseases that cause serious heart problems in patients, affecting the heart's function negatively [1].

People who have high cardiovascular risk should be routinely controlled because there are many known and unknown risk factors considered to cause this disease [2]. Therefore, various learning algorithms have been used by researchers in order to find out these risk factors so far. For example, Alizadehsani et al. examined the effectiveness of features information by using Gain and Confidence methods. They achieved 94.08% classification accuracy [1]. Chagas et al. examined the dealing between alcohol consumption and CAD severity. They divided alcohol consumption into three categories: no, moderate and heavy. According to their study, the relationship between moderate alcohol consumption and severity of CAD is less than the relationship between heavy alcohol consumption and severity of CAD [3]. Yadav et al. presented an application based on association rule data mining in order to identify the hidden knowledge from medical dataset for the detection of CAD [4]. Ghadiri and Saniee used the Particle Swarm Optimization algorithm with a boosting approach to extract rules for detecting CAD patients [5]. Alizadehsani et al. (2) examined a preprocessing algorithm, rule-based and feature-based classifiers on the Z-Alizadeh Sani dataset. They evaluated the performances of different classifiers with 10-fold cross validation [6]. Nithya et al. improved the classification accuracy of the Support Vector Machine by utilizing the fuzzy logic. They evaluated their proposed model on the PIMA and Z-AlizadehSani [7]. Arabasadi et al. introduced a hybrid method that is the combine of Genetic Algorithm and Neural Network. The proposed hybrid method achieved highly accurate results on the Z-Alizadeh Sani dataset [8]. Alizadehsani (3) et al examined the effectiveness of a preprocessing algorithm on the Z-Alizadeh Sani dataset. They extracted three new features from the

[1] http://www.who.int/topics/cardiovascular_diseases/en/

* Corresponding Author: Kastamonu University, Department of Computer Engineering, Kastamonu, 37100, Turkey. e-mail: kakyol@kastamonu.edu.tr, Tel: +90 366 280 2978

Doi: 10.21541/apjes.899055

dataset and used them to enrich the primary dataset. Then, Naive Bayes, Sequential Minimal Optimization, K-Nearest Neighbors and C4.5 algorithms were employed with 10-fold cross validation in their study [9]. Qin et al. introduced an algorithm that investigates the validity and materiality of feature selection on the Z-Alizadeh Sani dataset [10]. Babič et al. focused on predictive analysis, which consists of some classifiers, and descriptive analysis which consists of association and decision rules. According to their studies, the Support Vector Machine provided the best accuracy with 86.67% on the Z-Alizadeh Sani dataset. [11]. In another study, Pathak et al. investigated the CAD in women. According to their studies, it was lower than men but it rose steadily after fifth decade and there is an immediate need to better understand cardiac symptoms in order to facilitate the diagnosis and treatment of the disease in women [12]. Recently, Shahid and Singh proposed hybridized emotional neural networks with particle swarm optimization. They used four different feature selection methods to improve the performance of their study, and so offered 88.34% accuracy with 10-fold cross validation. [13]. Velusamy and Ramasamy applied the ensemble voting technique, which included K-Nearest Neighbor, Random Forest, and Support Vector Machine. Moreover, they conducted experiments on original and balanced datasets containing 5 selected attributes using the random forest-based Boruta wrapper and the attribute significance of the SVM. According to their studies, the authors achieved 98.97% and 100% classification accuracies on the original and balanced datasets, respectively [14]. Nasarian et al. studied the heterogeneous hybrid feature selection method and the performance of different classifiers on balanced datasets. The authors achieved 92.58% classification accuracy with the Synthetic Minority Oversampling Technique (SMOTE) and eXtreme Gradient Boosting (XGBoost) classifier they used in their proposed approach [15].

Understanding risk factors associated with CAD, adopting healthy lifestyle and preventive strategies, and early detection can be crucial to the management and elimination of such cardiovascular diseases. This disease has been proven to be the leading cause of death in both developed and developing countries [16]. Therefore, there is a need for a machine learning approach-based expert system that can diagnose heart disease more accurately in these countries with a shortage of medical resources and field experts. Moreover, the detection of properties important for this disease is used to increase the success of such a system. Being aware of the important attributes for this disease will help field-specialist in the analysis of routine laboratory test results of a patient coming internal medicine or another medicine unit except for the cardiology unit. By taking this data into consideration, the relevant physician will be able to guide the patient to the cardiology field-specialist and thus an important step will be taken for the patient's health. Based on these information, it is aimed to determine the significance of attributes for CAD by utilizing Stability Selection (SS) method in this study. Machine learning was carried out on the training and test sets which includes 22 features that intersect among the features that are found important in each fold with the SS method. So, it is thought that the results obtained from this study will be useful or give an idea to the field-specialists investigating the causes of this disease at least. And also, the best algorithm is investigated for detecting this disease.

In this context, the rest of this paper is organized as follows: Section 2 describes the material and methods. Section 3 offers the experimental results and discussion. Finally, Section 4 concluded the study with final remarks.

2. MATERIAL AND METHODS

2.1. Medical Dataset

The Z-Alizadeh Sani dataset[2] which consists of 303 people data in Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Center are used in this study. This dataset consists of 55 attributes including 51 test values and personal information such as age and weight, and 1 outcome attribute which has two possible categories, "CAD" or "Normal". If patient's CAD diameter value is greater than or equal to 50%, the patient is categorized as "CAD", otherwise, the patient is categorized as "normal".

2.2. Attribute Selection

Attributes set describes an information about any disease. But, all attributes may not be related to disease. Attribute selection is substantial step in order to avoid over-fitting and detect the best features. Feature selection is widely used in traditional machine learning based models [17–22], as well as in deep learning based models [23–28].

SS method provides information about the attributes for the result variable. In this method, the dataset is randomly shuffled many times. The combination of the "Least Absolute Contraction and Selection Operator" (LASSO) and its sequential regressions are performed to determine the significance of the feature [29]. Random Lasso, also known as SS, introduced in [30] offers consistent variable selection. Support Vector Regression generally has good generalization, and LASSO regression includes the L1 constraint in choosing a small subset of features in the dataset to explain the target variable [29].

2.3. Data mining

The purpose of data mining, which is a discipline that includes statistics and computer science algorithms, is to discover meaningful information from data [31]. Compared to statistical techniques, data mining techniques are often more powerful, flexible and effective for the knowledge discovery [32]. The performances of machine learning

[2] https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani#

algorithms are handled for data mining. These algorithms perform the knowledge inference from data by using mathematical and statistical methods. So, unknown data are classified or predicted by utilizing the model. In this study, the classification performances of Random Forest (RF), Logistic Regression (LR), Multi-layer Perceptron (MLP), and Diagonal Linear Discriminant Analysis (DLDA) algorithms which are introduced in [33], respectively are examined.

2.4. K-fold cross validation

The *k*-fold cross validation technique splits the dataset into sub-datasets in order to provide consistent data distribution. k-1 piece of the sub-datasets is used as training set, the remaining one sub-dataset is used as test set. This process is repeated until all sub-datasets are tested. So, the performances of the learning algorithms are evaluated k times. The average of these performances is considered as the overall success rate [34].

2.5. Performance metrics

Accuracy (Acc), Sensitivity (Sen) and Specificity (Spe) metrics are used to evaluate the performances of the algorithms. The Acc metric indicates the ratio of the number of correctly classified CAD and non-CAD samples to the number of all instances. The Sen metric is ratio of the number of correctly classified CAD patients to the number of all CAD patients. The Spe metric is ratio of the number of correctly classified control samples called non-CAD to the number of all control samples. These metrics are presented in between Equations 1 and 3 [35]:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+TN+FP}} \quad (1)$$

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

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (3)$$

Here, TP: true positive, TN: true negative, FP: false positive, FN: false negative.

3. EXPERIMENTS

In this study, experiments are carried out by using Python 2.7 imported "mlpy and sklearn" machine learning libraries. Firstly, the dataset is cleaned and prepared within the frame of three steps:

a) The records which include null value(s) are excluded from the raw dataset.
b) Categorical information like yes/no in the raw data file are converted to 1/0 categorical values.
c) All data in dataset are normalized into range from 0 to 1.

In this context, 158 records, which include 43 patients having CAD and 115 patients having non-CAD, are prepared for experimental studies. In later stage, the dataset is divided into 80-20% training and test sets within the frame of 5-fold cross-validation technique. That is, the sub-datasets are generated 5 times. Hence, 234 training and 59 test sets are obtained in each iteration. After this step, the SS attribute selection method is applied to these sub-datasets to detect the most effective attributes for target variable and avoid over-fitting. Randomized Lasso is used for computing the importance of attributes on each resampling. Its parameters settings are presented in Table 1.

Table 1. The Randomized Lasso model parameters setting.

Parameter	Meaning	Value type	Value
alpha	The regularization parameter	float	0.001
scaling	Scale	float	0.5
sample_fraction	The fraction of samples	float	0.75
n_resampling	Number of randomized models	integer	200
selection_threshold	A treshold value for a question of "which features should be selected?"	float	0.25
normalize	Normalization	boolean	True
max_iter	Maximum number of iterations	integer	500

4. RESULTS AND DISCUSSION

The significance values of attributes are presented in Figure 1. It is noted that the order of the attributes in this figure is the same as the public dataset used in this study. Also, the number of important attributes for each sub-dataset is presented in Table 2.

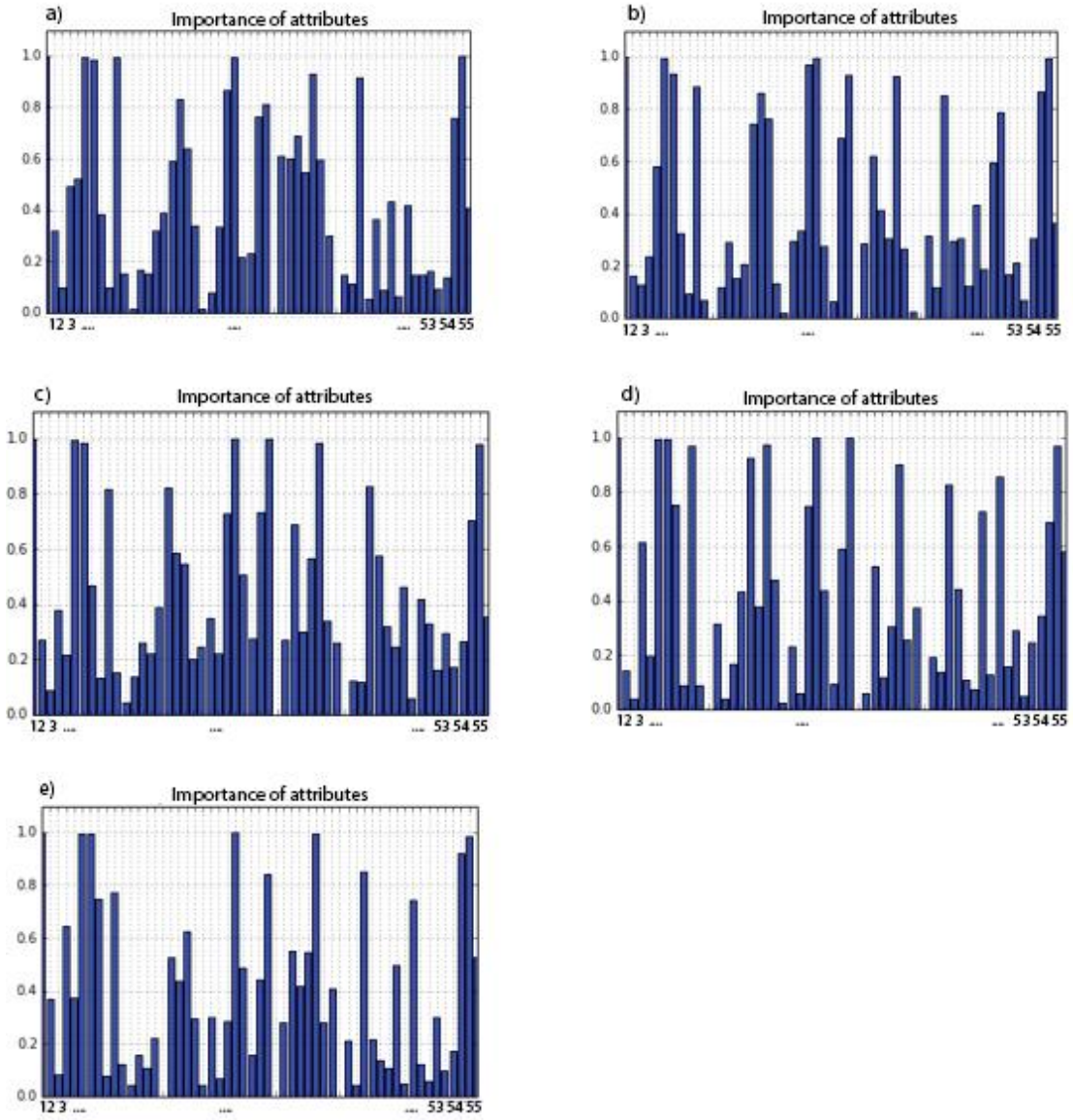


Figure 1. The significance values of attributes: a) Training data no. #1, b) Training data no. #2, c) Training data no. #3, d) Training data no. #4, e) Training data no. #5.

Table 2. Information about the significances of attributes.

Datasets	The number of important features
Training data no. #1	33
Training data no. #2	34
Training data no. #3	37
Training data no. #4	31
Training data no. #5	32

The study showed that 'Age', 'Atypical', 'Blood pressure', 'Current smoker', 'Diastolic murmur', 'Dyslipidemia', 'Diabetes mellitus', 'Ejection fraction', 'Erythrocyte sedimentation rate', 'Family history', 'Hypertension', 'Potassium', 'Nonanginal', 'Pulse rate', 'Q wave', 'Regional wall motion abnormality', 'Sex', 'St Depression', 'Triglyceride', 'Tinversion', 'Typical chest pain' and 'Valvular heart disease' attributes are found important for each sub-dataset. What is meant by this sentence is that empirical studies performed on all sub-datasets show that these 22 attributes given in Table 3 are found important in each fold. For example, the 'Current smoker' attribute is found important for CAD in the sub-datasets. On the other hand, 'Exertional CP' and 'BBB' attributes are insignificant for the sub-datasets. In addition, the important attributes which are found out as significant for all sub-datasets are presented in Figure 2.

Table 3. The weight information of the important attributes.

	Training data no. #1	Training data no. #2	Training data no. #3	Training data no. #4	Training data no. #5
Age	1	1	1	1	1
Atypical	0.745	0.705	0.735	0.54	0.535
Blood pressure	0.74	0.63	0.8	0.395	0.36
Current Smoker	0.345	0.48	0.335	0.76	0.77
Diastolic Murmur	0.85	0.695	0.96	0.725	0.27
Dyslipidemia	0.645	0.81	0.745	0.92	0.485
Diabetes mellitus	0.98	0.995	0.995	1	1
Ejection fraction	0.845	0.7	0.85	0.685	0.855
Erythrocyte sedimentation rate	0.49	0.435	0.4	0.655	0.48
Family history	0.985	0.865	0.9	0.975	0.775
Hypertension	0.995	0.985	0.91	0.985	1
Potassium	0.505	0.415	0.605	0.875	0.73
Nonanginal	0.835	0.99	0.94	0.98	0.925
Pulse rate	0.625	0.515	0.765	0.995	0.63
Q Wave	0.62	0.595	0.62	0.5	0.645
Regional wall motion abnormality	0.99	0.975	0.99	0.94	1
Sex	0.46	0.32	0.275	0.7	0.575
St Depression	0.575	0.625	0.255	0.335	0.485
Triglyceride	0.945	0.845	0.805	0.775	0.885
T inversion	0.9	0.995	0.91	0.895	1
Typical Chest Pain	1	0.99	0.99	1	1
Valvular heart disease	0.45	0.35	0.405	0.605	0.545

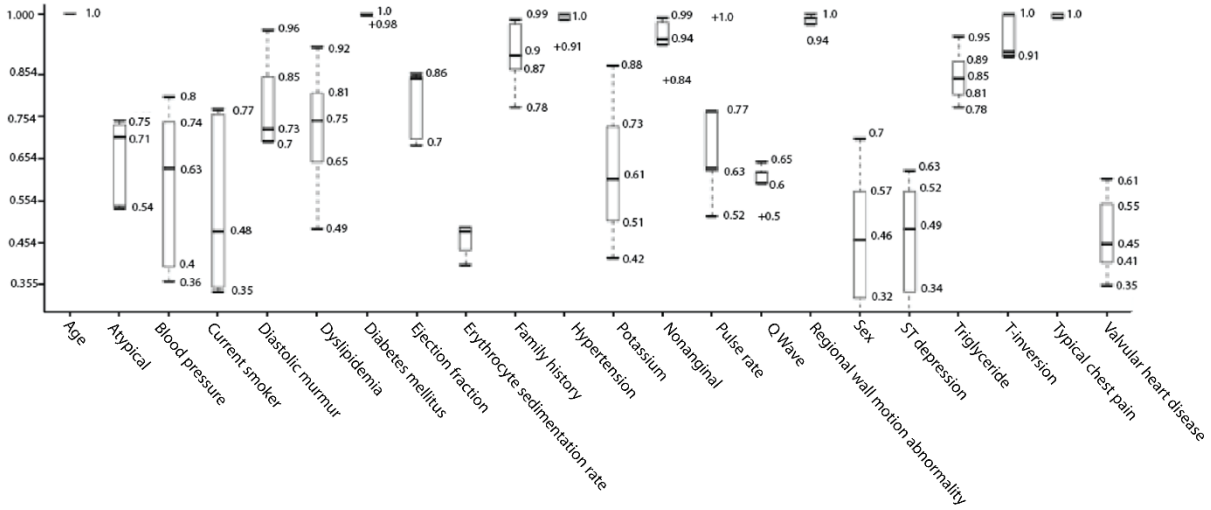


Figure 2. The weights of important attributes.

In further stage, the training and test sets which include best attributes are obtained respectively. By sending the sub-datasets as input data to the RF, LR, MLP and DLDA classifier algorithms, the machine learnings are carried out one by one. Table 4 presents detail settings about these algorithms' parameters. The prediction results are presented in Table 5 for sub-datasets including the best attributes. For example, 89.47% Acc, 92.86% Sen and 80.0% Spe measures are achieved on the test dataset no. 1 (Fold no. 1) by LR algorithm. According to Figure 3, the LR algorithm superior to others in terms of the Acc and Spe metrics. Also, the Sen metric of this algorithm quite near to the RF algorithm.

Table 4. The parameters setting of the learning algorithms.

Algorithm	Parameter	Value
RF	*n_estimators*	*10*
	criterion	*gini*
	max_features	*auto*
	bootstrap	*True*
	min_samples_split	2
	min_samples_leaf	1
LR	penalty	l2
	C	1.0
	solver	liblinear
	max_iter	100
MLP	solver	lbfgs
	hidden_layer_sizes	len(attributes), 2* len(attributes)+1
	activation	*relu*
	alpha	1e-5
	learning_rate	*'constant'*
	learning_rate_init	*0.001*
DLDA	delta	*0.1*

Table 5. Experimental results.

	Learning Algorithms	Acc %	Sen %	Spe %
Fold no. 1	RF	85.96	97.62	53.33
	LR	89.47	92.86	80.00
	MLP	82.46	88.10	66.67
	DLDA	87.72	92.86	73.33
Fold no. 2	RF	87.72	97.30	70.00
	LR	91.23	100.00	75.00
	MLP	64.91	100.00	0.00
	DLDA	85.96	89.19	80.00
Fold no. 3	RF	87.72	95.00	70.59
	LR	92.98	95.00	88.24
	MLP	77.19	87.50	52.94
	DLDA	92.98	95.00	88.24
Fold no. 4	RF	85.96	95.00	64.71
	LR	87.72	95.00	70.59
	MLP	75.44	90.00	41.18
	DLDA	85.96	90.00	76.47
Fold no. 5	RF	92.98	93.02	92.86
	LR	92.98	93.02	92.86
	MLP	84.21	86.05	78.57
	DLDA	92.98	93.02	92.86

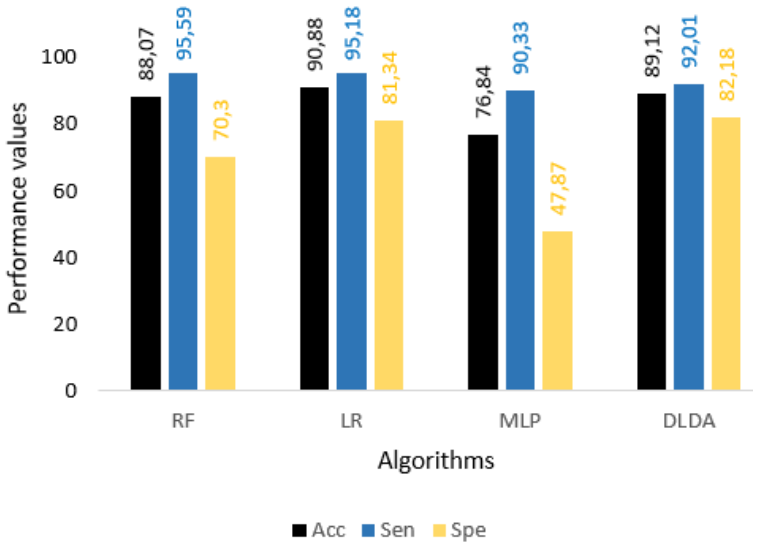


Figure 3. Average results of the algorithms.

In addition, the performance of the proposed study was compared with the existing methods given in Table 6. According to this table, the performance of the proposed study is slightly lower compared to other studies, except for [11,13].

However, the studies of [1,4,7–9] were not focused on feature selection. They only simply classified the data. On the other hand, the studies of [14,15] focused on feature selection.

Table 6. The comparison of the relevant studies.

Study	Method	Acc (%)	Sen (%)	Spe (%)
Alizadehsani et al. [1]	Gain and Confidence	93.40 ±5.53	95.83	87.36
Yadav et al. [4]	Association rule	93.75	95.65	91.53
Nithya et al. [7]	SVM and fuzzy logic	97.03	97.16	96.85
Arabasadi et al. [8]	Combination of Genetic Algorithm and Neural Network	93.85	97.00	92.00
Alizadehsani et al. [9]	Preprocessing and Sequential Minimal Optimization classifier	92.09	97.22	---
Babič [11]	SVM	86.67	---	---
Shahid and Singh [13]	PSO based emotional neural networks	88.34	91.85	78.98
Velusamy and Ramasamy [14]	Feature selection and ensemble voting technique Original dataset Balanced dataset	 98.97 100	 100 100	 96.3 100
Nasarian et al. [15]	Feature selection, SMOTE and XGBoost classifier	92.58	92.99	---
Proposed Method	**Feature selection and LR**	**90.88**	**95.18**	**81.34**

5. CONCLUSION

CAD is the most important type of cardiovascular diseases that cause serious heart problems in patients, affecting the heart's function negatively. Having knowledge about the important attributes for this disease will help the field-specialist in evaluating a patient's routine laboratory test results. Based on these information, it is aimed to determine the significance of attributes for CAD by utilizing SS method in this study. And also, the best algorithm is investigated for detection of this disease. Experimental results clearly show that LR, RF, MLP and DLDA machine learning algorithms are successful for best attributes datasets. The LR algorithm is just a little greater than the others within the frame of overall accuracy. It gives 90.88% accuracy. In this context, it can be said that the detection of best attributes is successful. And these attributes can be useful for field-specialists and health care systems in order to evaluate the condition of the patients. In future works, it is aimed to work with big datasets which include more attributes in order to obtain better achievement and more significant results. Also, it is aimed at experiments with the deep learning approaches that are between [33-35].

Acknowledgements

Author would like to thank Arabasadi et al. [8] for providing the Z-Alizadeh Sani dataset.

Author contributions:
Kemal Akyol: Conceptualization, Methodology, Writing – Original Draft, Software

Conflict of Interest: Author declares that he has no conflict of interest.

Financial Disclosure: The author declared that this study has received no financial support.

REFERENCES

[1] R. Alizadehsani, J. Habibi, M.J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian and Z.A. Sani, "A data mining approach for diagnosis of coronary artery disease", Comput. Methods Programs Biomed. vol. 111, no. 1, pp. 52–61, 2013.

[2] R. Roberts, "A genetic basis for coronary artery disease", Trends Cardiovasc. Med. vol. 25, no. 3, pp. 171–178, 2015.

[3] P. Chagas, L. Mazocco, J. da C.E. Piccoli, T.M. Ardenghi, L. Badimon, P.R.A. Caramori, L. Pellanda, I. Gomes and C.H.A. Schwanke, "Association of alcohol consumption with coronary artery disease severity", Clin. Nutr. vol. 36, no. 4, pp. 1036–1039, 2017.

[4] C. Yadav, S. Lade and M.K. Suman, "Predictive Analysis for the Diagnosis of Coronary Artery Disease using

Association Rule Mining", International Journal of Computer Applications, vol. 87, pp. 9-13, 2014.
[5] N. Ghadiri Hedeshi and M. Saniee Abadeh, "Coronary artery disease detection using a fuzzy-boosting PSO approach", Comput. Intell. Neurosci. vol. 2014, pp. 1-13, 2014.
[6] R. Alizadehsani, M.J. Hosseini, R. Boghrati, A. Ghandeharioun, F. Khozeimeh and Z.A. Sani, "Exerting Cost-Sensitive and Feature Creation Algorithms for Coronary Artery Disease Diagnosis", Int. J. Knowl. Discov. Bioinforma. vol. 3, no. 1, pp. 59–79, 2013.
[7] S. Nithya, C. Suresh and G. Dhas, "Fuzzy Logic Based Improved Support Vector Machine (F-Isvm) Classifier for Heart Disease Classification", ARPN Journal of Engineering and Applied Sciences, vol. 10, no. 16, pp. 6957-964, 2015.
[8] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei and A.A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm", Comput. Methods Programs Biomed. vol. 141, 19–26, 2017.
[9] R. Alizadehsani, M.J. Hosseini, Z.A. Sani, A. Ghandeharioun and R. Boghrati, "Diagnosis of coronary artery disease using cost-sensitive algorithms", in: Proc. - 12th IEEE Int. Conf. Data Min. Work. ICDMW, Brussels, Belgium, pp. 9–16, 2012.
[10] C.J. Qin, Q. Guan and X.P. Wang, "Application of ensemble algorithm integrating multiple criteria feature selection in coronary heart disease detection", Biomed. Eng. - Appl. Basis Commun. vol. 29, no. 6, pp. 1-11, 2017.
[11] F. Babic, J. Olejar, Z. Vantova and J. Paralic, "Predictive and descriptive analysis for heart disease diagnosis", in: Proc. 2017 Fed. Conf. Comput. Sci. Inf. Syst. FedCSIS 2017, Institute of Electrical and Electronics Engineers Inc., pp. 155–163, 2017.
[12] L.A. Pathak, S. Shirodkar, R. Ruparelia and J. Rajebahadur, "Coronary artery disease in women", Indian Heart J. vol. 69, no. 4, pp. 532–538, 2017.
[13] A.H. Shahid and M.P. Singh, "A Novel Approach for Coronary Artery Disease Diagnosis using Hybrid Particle Swarm Optimization based Emotional Neural Network", Biocybern. Biomed. Eng. vol. 40, no. 4, pp. 1568–1585, 2020.
[14] D. Velusamy and K. Ramasamy, "Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset", Comput. Methods Programs Biomed. vol. 198, pp. 1-13, 2021.
[15] E. Nasarian, M. Abdar, M.A. Fahami, R. Alizadehsani, S. Hussain, M.E. Basiri, M. Zomorodi-Moghadam, X. Zhou, P. Pławiak, U.R. Acharya, R.S. Tan and N. Sarrafzadegan, "Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach", Pattern Recognit. Lett. vol. 133, pp. 33–40, 2020.
[16] A.K. Malakar, D. Choudhury, B. Halder, P. Paul, A. Uddin and S. Chakraborty, "A review on coronary artery disease, its risk factors, and therapeutics", J. Cell. Physiol. vol. 234, no. 10, pp. 16812–16823, 2019.
[17] D. Effrosynidis and A. Arampatzis, "An evaluation of feature selection methods for environmental data", Ecol. Inform. vol. 61, pp. 1-10, 2021.
[18] K.P. Muhammed Niyas and P. Thiyagarajan, "Feature selection using efficient fusion of Fisher Score and greedy searching for Alzheimer's classification", J. King Saud Univ. - Comput. Inf. Sci., in press, 2021. https://doi.org/10.1016/j.jksuci.2020.12.009.
[19] K.K. Kavitha and A. Kangaiammal, "Correlation-based high distinction feature selection in digital mammogram", Mater. Today Proc., vol. 147, no. 5, e218-e227, 2020.
[20] P.D. Sheth, S.T. Patil and M.L. Dhore, "Evolutionary computing for clinical dataset classification using a novel feature selection algorithm", J. King Saud Univ. - Comput. Inf. Sci., in press, 2020. https://doi.org/10.1016/j.jksuci.2020.12.012.
[21] F. Amini and G. Hu, "A two-layer feature selection method using Genetic Algorithm and Elastic Net", Expert Syst. Appl. vol. 166, pp. 1-10, 2021.
[22] S.B. Chen, Y.M. Zhang, C.H.Q. Ding, J. Zhang and B. Luo, "Extended adaptive Lasso for multi-class and multi-label feature selection", Knowledge-Based Syst. vol. 173, pp. 28–36, 2019.
[23] A.U. Haq, A. Zeb, Z. Lei and D. Zhang, "Forecasting daily stock trend using multi-filter feature selection and deep learning", Expert Syst. Appl., vol. 168, pp. 1-8, 2021.
[24] M. Toğaçar, B. Ergen and Z. Cömert, "Classification of white blood cells using deep features obtained from Convolutional Neural Network models based on the combination of feature selection methods", Appl. Soft Comput. J. vol. 97, pp. 1-10, 2020.
[25] T. Niu, J. Wang, H. Lu, W. Yang and P. Du, "Developing a deep learning framework with two-stage feature selection for multivariate financial time series forecasting", Expert Syst. Appl. vol. 148, pp. 1-17. 2020.
[26] W. Tian, Z. Liu, L. Li, S. Zhang and C. Li, "Identification of abnormal conditions in high-dimensional chemical process based on feature selection and deep learning", Chinese J. Chem. Eng. vol. 28, no. 7, pp. 1875-1883, 2020.
[27] W. Kim, Y. Han, K.J. Kim and K.W. Song, "Electricity load forecasting using advanced feature selection and optimal deep learning model for the variable refrigerant flow systems", Energy Reports. vol. 6, 2604–2618, 2020.
[28] H. Shi, H. Li, D. Zhang, C. Cheng and X. Cao, "An efficient feature generation approach based on deep learning and feature selection techniques for traffic classification", Comput. Networks. vol. 132, pp. 81–98, 2018.
[29] F. Mordelet, J. Horton, A.J. Hartemink, B.E. Engelhardt and R. Gordân, "Stability selection for regression-based models of transcription factor-DNA binding specificity", Bioinformatics, vol. 20, no. 13, i117-125, 2013.
[30] N. Meinshausen and P. Bühlmann, "Stability selection", J. R. Stat. Soc. Ser. B. vol. 72, no. 4, pp. 417–473, 2010.

[31] C. Zucco, "Data Mining in Bioinformatics", Encycl. Bioinforma. Comput. Biol., Elsevier, vol. 1, pp. 328–335, 2019.
[32] M. Kantardzic, Data mining : concepts, models, methods, and algorithms, 3rd Edition, 2019.
[33] T. Hastie, R. Tibshirani and J. Friedman, Elements of Statistical Learning, 2nd edition, 2009.
[34] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence, vol. 2, pp. 1137–1143, 1995.
[35] S.A. Shaikh, "Measures Derived from a 2 x 2 Table for an Accuracy of a Diagnostic Test", J. Biom. Biostat. vol. 2, no. 5, pp. 1–4, 2011.