Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

A STUDY ON ANALYSIS OF CARDIOVASCULAR DISEASES

K. Akyol[1*], B. Şen[2]

[1]Department of Computer Engineering, Kastamonu University, Kastamonu - 37100, Turkey
[2]Department of Computer Engineering, Yıldırım Beyazıt University, Ankara - 06030, Turkey

Received: 04 February 2018 / Accepted: 14 April 2018 / Published online: 01 May 2018

ABSTRACT

Commonly seen in adults, cardiovascular diseases are important health problems. In order to investigate the causes of the diseases which affect the heart and the blood vessels, two datasets were used. First, one of these datasets is publicly available dataset provided by the University of California, Irvine Machine Learning Repository. The effects of biochemistry and hemogram laboratory test results for the Cardiovascular Diseases were analyzed by using the second dataset which was taken from the cardiology and other services of Yildirim Beyazit University Ataturk Training and Research Hospital. ICD-10 (International Statistical Classification of Diseases and Related Health Problems) booklet was taken as a reference for the patient and control groups. The successes of the classifier algorithms indicated that working with the datasets which have only limited number of attributes is not right step.

Keywords: Cardiovascular diseases; logistic regression; machine learning; medical data; random forest.

Author Correspondence, e-mail: kakyol@kastamonu.edu.tr

doi: http://dx.doi.org/10.4314/jfas.v10i2.20

1. INTRODUCTION

Mostly seen in adult population and being cause of many deaths, Cardiovascular Diseases (CD) are important health problems [1]. The risk prediction is a key element for identifying the CD. People who are at very high cardiovascular risk should be routinely controlled since there are many known and unknown risk factors considered to cause this disease. The risk calculators generate a “score” which estimates the probability of this disease using these factors. Several of the risk scoring studies are as follows: Framingham, Procam and Tekharf. In Framingham study, clinical risk factors were evaluated for prediction of the people’s cardiovascular disease risk in ten years [2]. In Procam study, a system was developed for preventing and early diagnosis of the disease [3]. In Tekharf study, the risk factors, especially in Turkish adults, were investigated for the disease [4]. Another study is that the risk of Coronary Artery Diseases (CAD) was predicted through Logistics Risk Score which has a considerableplace in the treatment of patients which have CD within the framework of identification of the risk factors for it [5]. The specialists of the field benefit from these score values in decision making process [6]. Researches on health are carried out through analysis of the data obtained from a particular group of patients, and the importance of such studies is great for the control of the disease. But, the successful characterization of the factors that causes to CD could not be achieved as well. Because, it is thought to be the effect of unknown confidential information, family background, socio-demographic characteristics of the disease. The correct identification of the risk factors related to the disease has great importance both for the prevention of disease and the treatment of patients. It could not be said that the risk factors such as smoking, obesity, high blood pressure, diabetes cause 100 % of the disease [1]. On the other hand, early detection should be performed for the people who are at very high cardiovascular risk. With this study, it is aimed to investigate the causes of CD, utilizing the datasets. Some of the studies in the relevant literature on the CD are as follows: Battes et al. reviewed 16 risk prediction models for secondary prevention and reported discrimination capability by incremental C-statistic. In their study, age, male, sex, current smoking, diabetes, low body mass index, CD events in the past years, history of heart failure as the key determinants of subsequent CD were identified [7]. Wang et al. identified the different levels effects of the conventional risk factors for CAD and they also found that the other risk factors can cause this disease as well [8]. Karaolis et al. performed the data-mining technique, using C4.5 decision tree algorithm to identify the risk factors for the disease. The most important risk factors for the disease were extracted utilizing the classification rules in their study [9]. Srinivas et al. researched the causes of heart attack, using decision tree, artificial neural networks and Bayesian classifiers. The study analyzes the Behavioral Risk Factor Surveillance System in order to test whether self-reported cardiovascular disease ratios are higher in Singareni coal mining regions in Andhra Pradesh state, India, compared to other regions after control for other risks. Dependent variables include self-reported measures of being diagnosed with CD or with a specific form of CD including chest pain, stroke and heart attack. Beside regular attributes, other general attributes Body Mass Index, physician supply, age, ethnicity, education, income, and others are used for prediction [10]. Abdullah and Rajalaxmi researched the prediction of CAD, using Random Forest algorithm so as to help doctors. Their studies can help the medical practitioners predict CAD with its various events and how it might be related with different segments of the population [11]. Bialy et al. aimed to apply an integration of the results of the machine learning analysis applied on different datasets for the CAD disease's detection. They applied Fast decision tree and pruned C4.5 tree algorithms to the resulted trees which were extracted from different datasets and compared. They obtained the 78.06% classification accuracy of the collected dataset. This accuracy is higher than the average of the classification accuracy of all separate datasets [12]. Jabbar et al. proposed a new algorithm which combines KNN and genetic algorithm to enhance the accuracy in diagnosis of heart disease [13]. Shu et al. summarized the basic principles of bioinformatics and core algorithms and tools applicable in order to investigate the cardiovascular diseases by capturing effectively the multilevel complexity, big data science which has emerged as a powerful strategy for the diseases [14]. Ave et al. introduced a device for early detection of cardiovascular with Photoplethysmogram signal which is blood volume measurement method based on optic principle. Volume change caused by vessel pressure can be detected with illuminating skin by the light from photodiode. Developed application in smartphone/tablet based on Android platform is designed with easily interpretable and attractive user interface targeting common people. Furthermore, this application is completed with many features that can optimize user experience in this PPG signal explication based on previous work [15]. Thakur and Ramzan introduced the basic challenges and scope of big-data in Cardiovascular and also identified the big-data capabilities for effective big-data based strategies. They used Hadoop program, which is scalable, cost-effective, flexible, fast, and resilient, as compared to other methods for analysis of big data [16]. Singh et al. provided a novel approach which is based on Structural Equation Modeling and Fuzzy Cognitive Map in order to solve the shortcomings of other methods and designed a very robust and reasonably accurate model. Canadian Community Health Survey, 2012 dataset was used to test the approach [17]. Guillen et al. focused on the innovative concepts developed in the framework of MyHeart project for prevention and disease management of cardio-vascular disease. 16 different concepts were tested, four of them were selected on the basis of user acceptance, technical feasibility and foreseen impact in their studies [18]. Hsiao et al. handled the environmental and outpatient records within Taichung Area for risk analysis of four specific categories of cardiovascular diseases, using deep learning approach. Autoencoder and Softmax were employed for feature extraction and classification. The output of Softmax for each sample is interpreted as the risk of these four specific categories of cardiovascular diseases [19]. Khare and Gupta applied association rule mining in order to explore the hidden relationships, which can help in understanding diseases and their causes in a better way, among data attributes for detecting of heart diseases [20]. Sabab et al. aimed the optimized cardiovascular disease prognosis by utilizing Support Vector Machine, Decision Tree and Naïve Bayes classification algorithms. They also introduced an efficient feature selection algorithm in order to improve the accuracy of proposed classifier models by reducing some lower ranked attributes. The dataset which contains total 14 attributes was collected from Department of Computing of Goldsmiths University of London [21].

It is seen that different methods and techniques were applied for this disease in the medical literature. In this context, the successes of the LR and RF algorithms, which are highly successful in knowledge discovery and classifying, have been discussed on two datasets within the frame of 5-fold cross-validation technique for this disease. The causes of poor success have been addressed. As it is well known in the literature, this study shows that there are many factors affecting this disease, not getting under weigh only biochemistry and laboratory test results.

The rest of this paper was organized as follows: In Section 2, the proposed methodology was given. In Section 3, experimental results were given in detail. Finally, in Section 4, conclusions and discussions were presented.

2. METHODOLOGY

2.1. Data Normalization

The quantity and quality of the data are very important for the prediction accuracy of any classification algorithm. Generally collected as a result of the patient care activity, the clinical datasets may contain redundant, incomplete, imprecise and inconsistent data [22]. Normalization involves scaling all values for a given attribute. So, all these values fall within a small specified range [23]. It can be said that normalization is a pre-process applied to perform the learning process more efficiently. Therefore, the data must be normalized before the analysis begins. Original data are transformed into a value between the minimum and maximum, i.e., zero and one with min-max normalization given in Equation 1.

$$x' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

In this equality [24]:

x' : the normalized value.

x_i : the processed value.

$x_{\min}$: the minimum value for variable x,

$x_{\max}$: the maximum value for variable x.

2.2. Classification and Learning

The classification based on the mathematical models is an important phase in the decision-making process and is used in many disciplines in order to find unknown patterns from the data. It consists of two steps. First, data which will be classified is obtained and prepared. Second, decision making process is applied to this data within the framework of the classification rules. Classification algorithms which are used commonly in many studies and benefited in this study are as follows:

Logistic Regression (LR): Being a mathematical modelling approach which describes the correlation between the attributes and the outcome variable, LR is a classification and an assignment tool. It is the most popular modelling procedure for analyzing epidemiologic data in which the outcome variable is categorical, i.e., true or false, so far [25].

Random Forest (RF): An ensemble learning method, RF algorithm is a very popular and effective machine learning technique introduced by Breiman [26] and this method is designed with the great numbers of decision tree. After the sub-datasets are generated from the original dataset, the trees are developed using the random selection feature [27].

2.3. Accuracy Analysis

The accuracy analysis is a process to determine the accuracy of assigned classes, utilizing classification algorithms. The successes of these algorithms are evaluated by comparing experimental results and precise results within the frame of the sensitivity (SE), the specificity (SP) and the Correct Classification Ratio (CCR) metrics which are given in Equation 2, 3 and 4 respectively.

$$\mathrm{SE} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN}) \tag{2}$$

$$\mathrm{SP} = \mathrm{TN} / (\mathrm{TN} + \mathrm{FP}) \tag{3}$$

$$\mathrm{CCR} = (\mathrm{TP{+}TN}) / (\mathrm{TP{+}FP{+}TN{+}FN}) \tag{4}$$

where, TP is the number of actual disease found as the disease, TN is the number of actual normal found as normal, FP is the number of actual normal found as disease, FN is the number of actual disease found as normal. The SE is the ratio of the number of actual correct positives in total positives. The SP is the ratio of the number of actual correct negatives in total negatives. The most commonly used in measurement of model success, the CCR can be expressed as the ratio of the number of accurately diagnosed samples to the number of total samples [28]. If these values are not at an ideal level, the improvements are performed by examining each process in the workflow.

3. DATA AND EXPERIMENTAL RESULTS

3.1. Dataset

Two datasets were used in order to investigate the effects of attributes in datasets on the CD. First one of these datasets is publicly available CD dataset named Cleveland Clinic

Foundation provided by the University of California, Irvine (UCI) Machine Learning Repository. There are 293 records in total. While the dataset has 76 raw attributes, only 14 of them were actually used. The rows, which have unknown values (?) in dataset were not processed. These attributes used was presented in Table 1.

Table 1. The information about attributes in the dataset no. 1 [31]

Attribute	Explanation
age	age in years
sex	sex (1 = male; 0 = female)
cp	chest pain type: Value 1: typical angina . Value 2: atypical angina Value 3: non-anginal pain, Value 4: asymptomatic
trestbps	resting blood pressure
chol	serum cholestoral in mg/dl
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg	resting electrocardiographic results
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment
ca	number of major vessels (0-3) colored by flourosopy
thal	3 = normal; 6 = fixed defect; 7 = reversable defect
num	diagnosis of heart disease (angiographic disease status) Value 0: < 50% diameter narrowing ,Value 1: > 50% diameter narrowing

The effects of biochemistry and hemogram laboratory test results such as cholesterol, HDL and LDL cholesterol for CD were analyzed by using the second dataset which was taken from the cardiology and other services of Yildirim Beyazit University Ankara Ataturk Training and Research Hospital (YBUAATR) between the dates 01.01.2011-11.10.2011 and 3774 records in total. This laboratory test information was given in Table 2. ICD-10 (International Statistical Classification of Diseases and Related Health Problems) booklet is the public coding sequence published by the World Health Organization for disease and health problems. ICD-10 was taken into consideration as a reference for the patient and control groups in this study.

Table 2. The information about biochemistry and hemogram laboratory test variables in the dataset no. 2

Test variable	**Data type**	**Description**
LDL	Continuous	LDL cholesterol value
HDL	Continuous	HDL cholesterol value
ALT	Continuous	Alanin Aminotransferaz value
AST	Continuous	Aspartat Transaminaz value
GLUCOSE	Continuous	Glucose value
HCT	Continuous	Hematocrit
HGB	Continuous	Hemoglobin
CREATININE	Continuous	Creatinine
LYM	Continuous	Lymphocytes value
MCH	Continuous	Mean Corpuscular Hemoglobin value
MCHC	Continuous	Mean Corpuscular Hemoglobin Concentration value
MCV	Continuous	Mean Corpuscular Volume value
MPV	Continuous	Mean Platelet Volume value
PCT	Continuous	Procalcitonin value
PDW	Continuous	Platelet Distrubition Width value
PLT	Continuous	Platelets
TRIGLYCERIDE	Continuous	Triglyceride value
CHOLESTEROL	Continuous	Cholesterol value
HYPERTENSION	Categorical	The situation of hypertension
HYPERLIPIDEMIA	Categorical	The situation of hyperlipidemia
DIABETES_MELLITUS	Categorical	The situation of diabetes mellitus

SMOKING	Categorical	The situation of smoking
FAMILY_BACKGROUND	Categorical	The situation of family background
DIAGNOSIS_OF_DISEASE	**Categorical**	**The situation of CAD**

3.2. Design Process and Experimental Results

The main objective of this research is to investigate the causes of CD, using both dataset 1 and dataset 2. Figure 1 describes the flow chart of the proposed method.

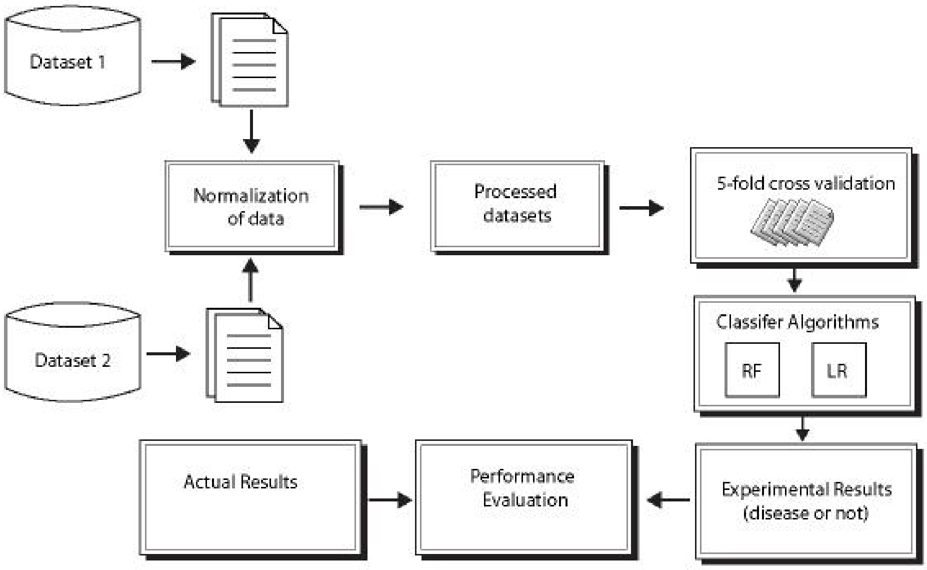

Fig.1. A general block diagram of the proposed method

As shown in this diagram, firstly, all attributes in both datasets were normalized into a range from 0 to 1. Next, both datasets were randomly divided into 80-20 % split of training and test set in accordance with k-fold cross validation technique with k=5 for experimental results. In other words, the algorithm iterates 5 times [29]. Hence, 234 and 3019 training records and 59 and 755 test records were obtained from dataset 1 and dataset 2 respectively in each iteration for the learning process. Following these processes, data classification was carried out on these sub-datasets utilizing RF and LR classification techniques in order to predict whether a person has a disease or not. The prediction results and the performance evaluations were presented in Table 3.

Table 3. Classification evaluation results for all models (with 5-fold cross-validation)

Datasets	**Classifier Algorithms**					
UCI Dataset	**LR**			**RF**		
Fold 1		**0**	**1**		**0**	**1**
	0	21	6	**0**	17	10
	1	15	17	**1**	12	20
	CCR: 64.41 %			CCR: 62.71 %		
	SE: 53.13 %			SE: 62.50 %		
	SP: 77.77 %			SP: 62.96 %		
Fold 2		**0**	**1**		**0**	**1**
	0	18	13	**0**	17	14
	1	12	16	**1**	10	18
	CCR: 57.63 %			CCR: 59.32 %		
	SE: 57.14 %			SE: 64.29 %		
	SP: 58.06 %			SP: 54.84 %		
Fold 3		**0**	**1**		**0**	**1**
	0	18	8	**0**	12	14
	1	22	11	**1**	18	15
	CCR: 49.15 %			CCR: 45.76 %		
	SE: 33.33 %			SE: 45.45 %		
	SP: 69.23 %			SP: 46.15 %		
Fold 4		**0**	**1**		**0**	**1**
	0	13	23	**0**	13	23
	1	7	16	**1**	6	17
	CCR: 49.15 %			CCR: 50.85 %		
	SE: 69.57 %			SE: 73.91 %		
	SP: 36.11 %			SP: 36.11 %		
Fold 5		**0**	**1**		**0**	**1**
	0	19	16	**0**	19	16
	1	14	10	**1**	12	12
	CCR: 49.15 %			CCR: 52.54 %		
	SE: 41.66 %			SE: 50.00 %		
	SP: 54.29 %			SP: 54.29 %		
	Average CCR: 53.90 %			**Average CCR: 54.24 %**		
YBUAATR Dataset						

Fold 1		**0**	**1**		**0**	**1**
	0	415	47	**0**	383	79
	1	240	53	**1**	195	98
	CCR: 61.99 %			CCR: 63.70 %		
	SE: 18.09 %			SE: 33.45 %		
	SP: 89.83 %			SP: 82.90 %		
Fold 2		**0**	**1**		**0**	**1**
	0	402	55	**0**	390	67
	1	244	54	**1**	195	103
	CCR: 60.40 %			CCR: 65.30 %		
	SE: 18.12 %			SE: 34.56 %		
	SP: 87.96 %			SP: 85.33 %		
Fold 3		**0**	**1**		**0**	**1**
	0	404	46	**0**	370	80
	1	258	47	**1**	195	110
	CCR: 59.74 %			CCR: 63.58 %		
	SE: 15.41 %			SE: 36.07 %		
	SP: 89.77 %			SP: 82.22 %		
Fold 4		**0**	**1**		**0**	**1**
	0	389	60	**0**	367	82
	1	243	63	**1**	203	103
	CCR: 59.87 %			CCR: 62.25 %		
	SE: 20.59 %			SE: 33.66 %		
	SP: 86.64 %			SP: 81.74 %		
Fold 5		**0**	**1**		**0**	**1**
	0	422	45	**0**	391	76
	1	232	56	**1**	195	93
	CCR: 63.31 %			CCR: 64.11 %		
	SE: 19.44 %			SE: 32.29 %		
	SP: 90.36 %			SP: 83.73 %		
	Average CCR: 61.06 %			**Average CCR: 63.79 %**		

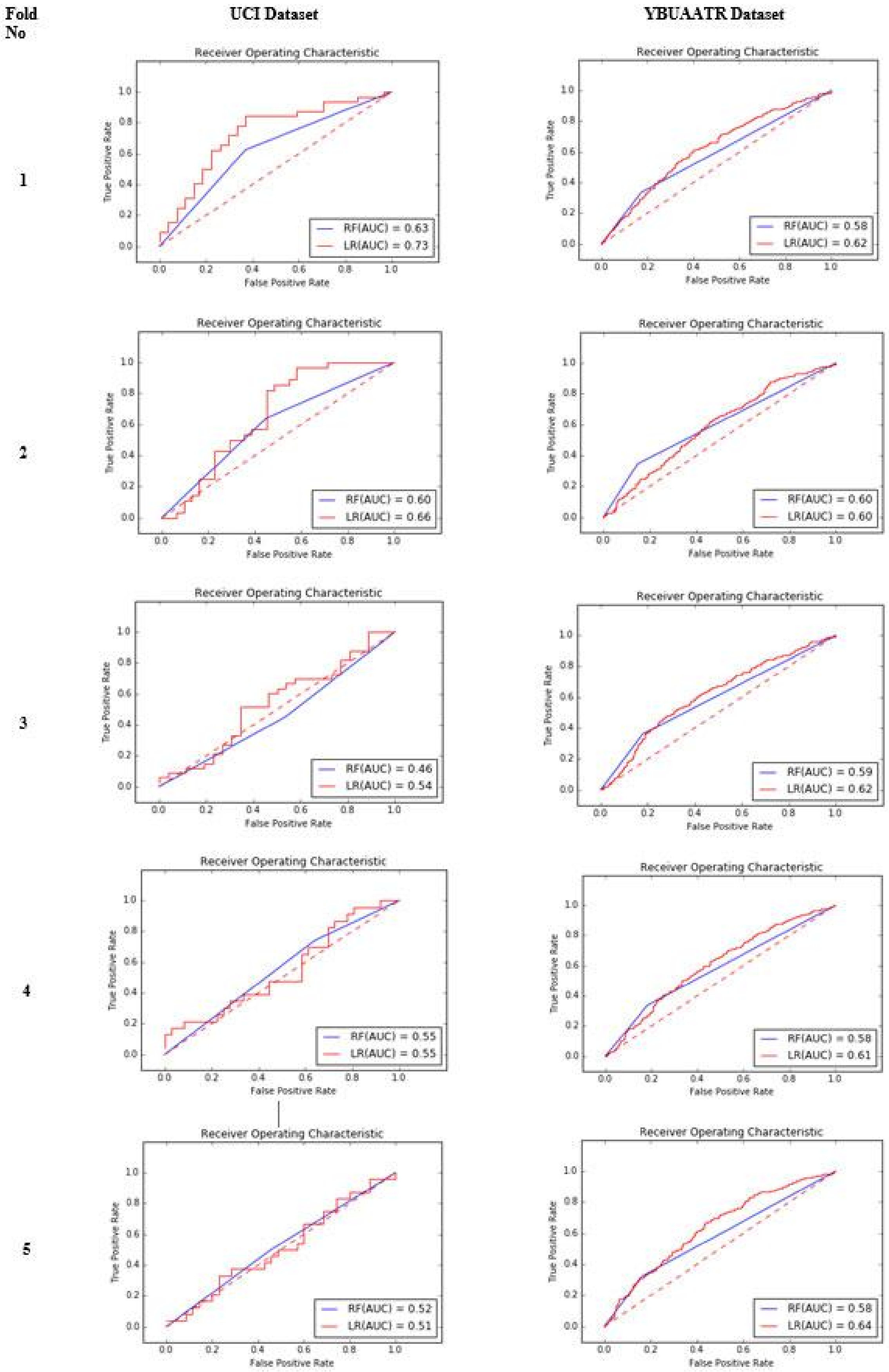

Fig.2. The ROC curve results for all models.

Since the output variable had two categorical values, the confusion matrix was 2x2 square matrix structure. The rows and columns represent the expected and observed values respectively. The overall accuracy of each model was presented at the bottom of the last columns of related matrix. For example, the correct classification ratio performed with LR is 64.41 % on fold 1. The SE and SP values were also available in the relevant tables. The ROC curve results for all models were given in Figure 2.

4. CONCLUSION AND DISCUSSION

The prediction of the causes of CD is a challenging process in health care systems because many known and unknown factors can cause the disease. In this study, the causes of CD were researched only using the datasets based on machine learning. Accuracy metrics showed that the successes for both datasets are around 60 %. As stated in [30-31], it is worth noting: many unknown, hidden attributes should be included in the dataset. Lifestyle factors may differ significantly between communities. Therefore, their influences on the risk may not accurately reflect the importance of these factors in different populations [31]. As seen in this study, it should be abstained from limited data and attribute space for the analysis of CD. All other information such as socio-demographic status, family background, smoking status of the patient should be regarded in the design. In addition, the patient-specific situations and the opinions of field specialists should be included in the decision support system.

5. ACKNOWLEDGEMENTS

The authors are thankful to the University of California, Irvine Machine Learning Repository and Yildirim Beyazit University Ankara Ataturk Training and Research Hospital.

6. REFERENCES

[1] Dorn BR, Dunn WA, Progulske-Fox A. Invasion of Human Coronary Artery Cells by Periodontal Pathogens. McGhee JR, ed. Infection and Immunity. 1999, 67(11), 5792-5798

[2] Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. Am J Cardiol. 1976, 38(1), 46-51.

[3] Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the Prospective Cardiovascular Münster (PROCAM) study. Circulation. 2002, 105(3), 310-315.

[4] Onat A, Keleş İ, Çetinkaya A, Başar Ö, Yıldırım B, Erer B, Ceyhan K, Eryonucu B, Sansoy V. Prevalence of Coronary Mortality and Morbidity in the Turkish Adult Risk Factor Study: 10-year Follow-up Suggests Coronary Epidemic. Türk Kardiyol Arş. 2001, 29(1), 8-19

[5] Chambless LE, Dobson AJ, Patterson CC, Raines B. On the use of a logistic risk score in predicting risk of coronary heart disease. Stat Med 9. 1990, 385-396

[6] Gilstrap LG, Wang TJ. Biomarkers and cardiovascular risk assessment for primary prevention: An update. Clin Chem, 2012, 58, 72-82, doi:10.1373/clinchem.2011.165712

[7] Battes L, Akkerhuis M, Van boven N, Boersma E, Kardys I. Cardiovascular risk prediction models in patients with stable coronary artery disease. Exp Clin Cardiol. 2014, 20, 117-129

[8] Wang Z, Hoy WE. Is the Framingham coronary heart disease absolute risk function applicable to Aboriginal people? Med.J.Australia. 2005, 182(2), 66-69

[9] Minas AK, Joseph AM, Demetra H, Constantinos SP. Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees, IEEE Transactions On Information Technology In Biomedicine, 2010, 14(3), 559-566, doi:10.1109/TITB.2009.2038906

[10] Srinivas K, Rao GR, Govardhan A. Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques. 5th IntConf on Computer Science and Education. 2010, 1344-1349, doi:10.1109/ICCSE.2010.5593711

[11] Abdullah AS, Rajalaxmi RR. A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier. International Conference on Recent Trends in Computational Methods, Communication and Controls (ICON3C 2012). 2012, 22-25.

[12] El-Bialya R, Salamay MA, Karam OH, Khalifa ME. Feature Analysis of Coronary Artery Heart Disease Data Sets, Procedia Computer Science. 2015, 65, 459-468, doi:10.1016/j.procs.2015.09.132

[13] Akhil jabbar M, Deekshatulu BL and Chandra P. Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA). 2013, 10, 85-94, doi:10.1016/j.protcy.2013.12.340

[14] Shu L, Arneson D, Yang X. Bioinformatics Principles for Deciphering Cardiovascular Diseases, Reference Module in Biomedical Sciences. 2017, doi:10.1016/B978-0-12-809657-4.99576-0

[15] Ave A, Fauzan H, Adhitya SR, Zakaria H. Early detection of cardiovascular disease with photoplethysmogram (PPG) sensor, 2015 International Conference on Electrical Engineering and Informatics, Denpasar, Indonesia, 10-11 Aug. 2015, doi:10.1109/ICEEI.2015.7352584

[16] Thakur S, Ramzan M. A systematic review on cardiovascular diseases using big-data by Hadoop, 6th International Conference Cloud System and Big Data Engineering, Noida, India, 14-15 Jan. 2016, doi:10.1109/CONFLUENCE.2016.7508142

[17] Singh M, Martins LM, Joanis P, Mago VK. Building a Cardiovascular Disease predictive model using Structural Equation Model & Fuzzy Cognitive Map, 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Vancouver, BC, Canada, 24-29 July 2016, doi:10.1109/FUZZ-IEEE.2016.7737850

[18] Guillen SG, Sala P, Habetha J, Schmidt R, Arredondo MT. Innovative Concepts for Prevention and Disease Management of Cardiovascular Diseases, Engineering in Medicine and Biology Society, 28th Annual International Conference on Engineering in Medicine and Biology Society, New York, USA, 30 Aug.-3 Sept. 2006, doi:10.1109/IEMBS.2006.259449

[19] Hsiao HCW, Chen SHF, Tsai JJP. Deep Learning for Risk Analysis of Specific Cardiovascular Diseases Using Environmental Data and Outpatient Records, 2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE), Taichung, Taiwan, 31 Oct.-2 Nov. 2016, doi:10.1109/BIBE.2016.75

[20] Khare S and Gupta D. Association rule analysis in cardiovascular disease, 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP), Mysore, India, 12-13 Aug. 2016, doi:10.1109/CCIP.2016.7802881

[21] Sabab SA, Md. Munshi AR, Pritom AI, Shihabuzzaman. Cardiovascular disease prognosis using effective classification and feature selection technique, 2016 International Conference on Medical Engineering, Health Informatics and Technology, Dhaka, Bangladesh, 17-18 Dec. 2016, doi:10.1109/MEDITEC.2016.7835374

[22] Madhavi P, Bamnote GR. Efficient Binary Classifier for Prediction of Diabetes Using Data Preprocessing and Support Vector Machine, Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014: 327 of the series Advances in Intelligent Systems and Computing: 131-140.

[23] Han J, Kamber M. Data Mining concepts and techniques, 2nd ed. Morgan Kaumann publication, An imprint of Elsevier, 2006, pp. 61-77.

[24] Jain YK, Bhandare SK. Min Max Normalization Based Data Perturbation Method for Privacy Protection, International Journal of Computer & communication Technology, 2011, 2(8), 45-50.

[25] Kleinbaum DG and Klein M. Logictic Regression A Self-Learning Text (3rd ed). New York, NY: Springer, 2010, pp. 4-5.

[26] Breiman L. Random forests. Mach Learn 2001, 45, 5-32

[27] Aggarwal CC. Data classification algorithms and applications. In: Kumar V, editor. Data Mining and Knowledge Discovery Series. Boca Raton, USA: CRC Press 2014, pp. 23-34.

[28] Baratloo A., Hosseini M., Negida A. and Ashal G.E., Part 1: Simple Definition and Calculation of Accuracy. Sensitivity and Specificity, Emerg (Tehran), 2015, 3(2), 48-49

[29] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of the 14th international joint conference on Artificial intelligence, 1995, 2, 1137-1143

[30] Gray D. Risk assessment gone mad: The rise of risk evaluation and mass public deception. Br J Cardiol 2009, 16, 117-118

[31] Kurian AK, Cardarelli KM. Racial and ethnic differences in cardiovascular disease risk factors: A systematic review. Ethn Dis 2007, 17, 143-152.

How to cite this article:
Akyol K, Şen B. A Study on Analysis of Cardiovascular Diseases. J. Fundam. Appl. Sci., 2018, *10(2), 277-292.*