Diagnosis of Coronary Heart Disease via Classification Algorithms and a New Feature Selection Methodology

Burak Kolukisa[1*], Hilal Hacilar[1], Gokhan Goy[1], Mustafa Kus[2], Burcu Bakir-Gungor[1], Atilla Aral[3], Vehbi Cagri Gungor[1]

[1]Department of Computer Engineering, Abdullah Gül University, Kayseri, Turkey

[2]Keydata Bilgi İşlem Teknoloji Sistemleri A.Ş., Ankara

[3]Department of Cardiovascular Surgery, Ankara University, School of Medicine, Ankara, Turkey

Emails: burak.kolukisa@agu.edu.tr, hilal.hacilar@agu.edu.tr, gokhan.goy@agu.edu.tr, mustafa.kus@keydata.com.tr,burcu.gungor@agu.edu.tr, aral@medicine.ankara.edu.tr, cagri.gungor@agu.edu.tr

***Abstract*— According to the World Health Organization (WHO), 31% of the world's total deaths in 2016 (17.9 million) was due to cardiovascular diseases (CVD). With the development of information technologies, it has became possible to predict whether people have heart diseases or not by checking certain physical and biochemical values at a lower cost. In this study, we have evaluated a set of different classification algorithms, linear discriminant analysis and proposed a new hybrid feature selection methodology for the diagnosis of coronary heart diseases (CHD). One of the advantages of the proposed method is its ability to work on real-time datasets. Throughout this research effort, we have tested the performance of our method using publicly available heart disease datasets (UCI Machine Learning Repository, Z-Alizadehsani). We have conducted comparative performance evaluations in terms of accuracy, sensitivity, specificity, F-measure, AUC and running time.**

Keywords- Cardiovascular Disease (CVD), Coronary Artery Disease (CAD), Data Mining, Machine Learning, Linear Discriminant Analysis, Feature Selection, Ensemble Methods, Classification

I. INTRODUCTION

According to the World Health Organization (WHO), 31% of the world's total deaths in 2016 (17.9 million) was due to cardiovascular diseases (CVD). WHO also predicts that the deaths due to CVDs will reach approximately 30 million in 2030 [1]. Coronary artery disease (CAD) or coronary heart disease (CHD) is one type of cardiovascular diseases in which the presence of atherosclerotic plaques in coronary arteries can lead to myocardial infarction or sudden cardiac death. Several tests such as echocardiogram (Echo), nuclear scan, electrocardiogram (ECG), angiography and exercise stress testing are widely used by medical doctors to diagnose heart diseases. ECG is a noninvasive method that is used to diagnose CAD, but it could lead to the undiagnosed condition of CAD. Angiography is a golden standard technique to diagnose heart diseases. But it requires expertise, it has the risks for the patient, it is costly and it is a consuming method [2]. Machine learning approaches make it possible to predict the risk for developing heart disease by checking certain values at a low cost. In this regard, researchers studied with different classifiers and feature selection methods on different heart disease datasets such as Cleveland dataset at UCI Machine Learning Repository, Z-Alizadehsani dataset [3-16]. In our previous study, we compared these existing studies in terms of classification methodologies, feature selection techniques, pre-processing; and also in terms of sensitivity, specificity, F-measure, accuracy, and Area Under Curve (AUC), as a performance measures [17]. A summary of this comparison can be found in Table I. Although all these existing studies provide valuable insights and foundations about CAD diagnosis, there is no internationally accepted standard approach for the CAD diagnosis. In addition, none of them presents a detailed performance evaluation of different classification methods and feature selection algorithms in terms of specificity, sensitivity, accuracy, AUC, F-measure and running time. The aim of this paper is to fulfill this gap and show that not only accuracy measure is important, but also other performance metrics, such as specificity, sensitivity, AUC and F-Measure are also critical in terms of reliable and diagnosis of CAD. Running time of a selected algorithm is also critical because if such a system is intended to be used for in intensive care units, a fast decision needs to be made.

In this study, we have evaluated different classification algorithms and linear discriminant analysis in terms of all evaluation criterias that we mentioned and we proposed a new hybrid feature selection methodology for the diagnosis of CAD. It is important to note that many of the algorithms are not stable when different data sets are used and their performances differ significantly when the number of samples changes. The application of feature selection methods resulted in good performance measures. To test our proposed methodology, we used publicly available resources, i.e., UCI Machine Learning Repository and Z-Alizadehsani Dataset. UCI Machine Learning Repository contains Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog CVD datasets. Since the Cleveland dataset has no missing values, we have analyzed the Cleveland dataset separately. Also, we have assembled all UCI datasets together and reanalyzed.

This work is organized in the following sections: Section II introduces publicly available CAD datasets, the proposed hybrid feature selection method and feature dimension reduction via linear discriminant analysis (LDA). Section III represents performance evaluations of different classification algorithms. Lastly, Section IV concludes the study.

TABLE 1. COMPARISION OF DIFFERENT CLASSIFICATION METHODS FOR CAD DIAGNOSIS

Reference	Method	FS	PP	KF	SN	SP	FM	AUC	ACC	TM	Dataset	Year
Kemal Polat et al [4]	Fuzzy-AIRS-KNN	No	-	15	92.30%	92.30%	-	-	87%	-	Cleveland	2007
My Chau Tu et al [5]	Bagging	No	Yes	10	74.93%	86.64%	-	-	81.41%	-	UCI	2009
Resul Das et al [6]	ANN Ensemble	No	Yes	-	80.95%	95.91%	-	-	89.01%	-	Cleveland	2009
Shouman et al [7]	Decision Tree	-	Yes	10	77.90%	85.20%	-	-	84.1%	-	Cleveland	2011
Alizadehsani et al [3]	SMO	Yes	Yes	-	97.22%	79.31%	-	-	92.09%	-	Z-Alizadehsani	2012
Karabulut et al [8]	ANN	-	-	10	95.6%	86.75	-	0.915	91.2%	-	UCI	2012
Shouman et al [9]	KNN	No	Yes	10	93.8%	99.5%	-	-	97.4%	-	UCI	2012
Nahar et al [10]	AR	No	No	-	-	-	-	-	99.38%	Yes	UCI	2013
Rajalaxmi et al [11]	BABC -Naïve Bayesian	Yes	Yes	-	-	-	-	-	86.4%	-	Cleveland	2014
Chetna Yadav et al [12]	TRM	Yes	No	-	96.65%	91.53%	-	-	93.75%	-	Z-Alizadehsani	2015
Randa El-Biary et al [13]	C4.5 Decision Tree	Yes	Yes	10	-	-	-	-	78.54%	Yes	Cleveland	2015
Luxmi Verma et al [14]	MLR	Yes	Yes	-	-	-	-	-	90.28%	-	Cleveland	2016
Frantisek Babic et al [15]	Decision Tree	Yes	Yes	-	-	-	-	-	73.87%	-	South Africa	2017
Frantisek Babic et al [15]	SVM	Yes	Yes	-	-	-	-	-	86.67%	-	Z-Alizadehsani	2017
Samuel et al [16]	ANN	Yes	Yes	-	%100	84.0%	-	-	91.1%	-	UCI	2017

FS: Feature Selection, PP: Pre-Processing, KF: K-Fold, SN: Sensitivity, SP: Specificity, FM: F-Measure, ACC: Accuracy, TM: Time

II. MATERIALS AND METHODS

A. PERFORMANCE METRICS AND DATASETS

Most of the existing studies aim to improve the prediction accuracy of CAD diagnosis. Accuracy is an important measure especially when we have a symmetric dataset where the number of false positive and false negative values are almost the same. For the CAD diagnosis problem, accuracy may not be enough to determine the performance of a classifier. For example, assume that false positive and false negative values are equal and we have more than two class labels, such as low, medium, high. In such a case, accuracy will not be enough to classify different cases. Hence, to overcome the disadvantages of accuracy, we focused on several measures, such as accuracy, sensitivity, specificity, F-measure, in addition to accuracy.

In this study, we have worked on Cleveland, UCI (Mix) and Z-Alizadehsani datasets. Cleveland dataset was the first proposed by Detrano [18] and it is widely used in the literature. It contains 303 data samples, in which only six samples have missing values. Each sample is categorized into one of the two following groups. If the vessels of a subject are narrowed less than 50%, these subjects are labeled as healthy if the vessels of a subject are narrowed more than 50%, these subject are labeled as sick. The Z-Alizadehsani dataset has been collected at Terhan's Shaeheed Rajaei Cardiovascular, Medical and Research Center. It contains 303 data samples and 55 attributes, in which attributes are divided into the following 4 groups, i.e. Demographic, Symptom and Examination, ECG, Laboratory and Echo as shown in Table 3. We created UCI (Mix) dataset via combining Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog CVD dataset available at UCI Machine Learning Repository, Figure 1 summarizes these datasets in terms of their sample sizes. All of these datasets have the same 14 attributes and in UCI (Mix) dataset, we have included the samples that are complete or missing only one feature. Missing features have 14 attributes and the samples missing only one feature. In this dataset, we replaced the missing feature of a sample with an average of this feature in other samples. UCI (Mix) contains 371 samples and 14 attributes. For each dataset, the number of attributes, the number of healthy (NOR) subjects and the number of sick (CAD) samples are shown in Table 2.

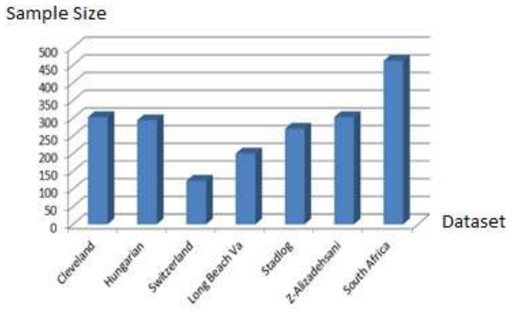


Fig. 1. Publicly available CVD dataset

TABLE 2. FEATURES OF PUBLICLY AVAILABLE HEART DISEASE DATASETS

	Attribute	CAD	NOR	Total
Z-Alizadehsani	55	216	87	303
Cleveland	14	165	138	303
UCI (Mix)	14	199	172	371

Tables 3 and 4 show the details of the features included in Z-Alizadehsani and Cleveland datasets, respectively.

TABLE 3. Z-ALIZADEHSANI DATASET DESCRIPTION

No	FT*	Attribute - Description	Values
1	Demographic	Age	30-86
2		Weight	48-120
3		Length	140-188
4		Sex	M,F
5		BMI (Body Mass Index Kg/m^2)	18-41
6		DM (Diabetes Mellitus)	Yes, no
7		HTN (Hyper Tension)	Yes, no
8		Current Smoker	Yes, no
9		Ex-Smoker	Yes, no
10		FH (Family History)	Yes, no
11		Obesity (MBI > 25)	Yes, no
12		CRF (Chronic Renal Failure)	Yes, no
13		CVA (Cerebrovascular Accident)	Yes, no
14		Airway Disease	Yes, no
15		Thyroid Disease	Yes, no
16		CHF (Congestive Heart Failure)	Yes, no
17		DLP (Dyslipidemia)	Yes, no
18	Symptom and examination	BP (Blood Pressure mmHg)	90 – 190
19		PR (Pulse Rate ppm)	50-110
20		Edema	Yes, No
21		Weak Peripheral Pulse	Yes, No
22		Lung Rales	Yes, no
23		Systolic Murmur	Yes, no
24		Diastolic Murmur	Yes, no
25		Typical Chest Pain	Yes, no
26		Dyspnea	Yes, no
27		Function Class	1,2,3,4
28		Atypical	Yes, no
29		Nonanginal CP	Yes, no
30		Exertional CP (Exertional Chest Pain)	Yes, no
31		Low Th Ang (Low Threshold Angina)	Yes, no
32	ECG	Q Wave	Yes, no
33		ST Elevation	Yes, no
34		ST Depression	Yes, no
35		T inversion	Yes, no
36		LVH (Left Ventricular Hypertrophy)	Yes, no
37		Poor R progression (poor R wave progression)	Yes, no
38		BBB	-
39	Laboratory and echo	FBS (Fasting Blood Sugar in mg/dl)	62–400
40		Cr (Creatine in mg/dl)	0.5–2.2
41		TG (Triglyceride in mg/dl)	37–1050
42		LDL (Low Density Lipoprotein in mg/dl)	18-232
43		HDL (High Density Lipoprotein in mg/dl)	15 -111
44		BUN (Blood Urea Nitrogen in mg/dl)	6–52
45		ESR (Erythrocyte Sedimentation Rate in mm/h)	1–90
46		HB (Hemoglobin in g/dl)	8.9–17.6
47		K (Potassium in mEq/lit)	3.0–6.6
48		Na (Sodium in mEq/lit)	128–156
49		WBC (White Blood Cell in cells/ml)	3700–18,000
50		Lymph (Lymphocyte in %)	7–60
51		Neut (Neutrophil in %)	32–89
52		PLT (Platelet in 1000/ml)	25–742
53		EF-TTE (Ejection Fraction in %)	15–60
54		Region RWMA (Regional Wall Motion Abnormality)	0,1,2,3,4
55		VHD (Valvular Heart Disease)	1-4

*FT: Feature Type

B. *CLASSIFICATION METHODS*

Classification is a supervised learning method. Classification algorithms firstly learn via analyzing the labels of samples and their nominal and/or numeric values as attributes and hence they create a model. Secondly, they make predictions this generated models. In this study, we have tested different classification algorithms on different datasets.

TABLE 4. CLEVELAND DATASET DESCRIPTION

No	Attribute - Description	Value
1	Age	29 - 77
2	Sex	M,F
3	CP (Typical, Atypical, Non-Anginal Pain, Asymptomatic)	1,2,3,4
4	Trestbps (Resting Blood Pressure)	94 - 200
5	Chol (Serum Cholestoral in mg/dl)	126 - 564
6	Fbs (Fasting Blood Sugar > 120)	Yes, No
7	Rectecg (Resting Electrocardiographic)	0,1,2
8	Thalach (Maximum Heart Rate Achieved)	71 - 202
9	Exang (Exercise Induced Angina)	Yes, No
10	Oldpeak (ST Depression Induced by Exercise Relative to Rest)	0 – 6.2
11	Slope (The Slope of The Peak Exercise ST Segment)	1,2,3
12	Ca (Number of Major Vessels Colored by Flourosopy)	0,1,2,3
13	Thal (Normal, Fixed Defect, Reversible Defect)	3,6,7
14	Num (Diagnosis of Heart Disease)	Yes, No

via incorporating ensemble classification methods, we also attempted to strengthen inaccurate learning which is caused by noisy data.

Naive Bayes (NB) is a simple probabilistic classifier which is easy to apply and it performs can well on data sets with a high number of instances. Random Forest (RF) is a tree-based algorithm that is widely used in machine learning problems and it is also applicable to both classification and regression problems. k-Nearest Neighbor (kNN) algorithm is an unsupervised algorithm that can handle discrete and continuous attributes and can be beneficial as the first step of supervised learning. The support vector machine (SVM) algorithm is capable of efficiently process high-dimensional data. Extreme Gradient Boosting (XGBOOST) is a machine learning algorithm for regression and classification problems that make a prediction using the ensemble of weak decision trees. In XGBoost classifier, if an attribute is highly used to make key decisions with decision trees, high relative importance is assigned to this feature.

C. *CROSS-VALIDATION*

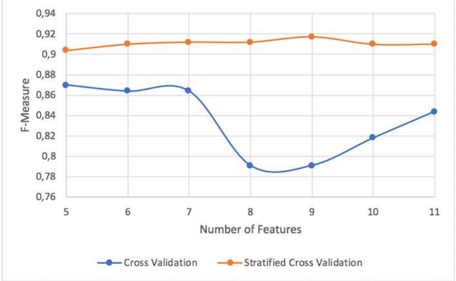


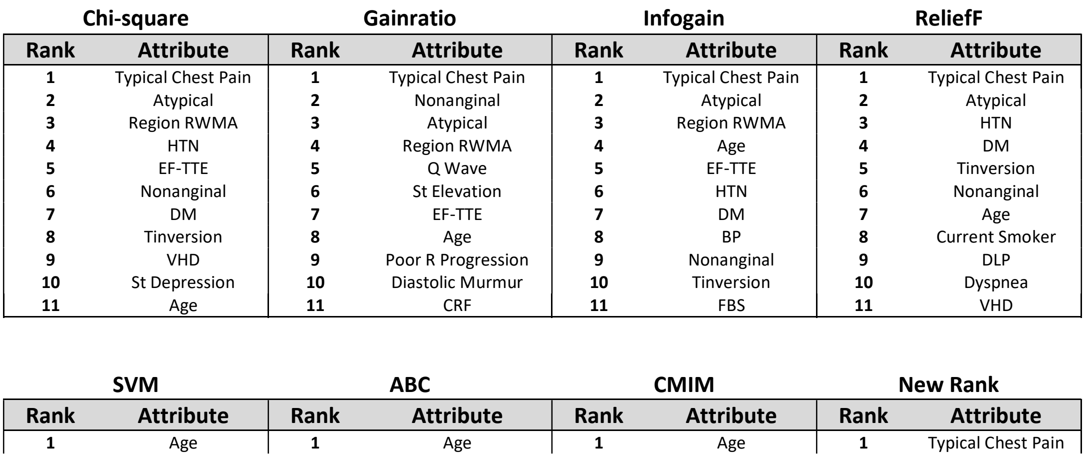
Fig. 2. Comparison between stratified cross-validation and normal cross-validation using SVM

The main goal of cross-validation is to prevent overfitting. In regular cross-validation, the proportions of the two types of classes distributed to the folds can be unbalanced. Therefore, the results of the classification algorithms in regular cross-validation may be incorrect. Whereas, in stratified cross-validation, each fold contains approximately the same proportions of samples from the two types of classes. When SVM classification algorithm is applied on Z-Alizadehsani dataset, Figure 2 emphasizes that higher levels of F-measures are obtained with stratified cross-validation in comparison to regular cross-validation for different numbers of features.

D. FEATURE SELECTION AND COMBINATION

The aim of the feature selection is to obtain a robust classification model via removing the features that are not related or less related to the class labels, or that have predictive power. In terms of diseases diagnosis, feature selection methodologies may help to reduce the time and the costs of the biological test. In our work, we used seven commonly used feature selection methods, i.e., information gain (IG), gain ratio (GR), relief-f (RF), chi-square (CS), SVM, artificial bee colony (ABC) and conditional mutual information maximization (CMIM). While information gain and gain ratio are filter-based feature selection methods, relief-F is an instance-based feature selection methods and chi-squared test as a statistical method. Information Gain is an entropy-based feature selection method, which works based on tree algorithms. It chooses the most meaningful features that are close to the root of the tree and it is usually stable. Gain Ratio is a modification of information gain that reduces its bias on highly branching features and tends to prefer unbalanced splits in which one partition is much smaller than the other. Chi-Square is a well known statistical test, which measures the variation between expected and observed values of samples and it decides whether each feature can represent the dataset. Relief-F selects top ranking features from the dataset by assigning different weights to each feature, compared to its neighbours. The support vector machine (SVM) algorithm is capable of efficiently process high-dimensional data. Artificial bee colony (ABC) is as simple as partifical swarm optimization and differential evolution algorithms. It requires only common control parameters such as colony size and maximum cycle number. Conditional Mutual Information Maximization (CMIM) feature selection method first ranks the features according to their conditional entropy and mutual information with the class to predict. Then it allows the addition of a new feature to the selected set of features if and only if the feature carries additional information. Using these seven different feature selection methods, we aimed to reduce overfitting and look at the data set from different points of view.

To perform these feature selection methods, we used scikit-learn library based on Python. In order to integrate the results of different feature selection methods, we proposed the following hybrid feature selection methodology. Firstly, we get seven different rankings of features by applying seven different feature selection techniques individually. Secondly, to calculate the final ranking of a feature, we take the average of seven rankings obtained via different feature selection methodologies. The rankings of features in different feature selection methods are shown in Figure 3 for Z-Alizadehsani dataset. Figure 4 shows the relationship between the number of features, running time and other performance measures when hybrid feature selection is applied and MLP is used as a classifier. We also compared the performances of the following two types of hybrid feature selection methodologies. While the first one integrates information gain, gain ratio, chi-square, and relief-f feature selection methodologies, the second one integrates all seven feature selection methodologies.



Chi-square		Gainratio		Infogain		ReliefF	
Rank	**Attribute**	**Rank**	**Attribute**	**Rank**	**Attribute**	**Rank**	**Attribute**
1	Typical Chest Pain	1	Typical Chest Pain	1	Typical Chest Pain	1	Typical Chest Pain
2	Atypical	2	Nonanginal	2	Atypical	2	Atypical
3	Region RWMA	3	Atypical	3	Region RWMA	3	HTN
4	HTN	4	Region RWMA	4	Age	4	DM
5	EF-TTE	5	Q Wave	5	EF-TTE	5	Tinversion
6	Nonanginal	6	St Elevation	6	HTN	6	Nonanginal
7	DM	7	EF-TTE	7	DM	7	Age
8	Tinversion	8	Age	8	BP	8	Current Smoker
9	VHD	9	Poor R Progression	9	Nonanginal	9	DLP
10	St Depression	10	Diastolic Murmur	10	Tinversion	10	Dyspnea
11	Age	11	CRF	11	FBS	11	VHD

SVM		ABC		CMIM		New Rank	
Rank	**Attribute**	**Rank**	**Attribute**	**Rank**	**Attribute**	**Rank**	**Attribute**
1	Age	1	Age	1	Age	1	Typical Chest Pain

2	Region RWMA	2	HTN	2	DM	2	Age
3	Typical Chest Pain	3	Typical Chest Pain	3	HTN	3	HTN
4	Tinversion	4	Q Wave	4	BP	4	Region RWMA
5	FBS	5	Tinversion	5	Typical Chest Pain	5	Tinversion
6	TG	6	FBS	6	Atypical	6	EF-TTE
7	PR	7	ESR	7	Nonanginal	7	Q Wave
8	St Elevation	8	K	8	Q Wave	8	Atypical
9	ESR	9	EF-TTE	9	Tinversion	9	ESR
10	HM	10	Region RWMA	10	ESR	10	K
11	Length	11	-	11	K	11	Nonanginal

Fig. 3. Hybrid feature selection generates a new ranking of attributes ranking via averaging the rankings of the attributes obtained in different feature selection algorithms on Z-Alizadehsani Dataset.

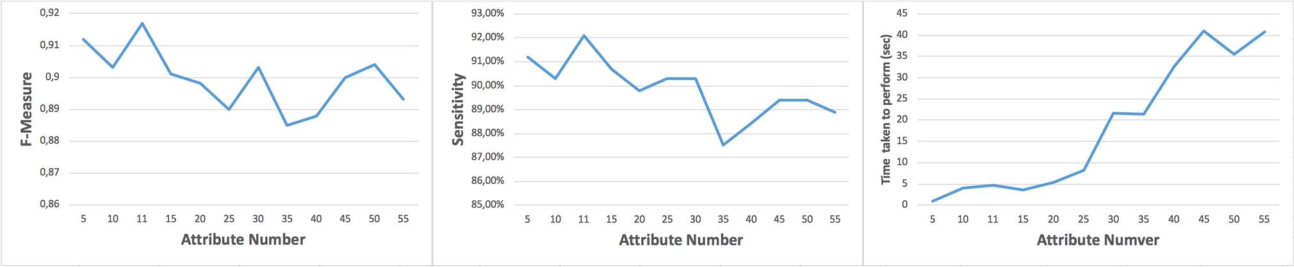


Fig. 4. Experimentation of MLP classification method with increasing number of features on Z-Alizadehsani Dataset for a first type of hybrid feature selection.

E. FEATURE SELECTION BASED ON MEDICAL DOCTORS RECOMMENDATIONS

In medical practice, a widely used method to diagnose cardiovascular diseases is to evaluate the patient in terms of the risk factors that are defined in the Framingham Heart Study (FHS). FHS is conducted by researchers from Boston University and sponsored by National Heart, Lung and Blood Institute (NHLBI). 5209 adult subjects from Framingham Massachusetts, USA participated in 1948 as the first generation of participants, and the study continues now with its third generation of participants. The participants have been observed to determine the primary factors that contribute to cardiovascular diseases. To detect a pattern between cardiovascular diseases and analyzed factors, the subjects have undergone extensive physical examinations and lifestyle interviews. This study is repeated in every 2 years. In 1971, 5124 people have been enrolled as the second generation. In 1994, different cohorts are added in order to diversify the sample. Furthermore, in 2002, the third generation is added to the study. As a result of these efforts, the study produced 1200 articles in well known medical journals in the last 50 years. Therefore, this study is accepted as a fundamental resource for cardiovascular diseases in clinical practice [19]. In 2008, the NHLBI adopted a risk calculator primarily based on FHS. This risk calculator also utilizes from other studies such as ATP III, PRO-CAM, QRISK, EURO-SCORE, and Reynolds. The calculator uses sex, age, total cholesterol, HDL cholesterol, untreated SBP, treated SBP, current smoker and diabetes factors in order to determine ten years cardiovascular disease risk score [20]. For the past two decades, it has been possible to estimate CHD risk by the use of regression equations derived from observational studies. Prediction models have typically been based on the logistic function, although the Weibull distribution has also been used. Formulations have often included age, sex, blood pressure, TC, HDL-C, smoking, diabetes, and left ventricular hypertrophy. The prediction of CHD has taken the form of sex-specific equations that were developed from a single study and applied to other populations or individuals. Age, TC, HDL-C, and blood pressure were used in the equations as continuous variables, in contrast to dichotomous variables (yes/no) such as smoking, diabetes, and left ventricular hypertrophy [21]. Therefore, in order to formulate a CVD or CAD diagnosis as a machine learning problem, one needs to make sure that these risk factors are included as attributes. According to the medical doctors' recommendation, we have separated the features into 2 groups, i.e., Framingham Heart Study (FHS) Risk factors and Clinically Important Findings (CIF). Tables 7 and 8 shows the details of the selected features based on medical doctors' recommendation in Cleveland and Z-Alizadehsani datasets respectively. Some of the features such as weight, BMI and obesity mean have similar meanings and hence, we have grouped them together.

TABLE 7. FEATURE SELECTION OF CLEVELAND DATASET BASED ON MEDICAL DOCTORS' RECOMMENDATION

No	Feature Type	Attribute
1	CIF	Cp
2	CIF	Exang
3	CIF	Oldpeak
4	CIF	Thal
5	FHS RF	Trestbps
6	FHS RF	Chol
7	FHS RF	Fbs
8	FHS RF	Age
9	FHS RF	Sex

TABLE 8. FEATURE SELECTION OF Z-ALIZEDEHSANI DATASET BASED ON MEDICAL DOCTORS' RECOMMENDATION

No	Feature Type	Attributes
1	CIF	Typcial Chest Pain
2	CIF	Exertional CP
3	CIF	Q Wave
4	CIF	Region with RWMA
5	FHS RF	Age
6	FHS RF	Sex
7	FHS RF	Weight, BMI, Obesity
8	FHS RF	DM, FBS
9	FHS RF	HTN, BP
10	FHS RF	Current Smoker, Ex-Smoker
11	FHS RF	FH
12	FHS RF	LDL
13	FHS RF	HDL

F. *DIMENSION REDUCTION USING LINEAR DISCRIMINANT ANALYSIS*

Dimension reduction techniques are widely used in the pre-processing step of several machine learning applications. The goal of dimension reduction is to project the dataset onto a lower-dimensional space with good class-separability in order to avoid overfitting and also to reduce computational costs. Linear Discriminant Analysis (LDA) is a commonly used dimension reduction technique and it is optimal in terms of maximizing the separation between several classes.

TABLE 9. RESULTS OF CLASSIFICATION ALGORITHMS

DT	*Method*	*FS*	*Number of FS Type*	*Feature Number*	*Sensitivity*	*Specificity*	*F-Measure*	*AUC*	*Accuracy*
Cleveland	*Bagging*	*No*	*-*	*14*	*87.3%*	*83.6%*	*0.847*	*0.891*	*82.83%*
	Bagging	*Yes*	*4*	*5*	*88.5%*	*76.8%*	*0.851*	*0.888*	*83.16%*
	Naïve Bayes	*No*	*-*	*14*	*86.7%*	*83.3%*	*0.851*	*0.904*	*83.49%*
	Naïve Bayes	*Yes*	*4*	*5*	*89.4%*	*86.6%*	*0.869*	*0.894*	*85.47%*
	Random Forest	*No*	*-*	*14*	*85.5%*	*82.2%*	*0.847*	*0.901*	*83.16%*
	Random Forest	*Yes*	*4*	*5*	*86.1%*	*82.7%*	*0.848*	*0.872*	*83.16%*
	KNN	*No*	*-*	*14*	*83%*	*80.3%*	*0.840*	*0.879*	*82.83%*
	KNN	*Yes*	*4*	*5*	*85.5%*	*73.2%*	*0.822*	*0.838*	*79.86%*
	MLP	*No*	*-*	*14*	*83.6%*	*80.6%*	*0.839*	*0.883*	*82.50%*
	MLP	*Yes*	*4*	*5*	*83.0%*	*81.9%*	*0.838*	*0.875*	*82.50%*
UCI Mix	*Logitboost*	*No*	*4*	*14*	*82.0%*	*79.6%*	*0.830*	*0.904*	*81.67%*
	Logitboost	*Yes*	*4*	*5*	*81.0%*	*79.0%*	*0.820*	*0.891*	*81.40%*
	Naïve Bayes	*No*	*4*	*14*	*84.0%*	*82.7%*	*0.860*	*0.898*	*85.44%*
	Naïve Bayes	*Yes*	*4*	*5*	*82.0%*	*80.1%*	*0.830*	*0.877*	*82.21%*
	Random Forest	*No*	*4*	*14*	*81.0%*	*78.6%*	*0.840*	*0.904*	*83.01%*
	Random Forest	*Yes*	*4*	*5*	*82.0%*	*80.0%*	*0.840*	*0.891*	*82.75%*
	XGBoost	*No*	*4*	*14*	*81.0%*	*78.6%*	*0.820*	*0.895*	*81.13%*
	XGBoost	*Yes*	*4*	*5*	*84.0%*	*81.0%*	*0.850*	*0.903*	*83.82%*
	SVM	*No*	*4*	*14*	*82.0%*	*81.0%*	*0.850*	*0.904*	*84.64%*
	SVM	*Yes*	*4*	*5*	*82.0%*	*91.0%*	*0.840*	*0.891*	*83.2%*
Z-Alizadehsani	*Bagging*	*No*	*-*	*55*	*90.7%*	*76.7%*	*0.905*	*0.871*	*86.46%*
	Bagging	*Yes*	*4*	*11*	*93.1%*	*81.3%*	*0.916*	*0.898*	*87.78%*
	Bagging	*Yes*	*7*	*11*	*91.2%*	*73.6%*	*0.904*	*0.902*	*86.13%*
	Naïve Bayes	*No*	*-*	*55*	*81.5%*	*63.3%*	*0.859*	*0.883*	*80.85%*
	Naïve Bayes	*Yes*	*4*	*11*	*87.5%*	*72.2%*	*0.896*	*0.908*	*85.47%*
	Naïve Bayes	*Yes*	*7*	*11*	*93.5%*	*85.1%*	*0.896*	*0.905*	*85.80%*
	Random Forest	*No*	*-*	*55*	*94.9%*	*86.3%*	*0.909*	*0.923*	*86.46%*
	Random Forest	*Yes*	*4*	*11*	*92.6%*	*80%*	*0.911*	*0.913*	*87.12%*
	Random Forest	*Yes*	*7*	*11*	*93.1%*	*71.3%*	*0.910*	*0.922*	*86.79%*
	KNN	*No*	*-*	*55*	*81.9%*	*61.4%*	*0.847*	*0.785*	*78.87%*
	KNN	*Yes*	*4*	*11*	*89%*	*70.1%*	*0.890*	*0.822*	*84.15%*
	KNN	*Yes*	*7*	*11*	*88.0%*	*75.9%*	*0.890*	*0.810*	*84.48%*
	MLP	*No*	*-*	*55*	*88.9%*	*73%*	*0.893*	*0.907*	*84.81%*
	MLP	*Yes*	*4*	*11*	*92.1%*	*80%*	*0.917*	*0.898*	*88.11%*
	MLP	*Yes*	*7*	*11*	*90.7%*	*78.2%*	*0.910*	*0.891*	*87.12%*
	Ensemble Classifier	*Yes*	*4*	*11*	*91.7%*	*78.2%*	*0.915*	*0.909*	*87.78%*
	Ensemble Classifier	*Yes*	*7*	*11*	*91.2%*	*80.5%*	*0.916*	*0.896*	*88.11%*
	XGBoost Classifier	*Yes*	*7*	*11*	*95.0%*	*85.0%*	*0.930*	*0.909*	*89.00%*
	SVM	*Yes*	*7*	*11*	*94.0%*	*84.0%*	*0.930*	*0.926*	*89.40%*
	Random Forest	*Yes*	*7*	*11*	*92.0%*	*79.0%*	*0.910*	*0.803*	*86.80%*
	Enemble Classifier with FLDA	*No*	*-*	*55*	*94.0%*	*87.4%*	*0.944*	*0.953*	*92.07%*

(DT: Dataset, FS: Feature Selection, AUC: Area Under the Curve)

In this study, we used the multiclass Fisher Linear Discriminant Analysis (FLDA), which is a supervised algorithm and computes the directions that will represent the axes that maximize the separation between multiple classes. To be able to run FLDA we changed the nominal attributes to numeric attributes.

G. *ENSEMBLE BASED METHODS*

Ensemble-based methods can often improve machine learning performance by combining single classifier's posterior probabilities or predicted values. This approach constructs a new model and then classify data points by taking a weighted average of each classifier's predictions. In this study, we used Naïve Bayes, Random Forest, KNN, MLP and SVM classifiers as a single classifier, and bagging, ensemble classifiers as an ensemble classifier. Table 9 shows the performance results of all these classifiers on UCI Cleaveland and Z-Alizadehsani datasets with different 10, 15, 20- fold cross-validation. Some ensemble classifiers performance results are better than single classifiers in terms of sensitivity, F-Measure, and accuracy.

III. Performance Results

Applying feature extraction, our proposed hybrid feature selection method and several classification algorithms including ensemble classifiers, this study presents comprehensive performance results obtained for CAD diagnosis dataset. Throughout this work, publicly available datasets, such as Machine Learning Repository and Z-Alizadehsani heart datasets have been utilized. In contrast to the performance metrics used in previous studies, we paid attention to the sensitivity, specificity, F-measure, AUC, running time.

According to the feature selection based on medical doctor recommendations in Cleveland dataset, we obtained 81.84% accuracy with SVM method using only CIF type of features. In Z-Alizadehsani dataset, we obtained 87.12% accuracy with SVM method using CIF and FHS RF labaled features. In Cleveland dataset, after we applied the first type of hybrid feature selection, the performances of some classification algorithms have increased. For example, when we run the Naive Bayes algorithm, the sensitivity has increased from 86.7% to 88.5%, specificity has increased from 83.3% to 86.2%, F-Measure has increased from 0.835% to 0.853% and also accuracy has increased from 83.49% to 85.14%. In UCI mix dataset, when we run XGBoost algorithm, the sensitivity has increased from 81.0% to 84.0%, the specificity has increased from 78.6% to 81.1%, F-Measure has increased from 0.820 to 0.850, accuracy has increased from 81.13% to 83.82%. In Z-Alizadehsani dataset, after we apply the second type of feature selection, the performance of some classification algorithms have increased compared to the first type of feature selection. For example, when we run Naïve Bayes algorithm, the sensitivity has increased from 87.5% to 93.5%, the specificity has increased from 72.2% to 85.1%, accuracy has increased from 85.47% to 85.80%. Unlike the feature selection algorithm, when we perform the linear discriminant analysis, the performances of some classification algorithms have increased noticeably. For example, when we run SVM algorithm, the sensitivity has increased from 91.7% to 95.8%, the specificity has increased from 78.2% to 85.1%, F-Measure has increased from 0.915 to 0.950, accuracy has increased from 87.78% to 92.74%. Performance results show that the feature selection methods generally improve the performance of the classifiers.

IV. Conclusions

With the development of information technologies, it has become possible to predict whether people have heart disease or not by checking certain physical, biochemical values at a lower cost. Although some studies provide valuable insights and foundations for CVD diagnosis, there is no internationally accepted standard machine learning approach for the CVD diagnosis. In addition, none of these studies present a detailed performance evaluation of different classification methods and feature selection algorithms in terms of accuracy, sensitivity, specificity, F-measure, AUC and running time. In this study, we have experimented a set of different classification algorithms, ensemble classifiers, linear discriminant analysis, proposed a new hybrid feature selection methodology and a feature selection methodology based on doctors' recommendation for the diagnosis of CAD. Utilizing from medical doctors' expertise and medical literature in feature selection process resulted in 81.84% accuracy in Cleveland Dataset. This value is less than the accuracy value obtained with the model which uses variables from our hybrid feature selection methodology. Also, in Z-Alizadehsani dataset, we reached 87.12% accuracy with FHS RF and CIF labeled data. This value is nearly the same with the accuracy value obtained via our hybrid model. Future work includes the investigation of the performance of the proposed approach in different datasets. As a result, it can be inferred that hybrid feature selection methodology increases the performance results of algorithms and gives better results than the area expert's feature selection. Thus, this methodology can also be considered as a tool to reveal important attributes in expertise areas. It is proposed to create a large heart disease data set obtained by today's technological capability. According to these performance metrics results, this system can be run in intensive care units.

Acknowledgements

This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Project number 3180177.

References

[1] "Cardiovascular diseases (CVDs)," *World Health Organization*. Available: http://www.who.int/mediacentre/factsheets/fs317/en/. [Accessed: 01-Nov-2018].

[2] Verma, Luxmi, Sangeet Srivastava, and P. C. Negi. "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data." *Journal of medical systems*40.7 (2016): 178.

[3] Alizadehsani, Roohallah, et al. "Diagnosis of coronary artery disease using cost-sensitive algorithms." *Data Mining Workshops*

(ICDMW), 2012 IEEE 12th International Conference on. IEEE, 2012

[4] Polat, Kemal, Seral Şahan, and Salih Güneş. "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 (2007): 625-631.

[5] Tu, My Chau, Dongil Shin, and Dongkyoo Shin. "Effective diagnosis of heart disease through bagging approach." *Biomedical Engineering and Informatics, 2009. BMEI'09. 2nd International Conference on*. IEEE, 2009.

[6] Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." *Expert systems with applications* 36.4 (2009): 7675-7680.

[7] Shouman, Mai, Tim Turner, and Rob Stocker. "Using decision tree for diagnosing heart disease patients." *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*. Australian Computer Society, Inc., 2011.

[8] Karabulut, Esra Mahsereci, and Turgay İbrikçi. "Effective diagnosis of coronary artery disease using the rotation forest ensemble method." *Journal of medical systems* 36.5 (2012): 3011-3018.

[9] Shouman, Mai, Tim Turner, and Rob Stocker. "Applying k-nearest neighbour in diagnosing heart disease patients." *International Journal of Information and Education Technology*2.3 (2012): 220-223.

[10] Nahar, Jesmin, et al. "Association rule mining to detect factors which contribute to heart disease in males and females." *Expert Systems with Applications* 40.4 (2013): 1086-1093.

[11] Subanya, B., and R. R. Rajalaxmi. "Artificial bee colony based feature selection for effective cardiovascular disease diagnosis." *International Journal of Scientific & Engineering Research* 5.5 (2014): 606-612.

[12] Yadav, Chetna, Shrikant Lade, and Manish K. Suman. "Predictive analysis for the diagnosis of coronary artery disease using association rule mining." *International Journal of Computer Applications* 87.4 (2014).

[13] El-Bialy, Randa, et al. "Feature analysis of coronary artery heart disease data sets." *Procedia Computer Science* 65 (2015): 459-468.

[14] Verma, Luxmi, Sangeet Srivastava, and P. C. Negi. "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data." *Journal of medical systems*40.7 (2016): 178.

[15] Babič, František, et al. "Predictive and descriptive analysis for heart disease diagnosis." *Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on*. IEEE, 2017.

[16] Samuel, Oluwarotimi Williams, et al. "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction." *Expert Systems with Applications* 68 (2017): 163-172.

[17] Kolukisa, Burak, et al. "Evaluation of Classification Algorithms, Linear Discriminant Analysis and a New Hybrid Feature Selection Methodology for the Diagnosis of Coronary Artery Disease." *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018.

[18] Detrano, Robert, et al. "International application of a new probability algorithm for the diagnosis of coronary artery disease." *American Journal of Cardiology* 64.5 (1989): 304-310.

[19] "Assessing Cardiovascular Risk: Systematic Evidence Review from the Risk Assessment Work Group," *National Heart Lung and Blood Institute*.Available: https://www.nhlbi.nih.gov/health-topics/ assessing - cardiovascular-risk. [Accessed: 01-Nov-2018].

[20] "Framingham: Past and Present," *Framingham Heart Study*. Available: https://www.framinghamheartstudy.org/. [Accessed: 01-Nov-2018].

[21] Wilson, Peter WF, et al. "Prediction of coronary heart disease using risk factor categories." *Circulation* 97.18 (1998): 1837-1847.