Available online at http://ijdea.srbiau.ac.ir

Int. J. Data Envelopment Analysis (ISSN 2345-458X)

Vol. 12, No. 4, Year 2024 Article ID IJDEA-00422, Pages 30-48
Research Article

International Journal of Data Envelopment Analysis

Science and Research Branch (IAU)

Prediction of Heart Disease Based on Data Mining Techniques

Arash Moradi, Mojtaba Alizadeh*
Department of Computer Engineering, Lorestan University, Khorramabad, Iran.

Received 3 April 2024, Accepted 1 October 2024

Abstract
Heart disease is one of the most popular health problems for human being. According to statistics published by the World Health Organization (WHO), two main diseases that caused about a quarter of deaths all around the world in 2016, were heart disease and stroke. Although a lot of articles suggested different ways of detection and treatment of the heart and its disease, no framework is proposed to leverage the machine learning techniques to do it. We tabulate 32 articles regarding the factors such as Algorithm, Database, Feature Selection, Evaluation Measures, Performance evaluation, Efficiency, and different used tools. The same factors also are leveraged for comparison including 17 papers that focus on ECG and HRV signals. The output of this comparison leads to a classification framework. All steps from pre-processing to classification are provided in the proposed framework. This survey and the proposed framework offer a roadmap that can help researchers to conduct future researches in the field.

Keywords: Heart Diseases, Data Mining, Machine Learning.

* Corresponding author: Email: alizadeh.mo@lu.ac.ir

1. Introduction

Based on the World Health Organization (WHO) [1] 15.2 million out of 56.9 million deaths were caused by two main diseases, Ischaemic heart disease and stroke, as shown in Figure 1. When the arteries responsible for providing the heart with oxygen and blood are completely clogged or narrowed, coronary artery disease (CAD) happens [2, 3]. Many disorders and issues that are commonly referred to as cardiovascular illnesses affect the heart and blood arteries. These conditions include angina, valvular heart disease, coronary heart disease, hypertensive heart disease, heart failure, myocardial infarction, and ischemic heart disease [4, 5].

One method to assess the CAD risk is to investigate the risk factors [6, 7]. Some of the risk factors for coronary artery disease (CAD) that have been identified by the literature include high blood pressure, high levels of low-density lipoprotein cholesterol (LDL-C), a family history of CAD, low levels of HDL-C, age, high total cholesterol, physical inactivity, high triglycerides, gender, diabetes mellitus, aging, obesity, smoking, and various genetic factors [8, 9]. Various meta-analysis studies have been conducted on other risk factors. As an example, it was discovered that some conditions including job strain [10], depression [11], anxiety [12], fried-food consumption [13], a diet poor in fruit and vegetable [14] increase the risk of CAD significantly.

Due to the substantial increase in coronary artery disease, its effects and consequences, and its high cost on society [15], the medical community is seeking solutions to prevent this disease, identify it early and treat it effectively through less expensive ways.

The increase in the usage of electronic health records (EHR), on the other hand, is innovative in the healthcare industry and opens up new opportunities for personalizing clinical diagnosis and decision-making [16]. Raw EHR data, on the other hand, may not be the best for analytical purposes or even clinical care. It is comprised of a diverse range of elements (such as clinical measures, diagnoses, drugs, and procedures) and large amounts of unstructured content. In other words, this industry is rich in data yet deficient in knowledge [17].

Healthcare organizations should utilize suitable decision support systems to diagnose patients' diseases accurately to prevent wrong decision, which could lead to unacceptable consequences also to minimize the cost. Traditionally, doctor intuition and experience playing an important role in Clinical decisions.

Medical diagnosis is very hard and complicated, and computer-aided diagnosis systems are designed to reduce observational oversight [18]. Since computers can store and process vast amount of data without distortion overtime, while carrying out complex computing operations quickly. Therefore, Computer-aided diagnosis could be helpful in conducting decision and classification procedures [19]. Due to their capacity to enhance healthcare quality like supporting and improving clinical decisions, Computer-based medical diagnosis systems are technologically advanced [20, 21].

Nevertheless, using of HER is increasing, knowledge-rich data can play an important role in Clinical decisions that leads to improving QoS (decrease medical errors, reduce unwanted practice variation, enhance the patient outcome, improve patient safety) via clinical decision support systems. [22-24]

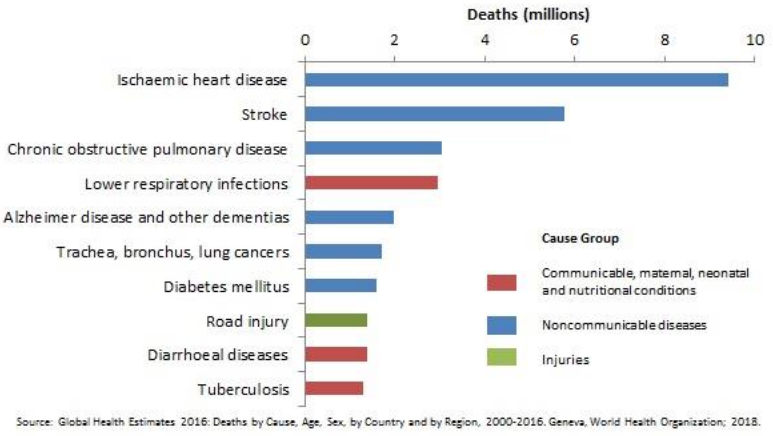

Figure 1. Global causes of human deaths in year 2016

Diagnostic system development is an intriguing topic in the field of biomedical informatics because of the prevalence of data mining as a tool for evaluating health care data, the death rate from coronary artery disease, and the significance of timely diagnosis, which consists of making the application domain understandable, producing a data set, preprocessing and transforming of data, and developing of the model [25] [26]. This research focuses on reviewing potential types of data mining methods that are used to process clinical attributes to distinguish between CAD patients and healthy.

This research aims to achieve three main goals including developing a framework to classify data mining applications for heart disease prediction, providing a comprehensive and systematic survey on the current researches in data mining applications fields, and creating a roadmap for researchers, who are working in the field.

2. Research methodology

The methodology of this study consists of three main phases including research definition, research methodology designing, and analyzing the research, which are described as follows.

In phase 1, goals and scope of the study are determined. As the research domain is reviewing academic papers that used data mining for heart disease diagnosing, the main goal of this study is to suggest future direction based on analyzing the exist researches in the field. The scope of the study is the paper that focus on the field of data mining application for prediction of heart disease and are published between year 2014 and 2020.

We proposed a framework to classify the selected papers based on common criteria, in phase 2. Finally, the selected articles are analyzed to reach conclusions and obtain some directions for future research. The analysis details have been presented in the Analysis of diagnosis heart disease papers based on the proposed framework.

Research on data mining and heart illnesses is not confined to any one field, and related studies are published in diverse journals. Table 1A summary of the journals that published papers in the field is provided in Table 1. The chosen papers are all about data mining methods used to identify cardiac conditions. The data type and applied features are used to categorize these investigations. Sex, age, blood pressure, blood sugar, kind of chest discomfort, serum cholesterol, maximal heart rate, and other significant characteristics are included in these

datasets. Furthermore, some researches focus on just ECG and Heart Rate Variability (HRV), to predict heart disease [27-29].

Table 1. Number of papers published by journals

Journal Name	Number of Papers
Knowledge-Based Systems	4
Indian Journal of Science and Technology	4
Computer methods and programs in biomedicine	4
Information sciences	3
Healthcare informatics research	3
Expert Systems with Applications	3
Biomedical Signal Processing and Control	2
Journal of medical systems	2
IEEE Trans. Biomed. Eng	1
Clinical Medicine Insights: Cardiology	1
Australasian physical & engineering sciences in medicine	1
International Journal on Computational Science & Applications (IJCSA)	1
Proceedings of the IEEE	1
CSI transactions on ICT	1
arXiv preprint arXiv	1
Applied Medical Informatics	1
BMC medical informatics and decision making	1
Cluster computing	1
IEEE Transactions on Biomedical Engineering	1
Entropy	1
Multimedia tools and applications	1
IEEE journal of biomedical and health informatics	1
PloS one	1
Journal of Cardiovascular Disease Research	1
Patient preference and adherence	1
Computers in biology and medicine	1
Peertechz J Biomed Eng	1
2017 IEEE Symposium on Computers and Communications	1
Applied Soft Computing	1
Advances in Natural and Applied Sciences	1
International Journal on Computational Science & Applications	1

3. Data mining and heart disease classification framework

A survey of the literature on data mining approaches in heart disease prediction served as the foundation for the development of the classification framework used in this study. Figure 2 illustrates the four general processes of data mining: datasets/data preprocessing, feature extraction/feature selection, data mining algorithm creation, and evaluation. Therefore, the framework was developed in agreement with these four steps.

As illustrated in Figure 2, the initial phase of the suggested approach entails locating the information source and getting it ready for the valuable data to be extracted. The next step is to find and extract a distinctive attribute or aspect of data that call feature extraction and then find the best feature among feature and omit irrelevant feature this process call Feature selection. Then,

apply data mining algorithms to the feature that selected -make the model- to discover the interesting patterns. Finally, the proposed model is evaluated based on performance criteria. The performance of classifier is relied on two factors: the quality of signals being preprocessed and the quality of features extraction. On the other hand, the modeling and evaluation phases are combined and can be repeated for many times to change parameters until optimum values are reached. Heart Diseases that have done can be classified according to their contribution in four areas as shown in the framework.

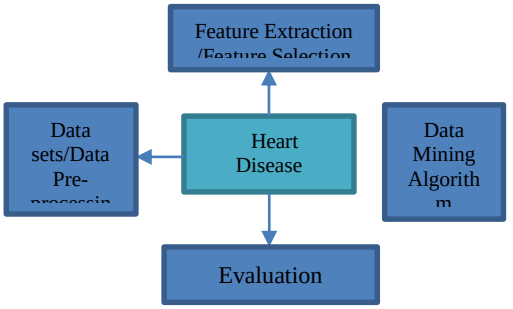

Figure 2: The proposed framework

4. Data sets/Data Preprocessing

As previously said, the initial stage of data mining techniques is to build or find data sets and prepare them for the extraction of usable information. Data mining is a "data-driven" process. The majority of the assessed papers employed standard datasets for research on heart disease prediction. Among the most well-known is the UCI repository, which houses a data set on heart disease. There are four divisions in this repository: University Hospital, Switzerland; Cleveland Clinic Foundation; V.A. Medical Center, Long Beach; and Hungarian Institute of Cardiology. As indicated in Table 2, the majority of articles employ a subset of 14 of the 76 features in the Cleveland dataset, which is the most widely used dataset.

Table 2. Cleveland principal attributes

1	#3	age
2	#4	sex
3	#9	cp
4	#10	trestbps
٥	#12	chol
6	#16	fbs
7	#19	restecg
8	#32	thalach
9	#38	exang
10	#40	oldpeak
11	#41	slope
12	#44	ca
13	#51	thal
14	#58	Num (the predicted attribute)

Some papers focus on electrocardiogram (ECG) signals and try to predict heart condition with the use of ECG. Physio net [30] databases are among most referenced database. This database contains Fantasia that is created by Normal and St.-Petersburg Institute of Cardiology Technics. They examined 12-lead arrhythmia for CAD patients.

Data quality has a decisive role in data mining, which is relied on data preprocessing. This procedure consists of three phases including filling missing values, removing outlier, and normalization phases. For example, for dealing with outlier sometimes it is best to completely remove or cap your data, assign a new value, or try a transformation. On the other hand, dealing with the missing value it can be critical to predict heart disease as property of heart disease datasets are sparse. There are several strategies for dealing with missing value, from deleting records with missing value to used Truncated mean in order to manage numerical missing values to fill the missing values.

By putting sensors on the chest or other body surface, ECG signals are acquired. Following the collection of ECG signals, these signals undergo preprocessing using the Daubechies wavelet 6 basis function, resulting in the removal of noise. [31, 32]. ECG signal split to its beat is then required. There are a number of techniques

for automatic ECG segmentation in the literature [33]. Because of development in ECG beat segmentation, this part is a routine procedure.

5. Feature Extraction /Feature Selection

A feature is a single, quantifiable attribute or characteristic of a phenomena under observation. The selection of discriminating, independent, and informative features is a crucial stage for algorithms to be effective in recognizing patterns. In order to create more manageable groupings that are nonetheless correct data sets, the dimensions of an initial set of raw variables are lowered in features extraction. Variables and features are integrated in this phase to minimize the quantity of data that needs to be processed [34].

A subset of relevant features is utilized to develop the model during the feature selection phase. Since we can eliminate these data without losing any information, we can presume that some of the data are redundant and unnecessary [35]. Model simplification is achieved by using feature selection approaches. These strategies serve four purposes: decreasing overfitting, avoiding the curse of dimensionality, simplifying models, and shortening training periods. A subset of features is preserved in feature extraction while new features are created based on the original feature functions. This is the distinction between feature extraction and feature extraction. When we have a large number of features but limited samples, we use feature selection strategies.

One of the features in an ECG signal, are P-QRS-T waves. The intervals and amplitudes of ECG signals are selected to distinguish human's heart functions [36]. Few techniques and methods such as Discrete Wavelet Transform (DWT) [37] and Discrete Cosine Transform (DCT) [38] are utilized to extract features from ECG. According to literature, researches that used Cleveland dataset, attributes are used as feature, however, other studies used technique such as particle swarm optimization algorithm, generic algorithm, Principal component analysis (PCA) to reduce data size.

6. Data Mining Algorithm (model development)

We discover new knowledge from data by utilizing data mining techniques. There are six categories of data mining techniques including Prediction, Outlier detection, Clustering, Association, Classification, Regression and Visualization [39]. Each of these categories is supported by special algorithmic approaches to extract the relevant relation between data.

An example, developing classification model can be divided into three main phases, training, validation, and testing. The classifier is created based on training sets that are obtained from datasets tuples and their class labels. In validation phase, the hyper-parameters of a classifier are tuned. Finally, in the test phase, we validate the model that is created in the training phase.

Several tools (sometimes referred to as data modeling or data analysis) are employed in the model-development process. These tools enable developers to experiment with a large number of methods and make use of them without having to start from scratch. MATLAB, Weka, Tanagra, R, TensorFlow, and Fast Miner are a few of the tools.

7. Performance evaluation

For tests that a specific model is suitable for its intended use or not, we use validation techniques. Two basic problems

in data mining, i.e., performance estimation and model selection motivate validation techniques. There are approaches to assess a data mining model's characteristics and quality, including the usage of different statistical validity measures in order to specify whether problems exist in the model or in the data. Accuracy is an important criterion to validate and test data mining models. In addition to accuracy, Sensitivity, ROC Curve (AUC), Specificity, Recall, Precision, and F-Measure and Area under the are among most used statistical measurements for evaluation performance of models in the literature as shown in Table 3.

In real applications, there are just some examples such as K-Fold Cross-Validation and Cross Validation. In fact, all presented examples in the dataset are finally used for both testing and training is the key advantage of using K-Fold Cross-Validation. The procedure of K-Fold Cross-Validation is as follows:

1. Shuffle the dataset randomly

2. The shuffled dataset is divided into k groups

3. For each group, we apply following actions:

A. We test the data and hold the group

B. The rest of the groups are used as a training data set

C. The model is evaluated

D. The group receives the evaluation score.

We summarize the model's skill based on the evaluation scores

Accuracy is a metric to evaluate classification models, which means the percentage of accurate predictions. For evaluation, most of the paper used accuracy alongside with sensitivity and specificity. Table 3 shows the performance evaluation metrics for binary classification.

Table 3. Performance Evaluation For binary classification

Criterion	**Calculation Formula**	**Focus of Evaluation**
Accuracy	$\frac{tp+tn}{tp+fp+tn+fn}$	The accuracy of a prediction is its rate of correctness.
Specificity (sp)	$\frac{tn}{tn+fp}$	The quantity of negative patterns that are categorized as proper patterns is known as specificity.
Error Rate (err)	$\frac{fp+fn}{tp+fp+tn+fn}$	The rate of inaccurate prediction relative to all cases analyzed is known as the error rate.
Recall (r)	$\frac{tp}{tp+tn}$	The rate of correctly recognized negative patterns is called recall.
Sensitivity (sn)	$\frac{tp}{tp+fn}$	The rate of correctly recognized positive patterns is known as sensitivity.
F-Measure (FM)	$\frac{2*p*r}{p+r}$	F-Measure is the mean value between recall and positive values.
Precision (p)	$\frac{tp}{tp+fp}$	The accurately anticipated positive patterns in a positive class are known as precision.
Where TN = True Negatives, FN = False Negatives, TP = True Positives, and FP = False Positives		

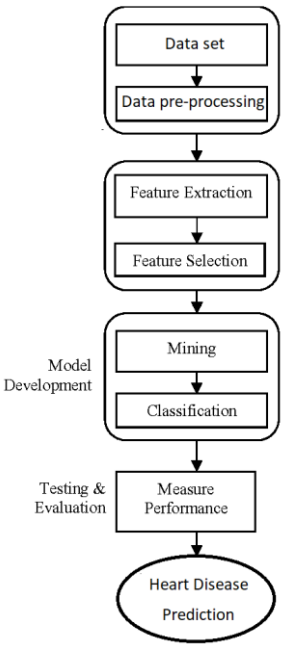

Figure 1. Different stage of Data Mining

8. Analysis of diagnosis of heart disease papers

This study reviews the applications of data mining in the diagnosis of heart disease. The review of the literature is organized according to the framework as shown in **Error! Reference source not found.**. 49 papers are chosen according to criteria and read carefully for determining four main steps according to the framework for each paper.

Generally, papers that review in this research can be divided into group according to the data used. The first category (Table 1) is papers used database such as Cleveland that key attributes are shown in **Error! Reference source not found.**. However, some papers only focused on electrocardiogram (ECG).

As it is described in [], all reviewed papers are in the field of ECG signals classification. In the first category for each paper researcher look for methods and Algorithm, data, Feature Selection, Evaluation measures, Performance evaluation, Efficiency, and tools. However, for papers that focus on ECG, Feature extraction from row ECG signals is prime part of works so the feature extraction are also added to the literature review.

The majority of the research studies employed standard databases from Physio Net, UCI medical datasets, or similar databases. Scholars, educators, and students worldwide employ it as the primary source of machine learning datasets. The primary databases for cardiac disease diagnosis are those from Cleveland, Hungarian, Long Beach, and Switzerland. Physio Net provides access to enormous recordings of physiological signals. Established in 1999, the goal of the Physio Net resource is to promote ongoing and new studies into complex physiological and biological signals.

A crucial component of data mining is preprocessing the data. Preprocessing and segmentation using common tools and methodologies now aid researchers in obtaining satisfying findings; yet, the number of published research studies in this topic is inadequate. Preprocessing of the ECG data mostly involves denoising and subtracting baseline drift using techniques like Daubechies wavelet. Preprocessing is therefore not included in this literature analysis due to space constraints and standard operating procedures.

As shown in Figure 1 after preprocessing Feature Extraction and Feature Selection are the next steps in Heart Disease Prediction. Feature Extraction is a crucial step for ECG and HRV signal analyzing. Feature extraction is a process of

transforming high-dimensional data to a low-dimensional using mapping techniques. Moreover, feature extraction can generate a new feature based on transformation and combination the original feature. There are two groups of feature extraction, linear and non-linear [40]. Wavelet Transform is a popular method to extract feature among other techniques based on the literature. Feature selection ranks the present attributes based on their predictive significance. It often addresses three types, namely filtering, embedded and wrapper [40].

The first step in this literature review was to look for used methods and Data Mining algorithms. The most frequently used techniques were SVM, Decision Tree, and Artificial Neural Network. According to literature, most studies applied K-fold cross validation method to in evaluation phase. In addition, performance evaluation was done mostly with help of accuracy, sensitivity, and specificity.

There are numerous data mining tools available in the literature that can help us to develop a model. The last part of this literature review is to review and analyze tools used in each research. Tools like MATLAB, Weka, R and etc., dramatically lowers the barrier to write a research paper.

9. Conclusion, research implications and limitations

Although data mining techniques play a significant role in the diagnosis of cardiac disease, the topic lacks a thorough study. In order to determine future prospects for our study, we examined and analyzed recent research in the subject. Additionally, a paradigm is put forth for using data mining to diagnose cardiac disease. The study's findings allow the following observations to be made:

- Data is a very crucial part for developing data mining model. As discussed in
  - Table 1, most studies used the Cleveland or similar database, which literature, many researches utilized standard databases such as Cleveland and MIT-BIH. As the amount of data daily in the healthcare industry is increasing instantly, there is need for a bigger and up-to-date database. These amount of healthcare data show very promising future for data driven research and product in healthcare, especially in heart disease diagnosis. The only issue for conduction researches in the field, is privacy issues of collecting the healthcare data.
- According to the literature, the heart disease diagnosis is become matured, and the trend continues in future, so that, the future researches will focus on practical applications. The promising result causes some researcher to look for developing Clinical Decision Support System (CDSS) [41-44]. These systems can assist healthcare professional with decision-making duties.
- Deep learning, which is regarded as a ground-breaking technique that radically altered the perspective of data mining and machine learning research, is a popular topic. The primary benefit of utilizing deep learning techniques is their capacity to automatically extract and choose valuable features. Convolutional networks are one type of deep learning technique. The literature claims that deep learning was utilized for training purposes rather than feature extraction. Deep learning techniques are only employed in ECG and HRV systems that have access to sufficient data for training, as they require at least two or three orders of magnitude more data.
- Many different heart related diseases are investigated in the literature. In one hand, based on

  focuses just on distinguishing the existing of heart diseases. On the other

hand, ECG and HRV signal used for diagnoses diverse heart related diseases such as Sudden cardiac death, normal and Coronary Artery Disease conditions [45, 46], congestive heart failure [47], myocardial infarction [37, 48] and etc.

Table 1. Literature review

#	Year	Ref	Method & Algorithm	Data & Database	Feature Selection	Evaluation measures/ Performance evaluation	Efficiency	Tools
1	2015	[41]	Hybrid/ Evolutionary/ algorithm/ EPSO + ABC /SVM classifier	UCI medical datasets	EPSO + ABC	5-fold cross-validation/accuracy Full set of features: 82.59/ Selected features 93.33%/ Friedman test/p-value=0.00657	Selected features 93.33%/ /Friedman test/p-value=0.00657	Dev C++ / LibSVM
2	2015	[42]	NB, LR, QDA, IBL and SVM framework	UCI SPECT, SPECTF, Cleveland, and Statlog/ ricco Eric	No feature selection	ANOVA statistics/Ten-fold cross validation/accuracy- Confusion Matrix	Cleveland achieved best 84.16% accuracy	-
3	2015	[43]	fuzzy logic and decision tree - CART	KNHANES-VI	No feature selection	Confusion matrix/ ROC curve	Accuracy 69.51% Sensitive 93.10%	IBM SPSS/ MATLAB R2009b fuzzy tool box
4	2015	[44]	Genetic-fuzzy combination for heart disease diagnosis	UCI Machine learning repository	The genetic algorithm	The stratified k fold technique (k = 10) /accuracy, specificity, sensitivity, confusion matrix	accuracy was 86%	-
5	2014	[45]	Prediction model based on Fuzzy Rule Adaptive Coronary	G Medical Center in Korea	No feature selection	Divided datasets into training sets (70 %) and testing sets (30 %)	accuracy rate of 69.22 %,	MATLAB Toolbox FIS Editor.
6	2015	[46]	DSS with Random Forest algorithm	PhysioNet	No feature selection	10-fold cross-validation/ specificity / sensitivity/accuracy	Decompensation's prediction: 71.9%/ Severity assessment: 81.3%	Matlab/ HRV toolkit from PhysioNet
7	2015	[47]	CAD with SFCM classification/GMM/Statistical feature selection	Cleveland clinic foundation	SFS/MLR	10-fold cross validation	SFCM + GMM accuracy: 79%/ Using SFS: 82%/DSA-based accuracy: 88%	-
8	2015	[48]	NB, DT, SVM, LoR, MLP and Adaboost/cardiovascular diseases dataset	UCI repository, Cleveland/Hungarian dataset/ LongBeach/ Switzerland	Genetic Search (GS); InfoGain (IG)	-	Adaboost with DT as classifier: 98% accuracy for Cleveland dataset	Weka tool
9	2014	[49]	CART / Random Forests /Clinical Decision Support System	St. Maria Nuova Hospital, Florence, Italy	No feature selection	ten-fold cross validation	Random Forest: accuracy 83.3%/CART: 87.6% accuracy	Matlab R2010b./ Microsoft .NET
10	2014	[50]	LoR, ANN, and multivariate adaptive regression splines.	UCI repository	LoR/MARS/ RS techniques	60% used for the model building set and 40% retained as the validation set	MARS–LoR with 6 feature/ RS–LoR with 10 features: accuracy rate of AIR: 83.93%	Qnet97/ SPSS/ MARS/ RESE
11	2015	[51]	conjugate gradient algorithm and neuro-fuzzy classifier	Cleveland	MLR/FS	"Hold-out" method, 10-fold cross-validation / SE/ SP	MLR/NFC: 84.2% accuracy	Matlab/ SPSS
12	2015	[52]	ML unsupervised and supervised	LV volume/pressure data	No feature selection	cross-validation (10-fold)	SVM: 4.06% test error	Weka

13	2015	[53]	SVM (multiclass performance classification)	Cleveland	No feature selection	Cross-validation (10-fold cross)/ precision/Recall	SVM: 90% accuracy	
14	2015	[54]	Adaptive Neuro Fuzzy Inference System	Not publicly available.	PSO	Confusion matrix	Diabetic / Heart: 98% accuracy	MATLAB
15	2016	[55]	Decision Tree/Naive Bayes/Neural Networks	Cleveland	Using 8 attributes out of the 13.	Accuracy and time complexity.	ANN: 100% accuracy	WEKA/ C# and Python platform.
16	2016	[56]	CAD/PSO	Cleveland/ Indira Gandhi Medical College dataset	PSO	cross validation (10-fold cross)/ confusion matrix	MLR: 88.4 % accuracy	Weka
17	2016	[57]	Optimized Particle Swarm Optimization (PSO) dimension reduction + Optimized Artificial Neural Network	UCI repository	PSO for Feature Reduction	Performance plot, Regression, ROC Value and Confusion Matrix. F-measure	CPSO-ANN has best performance of 97.2%	-
18	2016	[58]	CVD (ScoreCard)/lasso logistic regression techniques	University of Kentucky (UK)/ Texas Medical Center (TMC) datasets	Lasso using continuous subset selection	5-fold cross-validation/ PPV/ SP/AUC, SE/ NPV	AUC: 0.8403 and 0.9412 accuracy	R statistical software
19	2016	[59]	CVD	CVD/ University of Washington dataset	No feature selection	Select and divide training, validating, and test sets	CVD: 98.57% accuracy	-
20	2016	[60]	SVM/Apriori algorithm	Z-Alizadeh Sani dataset	SVM	Cross validation (10-fold cross validation)/ Precision/ F-Measure/Recall and	LAD, LCX, RCA: 86.14%, 83.17%, and 83.50% accuracy rates	RapidMiner
21	2016	[61]	Weighted Associative Classifier (WAC)/ C5.0 Prediction Tree	Cleveland	No feature selection	Accuracy and computation time	94.54% accuracy	MapReduce
22	2016	[62]	K-star algorithm	Cleveland	No feature selection	cross-validation (k-fold)/ F-measure/SE/positive prediction value/SP/AUC/negative prediction value	SE: 80.1% / SP: 95%/PPV: 80.1%/NPV: 95%/AUC: 87.5%/ F-measure: 80.1%. accuracy	-
23	2016	[63]	C4.5/SMOTE/multiclass/ level diagnosis of coronary heart disease/system for clinical data interpretation for type	Cleveland	information gain (IG)	cross-validation (k-fold)/ F-measure/SE/positive prediction value/SP/AUC/negative prediction value/confusion matrix	Sensitivity: 74.7%/specificity: 93.7%/PPV: 74.2%/NPV:93.7%/AUC: 84.2%	-
24	2016	[64]	LoR/SVM/MLP/KNN	Palo Alto Medical Foundation (Sutter-PAMF)		cross validation (six-fold)	logistic regression: AUC 0.766 to 0.791, SVM: AUC 0.736 to 0.791, neural network: AUC 0.779 to 0.814, KNN: AUC 0.637 to 0.785	Theano/CUDA/Scikit-Learn
25	2017	[65]	SVM/Ensemble Machine Learning/ MLP/DT/ K-NN/NB/ RBF/SCRL	Cleveland	-	Cross-Validation (10-Fold)	SVM: 84.81% acuracy	-
26	2017	[66]	ANN/Fuzzy AHP	Cleveland	No feature selection	cross-entropy/percent error/ROC plot/Performance plot	91.10% accuracy	MATLAB/Microsoft Excel/SPSS

27	2017	[67]	Hybrid neural network-Genetic algorithm	Z-Alizadeh Sani dataset	Gini index: SVM and PCA	cross validation (10-fold)	93.85% accuracy	-
28	2017	[68]	SVM/MLPE/GAM	'statlog' from the UCI		cross validation (10-flods)	SVM: 85% accuracy	R statistical software
29	2017	[69]	SVM/RF/GBRM/LDA	Cleveland	Mean Decrease in Gini Index	Accuracy/Specificity/sensitivity	91.89% accuracy	R Language
30	2017	[70]	SVM/Multi-Layer Perceptron neural network	Cleveland	-	Positive Prediction Value (PPV)/testing time/ specificity/sensitivity/Negative Prediction Value (NPV)/accuracy	MLP:98% and in case of adding four types of heart diseases classification 81%. accuracy	MATLAB
31	2018	[71]	MKL/ANFIS	-	-	MSE/ Specificity/Sensitivity	MKL and ANFIS: 98%, specificity: (99%) accuracy	-
32	2018	[72]	SVM/LoR/RF/EHR method	Boston Medical Center (BMC)	K-LRT	Cross-validation/ROC curve	AUC: 81.62% accuracy	-

Abbreviations: Endocrine-based Particle Swarm Optimization (EPSO), Support Vector Machines (SVM), Left Circumflex (LCX), Artificial Bee Colony (ABC), University of California, Irvine (UCI), Linear Regression (LR), Naive Bayes (NB), Quadratic Discriminant Analysis (QDA), Instance Based Learner (IBL), Receiver Operating Characteristic (ROC), Neural Network (NN), Classification and Regression Tree (CART), Differential Search Algorithm (DSA), The Generalized Minkowski Metrics (GMM), Multi-Layer Perceptron (MLP), Logistic Regression (LoR), Rough Set Exploration System (RESE), Chronic Heart Failure (CHF), Least Squares Support Vector Machine (LS-SVM), Heart failure (HF), Left Anterior Descending (LAD), Adaptive Neuro-Fuzzy Inference System (ANFIS), Right Coronary Artery (RCA), Area Under the Curve (AUC), Single Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF), Multi-Layer Perceptrons ensembles (MLPE), Generalized Additive Model (GAM), Generalized Boosted Regression Modelling (GBRM), Linear Discriminant Analysis (LDA), Multiple Kernel Learning (MKL), Electronic Health Records (EHR),

Table 2. Literature review of papers that focus on ECG and HRV signal

#	Year	Ref	Method/Algorithm	Data/Database	Feature Extraction	Feature Selection/number of features	Evaluation measures/ Performance evaluation	Efficiency	Tools
1	2014	[73]	SVM/LDA/RBF/MLP/ KNN	MIT/BIH database in PhysioNET	wavelet packet transform /several nonlinear parameters / standard HRV measures	59 features	leave-one-out cross validation/sensitivity/specificity/ Accuracy	Sensitivity: 82.75%/ Specificity: 96.29%/Accuracy: 91.56%	MATLAB/ IBM SPSS
2	2014	[74]	SVM	MITDB/CUDB/VFDB	Temporal/Morphological Parameters/Spectral parameter/ Complexity parameters	filter-type FS/Fisher criterion/mRMR Criterion	Five-fold cross-validation /bootstrap resampling/ ROC/curve/SE/SP/AUC/PP/Accuracy	SVM with FS: 98.4% accuracy/ SVM without FS: 98.6% accuracy	-
3	2015	[75]	SVM	Normal Sinus Rhythm and MIT-BIH SCD Holter databases	Nonlinear Feature Extraction/ Discrete Wavelet Transform	Ranked using t-value	cross validation (10-fold)	KNN: 92.11%, SVM: 98.68% accuracy	-
4	2015	[76]	SVM/AdaboostM1/RF/ Decision tree/ NB	the University Hospital Federico II.	HRV/Pattern Recognition	Random Forest/ Chi-squared statistics	cross validation (10-fold)/ ROC curves	RF: 85.7%, SVM: 90.1% accuracy	MATLAB
5	2015	[77]	LS-SVM/Heart sound feature	Chongqing University of Medical Sciences	CR	Heart sounds characteristics/ ROC	cross validation (double-fold)	95.39% accuracy	SPSS
6	2016	[78]	CAD/LS-SVM/Flexible Analytic Wavelet Transform	Iqraa Hospital Calicut database	K-NN/ FzEn	Wilcoxon /ROC/ Entropy features/ Bhattacharya space algorithm	Cross-validation (10-fold)/accuracy/specificity/sensitivity	LS-SVM: 100% accuracy	Matlab/ FAWT Toolbox
7	2016	[37]	k-Nearest Neighbours /MI	Physiobank	DWT/the extraction of different nonlinear features	T-test and ANOVA	Cross validation (10-fold)	98.80% and 99.97% accuracy for locating inferior posterior infarction	-
8	2016	[79]	CHF/K-NN/C4.5/SVM/RF	BIDMC/PTB/ PhysioNet	Autoregressive Burg method is applied for extracting features.	Autoregressive Burg method	Cross validation (10 -fold)/accuracy/ROC/sensitivity/specificity	100% classification accuracy	-
9	2016	[80]	Deep learning and Active Learning (AL)	MIT-BIH/ SVDB/ INCART	Stacked denoising autoencoders	-	cross-validation (5-fold)/SP/Se/positive predictive/Sp	MIT-BIH: 99.9%, INCART: 99.91%, SVDB: 99.58 % accuracy	-

10	2016	[81]	WPE/RF	MIT–BIH Arrhythmia database	WPD/RR/Entropy	No feature selection	cross validation (10-fold)	94.61% accuracy	Matlab
11	2016	[82]	1-D CNN	MIT and BIH arrhythmia	CNN	-	Accuracy/Sen/Positive predictivity/sensitivity	VEB: 98.6%, SVEB: 96.4% accuracy	C++ over MS Visual Studio
12	2017	[83]	CNN/MI ECG beats	Physikalisch-Technische Bundesanstalt diagnostic ECG database	no feature extraction	No feature selection	cross-validation (10-fold)/sensitivity/ specificity/accuracy	With noise: 93.53% and 95.22% accuracy without noise	-
13	2017	[84]	CAD/KNN and DT	St. Petersburg Institute of Cardiological and Fantasia	HOS bispectrum and cumulant features	Principal Component Analysis (PCA)	cross validation (ten-fold)/SE/SP/ROC	KNN (with 13 bispectrum features): 98.17%, KNN (using DT with 31 cumulant features): 98.99% accuracy	-
14	2017	[32]	CNN	MIT-BIH/ Creighton University ventricular tachyarrhythmia	no need to for features extraction	No feature selection	cross-validation (ten-fold)/SP/accuracy/SE	Two seconds of ECG: 92.50%, and five seconds of ECG: 94.90% accuracy	-
15	2017	[85]	CNN	Physionet Fantasia and St.-Petersburg Institute	no need to for features extraction	No feature selection	cross-validation (ten-fold)/SP/accuracy/SE	Two seconds durations of ECG signal segments: 94.95%-, and 5-seconds segments: 95.11% accuracy	-
16	2017	[86]	CAD using optimized SVM	the Long-Term ST & Normal Sinus Rhythm RR Interval PhysioNet	Linear, non-linear and frequency domains	The Principal Component Analysis (PCA)	cross-validation (ten-fold)/SP/accuracy/SE	Accuracy: 99.2% Sensitivity: 98.43% Specificity: 100%	-
17	2017	[31]	LS-SVM/CAD	St. Petersburg Institute of Cardiological and Fantasia	FAWT/CIP		cross-validation (ten-fold)/SP/accuracy/ t-test and Matthews Correlation Coefficient (MCC)/ specificity/sensitivity	Classification for Morlet: 99.60%, Radial Basis Function (RBF): 99.56% accuracy	Matlab

Abbreviations: Higher-Order Statistics and Spectra (HOS), k-Nearest Neighbors (KNN), Creighton University Ventricular Tachycardia Database (CUDB), Linear Discriminant Analysis (LDA), St.-Petersburg Institute of Cardiological Technics 12-lead arrhythmia database (INCART), Multi-Layer Perceptron (MLP), Radial Basis Functions (RBF), Support Vector Machines (SVM), Medical Imaging Technology Database (MITDB), the MIT-BIH Malignant Ventricular Arrhythmia Database (VFDB), Flexible Analytic Wavelet Transform (FAWT), Sensitivity (SE), MITBIH Supraventricular Arrhythmia Database (SVDB), Specificity (SP), Area Under the ROC Curve (AUC), the positive predictivity (PP), Heart rate variability (HRV), Random Forests (RF), Myocardial Infarction (MI), Convolutional Neural Network (CNN), Myocardial Infarction (MI)

References

[1] (24 May 2018). *The top 10 causes of death*. Available: http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] S. Patidar, R. B. Pachori, and U. R. Acharya, "Automated diagnosis of coronary artery disease using tunable-Q wavelet transform applied on heart rate signals," *Knowledge-Based Systems*, vol. 82, pp. 1-10, 2015.

[3] P. Libby and P. Theroux, "Pathophysiology of coronary artery disease," *Circulation*, vol. 111, pp. 3481-3488, 2005.

[4] S. Mendis, P. Puska, and B. Norrving, *Global atlas on cardiovascular disease prevention and control*: World Health Organization, 2011.

[5] I. Abubakar, T. Tillmann, and A. Banerjee, "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013," *Lancet*, vol. 385, pp. 117-171, 2015.

[6] V. Khatibi and G. A. Montazer, "A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment," *Expert Systems with Applications*, vol. 37, pp. 8536-8542, 2010.

[7] P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation*, vol. 97, pp. 1837-1847, 1998.

[8] S. Mittal, *Coronary heart disease in clinical practice*: Springer Science & Business Media, 2005.

[9] G. W. Barsness and D. R. Holmes, *Coronary Artery Disease: New Approaches Without Traditional Revascularization*: Springer Science & Business Media, 2011.

[10] M. Kivimäki, S. T. Nyberg, E. I. Fransson, K. Heikkilä, L. Alfredsson, A. Casini, *et al.*, "Associations of job strain and lifestyle risk factors with risk of coronary artery disease: a meta-analysis of individual participant data," *Canadian Medical Association Journal*, vol. 185, pp. 763-769, 2013.

[11] J. Barth, M. Schumacher, and C. Herrmann-Lingen, "Depression as a risk factor for mortality in patients with coronary heart disease: a meta-analysis," *Psychosomatic medicine*, vol. 66, pp. 802-813, 2004.

[12] A. M. Roest, E. J. Martens, P. de Jonge, and J. Denollet, "Anxiety and risk of incident coronary heart disease: a meta-analysis," *Journal of the American College of Cardiology*, vol. 56, pp. 38-46, 2010.

[13] L. E. Cahill, A. Pan, S. E. Chiuve, Q. Sun, W. C. Willett, F. B. Hu, *et al.*, "Fried-food consumption and risk of type 2 diabetes and coronary artery disease: a prospective study in 2 cohorts of US women and men–," *The American journal of clinical nutrition*, vol. 100, pp. 667-675, 2014.

[14] L. Dauchet, P. Amouyel, S. Hercberg, and J. Dallongeville, "Fruit and vegetable consumption and risk of coronary heart disease: a meta-analysis of cohort studies," *The*

Journal of nutrition, vol. 136, pp. 2588-2593, 2006.

[15] D. Lloyd-Jones, R. Adams, M. Carnethon, G. De Simone, T. B. Ferguson, K. Flegal, *et al.*, "Heart disease and stroke statistics—2009 update. A report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee," *Circulation*, 2008.

[16] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, p. 395, 2012.

[17] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, pp. 43-48, 2011.

[18] R. A. Castellino, "Computer aided detection (CAD): an overview," *Cancer Imaging*, vol. 5, p. 17, 2005.

[19] R. M. Rangayyan and N. P. Reddy, "Biomedical signal analysis: a case-study approach," *Annals of Biomedical Engineering*, vol. 30, pp. 983-983, 2002.

[20] D. L. Hunt, R. B. Haynes, S. E. Hanna, and K. Smith, "Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review," *Jama*, vol. 280, pp. 1339-1346, 1998.

[21] D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, M. López, I. Alvarez, F. Segovia, *et al.*, "Computer-aided diagnosis of Alzheimer's disease using support vector machines and classification trees," *Physics in Medicine & Biology*, vol. 55, p. 2807, 2010.

[22] R. Wu, W. Peters, and M. Morgan, "The next generation of clinical decision support: linking evidence to best practice," *Journal of healthcare information management: JHIM*, vol. 16, pp. 50-55, 2002.

[23] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success," *Bmj*, vol. 330, p. 765, 2005.

[24] K. Singh and A. Wright, "Clinical decision support," in *Clinical Informatics Study Guide*, ed: Springer, 2016, pp. 111-133.

[25] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, "Advances in knowledge discovery and data mining," 1996.

[26] E. Turban, J. E. Aronson, T.-P. Liang, and R. Sharda, *Decision Support and Business Intelligence Systems (8th Edition)*: Prentice-Hall, Inc., 2006.

[27] A. Sajadieh, V. Rasmussen, H. O. Hein, and J. F. Hansen, "Familial predisposition to premature heart attack and reduced heart rate variability," *American Journal of Cardiology*, vol. 92, pp. 234-236, 2003.

[28] J. M. Dekker, R. S. Crow, A. R. Folsom, P. J. Hannan, D. Liao, C. A. Swenne, *et al.*, "Low heart rate variability in a 2-minute rhythm strip predicts risk of coronary heart disease and mortality from several causes: the ARIC Study," *Circulation*, vol. 102, pp. 1239-1244, 2000.

[29] Z. Binici, M. R. Mouridsen, L. Køber, and A. Sajadieh, "Decreased nighttime heart rate variability is associated with increased stroke

risk," *Stroke*, vol. 42, pp. 3196-3201, 2011.

[30] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, pp. e215-e220, 2000.

[31] M. Kumar, R. B. Pachori, and U. R. Acharya, "Characterization of coronary artery disease using flexible analytic wavelet transform applied on ECG signals," *Biomedical signal processing and control*, vol. 31, pp. 301-308, 2017.

[32] U. R. Acharya, H. Fujita, O. S. Lih, Y. Hagiwara, J. H. Tan, and M. Adam, "Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network," *Information sciences*, vol. 405, pp. 81-90, 2017.

[33] I. Beraza and I. Romero, "Comparative study of algorithms for ECG segmentation," *Biomedical Signal Processing and Control*, vol. 34, pp. 166-173, 2017.

[34] *Feature Extraction*. Available: https://deepai.org/machine-learning-glossary-and-terms/feature-extraction

[35] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, *et al.*, "Application of high-dimensional feature selection: evaluation for genomic prediction in man," *Scientific reports*, vol. 5, p. 10312, 2015.

[36] S. Karpagachelvi, M. Arthanari, and M. Sivakumar, "ECG feature extraction techniques-a survey approach," *arXiv preprint arXiv:1005.0957*, 2010.

[37] U. R. Acharya, H. Fujita, V. K. Sudarshan, S. L. Oh, M. Adam, J. E. Koh, *et al.*, "Automated detection and localization of myocardial infarction using electrocardiogram: a comparative study of different leads," *Knowledge-Based Systems*, vol. 99, pp. 146-156, 2016.

[38] U. R. Acharya, H. Fujita, M. Adam, O. S. Lih, V. K. Sudarshan, T. J. Hong, *et al.*, "Automated characterization and classification of coronary artery disease and myocardial infarction by decomposition of ECG signals: A comparative study," *Information Sciences*, vol. 377, pp. 17-29, 2017.

[39] E. W. Ngai, Y. Hu, Y. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems*, vol. 50, pp. 559-569, 2011.

[40] W. Wiharto, H. Kusnanto, and H. Herianto, "System Diagnosis of Coronary Heart Disease Using a Combination of Dimensional Reduction and Data Mining Techniques: A Review," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 7, pp. 514-523, 2017.

[41] G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, "A machine learning system to improve heart failure patient assistance," *IEEE journal of biomedical and health informatics*, vol. 18, pp. 1750-1756, 2014.

[42] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An

integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," *Expert Systems with Applications,* vol. 68, pp. 163-172, 2017.

[43] M. P. McRae, B. Bozkurt, C. M. Ballantyne, X. Sanchez, N. Christodoulides, G. Simmons, *et al.*, "Cardiac ScoreCard: a diagnostic multivariate index assay system for predicting a spectrum of cardiovascular disease," *Expert Systems with Applications,* vol. 54, pp. 136-147, 2016.

[44] S. Bashir, U. Qamar, and F. H. Khan, "BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting," *Australasian physical & engineering sciences in medicine,* vol. 38, pp. 305-323, 2015.

[45] A. D. Dolatabadi, S. E. Z. Khadem, and B. M. Asl, "Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM," *Computer methods and programs in biomedicine,* vol. 138, pp. 117-126, 2017.

[46] U. R. Acharya, V. K. Sudarshan, J. E. Koh, R. J. Martis, J. H. Tan, S. L. Oh, *et al.*, "Application of higher-order spectra for the characterization of coronary artery disease using electrocardiogram signals," *Biomedical Signal Processing and Control,* vol. 31, pp. 31-43, 2017.

[47] Z. Masetic and A. Subasi, "Congestive heart failure detection using random forest classifier," *Computer methods and programs in biomedicine,* vol. 130, pp. 54-64, 2016.

[48] U. R. Acharya, H. Fujita, S. L. Oh, Y. Hagiwara, J. H. Tan, and M. Adam, "Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals," *Information Sciences,* vol. 415, pp. 190-198, 2017.