Machine Learning Techniques for Heart Disease Datasets: A Survey

Younas Khan[1], Usman Qamar[1], Nazish Yousaf[*1, 2], and Aimal Khan[1]

[1]Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering, National University of Sciences and Technology, Islamabad, Pakistan

[2]Department of Computer Sciences, University of Wah, Wah Cantt, Pakistan

younas.khan15@ce.ceme.edu.pk, usmanq@ceme.nust.edu.pk, nazish.yousaf15@ce.ceme.edu.pk, aimalkhan@ceme.nust.edu.pk

ABSTRACT

Heart Failure (HF) has been proven one of the leading causes of death that is why an accurate and timely prediction of HF risks is extremely essential. Clinical methods, for instance, angiography is the best and most effective way of diagnosing HF, however, studies show that it is not only costly but has side effects as well. Lately, machine learning techniques have been used for the stated purpose. This survey paper aims to present a systematic literature review based on 35 journal articles published since 2012, where state of the art machine learning classification techniques have been implemented on heart disease datasets. This study critically analyzes the selected papers and finds gaps in the existing literature and is assistive for researchers who intend to apply machine learning in medical domains, particularly on heart disease datasets. The survey finds out that the most popular classification techniques are Support Vector Machine, Neural Networks, and ensemble classifiers.

CCS Concepts

• Computing methodologies →Supervised learning by classification • Computing methodologies→ Neural networks.

Keywords

Heart failure; heart diseases; risk prediction; neural network; deep learning; machine learning; healthcare.

1. INTRODUCTION

Learning is the procedure of developing a model after knowledge is covered from data, while machine learning is the complex computation procedure of automatically recognizing patterns and intelligent decision making on the basis of trained data samples. Machine learning falls under the umbrella of artificial intelligence; it is the ability of a machine to learn from a large set of data and predict, cluster or classify similar but unseen or new data based on its learning or training. Some famous machine learning techniques include Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree, self-organization map, and k-means clustering etc. There are ensemble approaches as well which integrate the outcomes of individual classification techniques and produce an overall better performance.

ICMLC '19, February 22–24, 2019, Zhuhai, China

ACM ISBN 978-1-4503-6600-7/19/02…$15.00

DOI: https://doi.org/10.1145/3318299.3318343

One of the leading causes of death all around the world is heart failure, it is seen as a major disease in old, and middle ages. Coronary Artery Disease (CAD) particularly, is termed a widespread cardiovascular illness having high rates of mortality. Clinical diagnosis techniques, such as angiography, are recommended by physicians as they call it the best diagnosis for CAD. On the contrary, it is very costly and possesses side effects. In order to present alternatives to such clinical diagnosis, much research has been conducted in the field of machine learning [1]. Heart disease has gradually become a communal problem of public health around the globe, it is because of unawareness, unhealthy consumption, and poor lifestyle. Today, its accurate prediction and diagnosis are great challenges for hospitals and practitioners. The development in computing technologies has assisted healthcare facilities in the collection and storage of data for founding clinical decision making. Hospitals, located in many developed countries, collect and store patients' data in the digital and manageable form [36]. Coronary Heart Disease (CHD) is a proven significant public health problem not only in some parts of the world but globally. Algorithms that incorporate the evaluation of clinical biomarkers with numerous reputable conventional risk factors can assist physicians in predicting CHD and make clinical decision making easier and reliable [2].

Machine Learning (ML) is a process used to interpret datasets by using computers that acquire knowledge from experiences. ML for health-informatics has emerged as an interdisciplinary science of dealing with healthcare data using complex computational techniques [37]. More often than not the healthcare datasets are colossal, that is where data mining comes into play. Data mining transforms the huge sets of data into information which are later used to make better predictions and decisions [38].

There are numerous studies which have done a similar job i.e. to discover which ML techniques have been used for diagnosing heart disease. For instance [39], which investigates and compares various data mining classification procedures, and ensembles. ML techniques can ameliorate the results of cardiac arrest forecast, but, there is a further need for research to improve the prediction and generalization of ML techniques [40].

This survey presents research studies focusing on ML techniques applied to heart disease datasets and which are published since 2012. The study intends to find gaps in the existing literature and suggests that a general solution for various healthcare problems is needed to be proposed. There are a few studies which seem to have done a similar job, however, there is no work done

particularly on heart disease datasets. Hence, we feel the need of conducting thorough research to objectively and critically analyze past papers. This survey provides the current gaps in research and is helpful for researchers who want to apply machine learning and data sciences techniques for diagnosing heart failure.

Here is an overview of the paper: section 2 depicts the methodology of the paper, section 3 presents a summary of the related work, whereas section 4 demonstrates analysis, section 5 illustrates a critical analysis of the papers and discussion and finally section 6 describes the conclusion and future work.

2. RESEARCH METHODOLOGY

The collection and selection processes of papers have been done in a way which meets a predefined criteria. Four keywords have been used to narrow down the search process i.e. Heart failure, Heart Diseases, Risk Prediction, and Neural Network. Research papers published by Elsevier, IEEE, and ACM, only are considered. Figure 1 presents an overview of the papers selected from the stated publishers and the details of individual research studies for each publisher are given in Table 1.

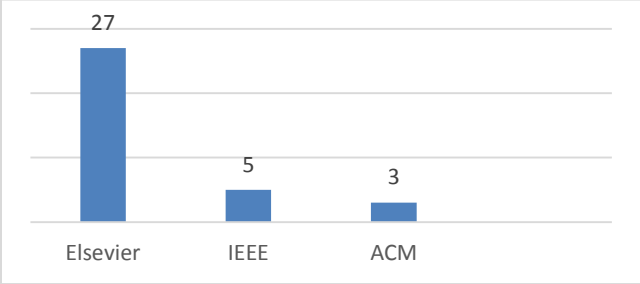

Figure 1. Selected researches per publisher

In order to ensure the quality of the survey, only state of the art papers and techniques are brought forth and only journal papers from 2012 to present have been selected for this survey as shown in ***错误!未找到引用源。***. At the end of the selection process, a total of 35 journal papers have been opted for.

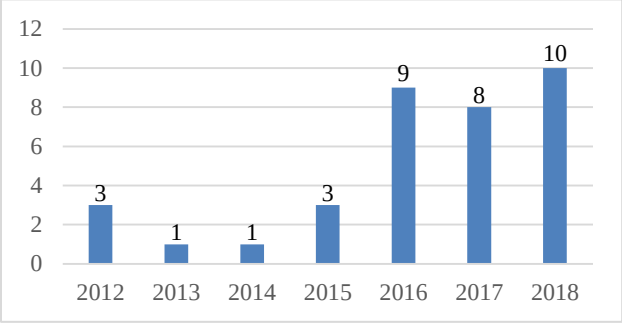

Figure 2. Selected researches per year

Papers which have presented an assessment of its technique and have presented validation by implementing it on at least one dataset are considered and those which do not meet this criterion are excluded. A thorough analysis of their methodologies and results has been conducted to make sure that this criterion is met.

Table 1. Details of Research Studies per Publisher

Publisher	Selected Research Studies	No. of Researches
Elsevier	[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [20] [21] [22] [23] [25] [26] [27] [28] [29]	27
IEEE	[24] [30] [31] [32] [35]	5
ACM	[19] [33] [34]	3
Total		*35*

3. RELATED WORK

In this section, the main findings of the selected papers have been discussed after analyzing them. It presents a brief of the methodologies brought forth in the selected papers. The research papers have been categorized on the basis of the classification techniques. Some of the selected papers have applied their techniques to various datasets, these datasets include, Diabetes datasets, Heart disease datasets, Breast Cancer datasets, Liver Disease and Hepatitis datasets. However, in this survey those papers have been selected which implement their techniques on heart disease datasets, for instance: UCI, PHP, BIDMC CHF dataset, PTB Diagnostic ECG and others.

The following techniques are found to be most popularly used for heart disease and HF classification.

3.1 Support Vector Machine (SVM)

Sung and Yuan Lee makes use of Genetic Algorithm (GA) for feature selection and SVM for classification. The results of this technique validate its effectiveness. When SVM is applied without GA, it outclasses two other techniques in literature by producing an accuracy of 96.38%, when GA is applied the accuracy is improved by 3.14% [1].

A.D. Dolatabadi et al., [2] proposes a technique for an automatic diagnosis of Coronary Artery Disease (CAD) and normal conditions by making use of Heart Rate Variability (HRV) signals that are derived from an electrocardiogram (ECG). The technique makes use of Principal Component Analysis (PCA) for dimensionality reduction and then applies SVM for classification purpose.

Zeinab A. et al., [5] presents a system for optimizing parameters and settings of multiple algorithms of HRV feature extraction, it selects the best subset of features and tunes SVM parameters at the same time for maximizing prediction performance.

A. Mustaqeem et al., [12] collects enough data to form its own dataset; accuracy and kappa statistics are two evaluation measures used to assess the performance of the proposed algorithm.

H. Fujita et al., [15] ranks the features, which are clinically significant and are obtained by numerous means, using their t-value and feeds them to classifiers like k-nearest neighbor (kNN), decision tree and SVM. The technique predicts Sudden Cardiac Death (SCD) one, two, three and four minutes prior to the SCD with an accuracy of 97.3%, 89.4%, 89.4%, and 94.7%, respectively.

U. Rajendra et al.,[19] feeds nonlinear features to these classifiers: Decision Tree (DT), kNN, and SVM, the mix of Discrete Wavelet Transform (DWT) and a nonlinear analysis of the ECG signals is empowered to predict SCD accuracies of 92.11% (kNN), 98.68% (SVM), 93.42% (kNN) and 92.11% (SVM) before one, two, three and four minutes before SCD occurrence, respectively.

Y. Zheng [22] employees Least Square-SVM (LS-SVM) in order to implement an intelligent diagnosis, an analysis of the results depict that the prediction accuracy of chronic heart failure is 95.39%.

The technique proposed in [25] aims to predict the severity of heart failure in patients and uses SVM with Gaussian Radial, and it solves an optimization issue by maximizing hyperplanes amongst the designed classes. Another study, [27] designs a system based on SVM and genetic algorithm, which is coupled with a 10 fold cross-validation technique for selecting features and optimizing parameters of the classifier.

3.2 Neural Networks

Zeinab A. et al., [3] proposes a highly accurate hybrid algorithm for diagnosing CAD. The proposed algorithm increases the performance of the neural network by around 10%; it assigns the initial weights via GA and then feeds it to an NN.

Roshan et al., [8] feeds the extracted features to two classification algorithms i.e. NN and LS-SVM, the higher accuracy is achieved by using the first approach with a blend of principal components of ECG beats with approximately 98.11% average accuracy.

Oluwarotimi et al., [10] consults a seasoned cardiac clinician to select 13 features and calculates their weights to train an ANN in order to predict the risks of HF. The technique is then applied to 297 HF patients. It predicts HF risks with an accuracy of 91.1%.

D. Tay et al., [17] introduces an algorithm that is inspired from neurons, state of the art testing is conducted on Honolulu Heart Program (HHP) dataset and its results are compared with SVM and Evolutionary Data-Conscious Artificial Immune Recognition System (EDC-AIRS). Results depict that the proposed algorithm outclasses both SVM and (EDC-AIRS).

Altan, G. et al., [18] focuses on diagnosing CHF and CAD with multilayer perceptron NN; its classification performance is state of the art. Leema, Khanna, and Kannan [23] proposes a Computer Aided Diagnostic system which makes use of ANN that is trained by Particle Swarm Optimization (PSO) and Gradient Descent Backpropagation in order to classify clinical datasets.

The techniques presented in [28] though is formed on Convolutional Neural Network, but can be distinguished from it as it proposes to use many options for segmentation, and a huge number of settings is tested.

A CNN based technique has been presented in [29], which has 2 convolution layers and 2 pooling layers, moreover, it makes use of LSTM (Long Short-Term Memory) i.e. an RNN but with a hidden memory layer and an output layer. The memory block in LSTM has 3 gates units which are called input, forget and output and a self-recurrent connecting neuron.

In [31], a neural network based technique for the prediction of heart failure, on patient's electronic medical data, has been presented. However, especially a one-hot encoding and word vectors have been employed for modeling and predicting HF with LSTM.

In [32] a novel technique has been presented which is based on CNN, the tool aims to classify healthy and pathological people by making use of an auditory sensor for FPGA (Field Programmable Gate Array, it is used to decompose audio in real time to frequency bands).

3.3 Decision Trees

Maryam T. et al., [4] uses a dataset of 1159 healthy, 405 negative angiography and 782 positive angiography participants, 10 variables out of 12 are entered into a Decision Tree and the model is able to identify the risk factors with an accuracy of 94%. Whereas, a Bagged Decision Tree ensemble is used by [11] for classifying two groups with a 5 fold cross-validation and a 20% holdout validation, yielding 98.1% and 99.5% respective accuracies.

K. Sudarshan [21] makes use of dual tree complex wavelets transform or DTCWT in order to automatically differentiate ECG signals CHF from normal. The proposed methodology has been tested on ECG segments of 2 seconds. The study uses five different techniques for feature ranking, which are fed to kNN, and decision trees for classification. The presented method has achieved an accuracy of 99.86%. Since the method is based on merely 2 seconds of ECG signals, the clinicians would have sufficient time for further investigations.

3.4 Rules Based, Rough Sets, and Fuzzy Logic

Purushottam, Kanak, and Richa [6] plans a framework for foreseeing the levels of risks of patients based on provided parameters. The major influence of this Rules Based algorithm is to assist a non-specialized physician in making the correct decision regarding the risk levels. While, Debabrata Pa et al., [7] helps in detecting CAD at an early stage, through rules formulation from doctors and a fuzzy expert system technique.

Nguyen, Phayung, and Herwig [9] makes use of fuzzy logic and presents a technique called interval type-2 fuzzy logic system (IT2FLS) which utilizes a blended learning procedure consisting of fuzzy c-mean clustering and parameters tuned in by Firefly and GA. Since there is high dimensionality involved the computation cost of this technique is high, however, an assessment of the experiments conducted by IT2FLS reveals that it outclasses other ML techniques, for instance, NB, SVM, and ANN.

A fuzzy clinical decision support system has been presented in [33], which has been founded on C5.0 decision tree for classifying CAD and healthy conditions. The proposed algorithm is able to achieve an accuracy of 90.50%. One of the key features delivered by the system is the automatic detection of fuzzy rules without any aid from experts.

The algorithm presented in [34] makes use of a mining method to conduct the diagnosis. The algorithm consists of three parts i.e. firstly a fuzzy membership function is built by utilizing statistical methods and medical guidelines, secondly a decision tree creates the rules and lastly, fuzzy inferencing is used to predict heart disease.

3.5 Regression Models

A. Mustaqeem et al., [14] proposes the application of classification algorithm called Contrast Pattern Aided LR (CPXR (Log)). It develops and validates prognostic risk models in order to forecast survival of one, two and five years in HF by implementing Electronic Health Records (EHRs).

Long, N. C., et al.,[25] proposes MulSLR or Multilinear Sparse Logistic Regression, which can be seen as an extension of Sparse-LR. It differs from conventional LR in a sense that it solves for K classification vectors instead of one. The convergence problem is tackled by block proximal descent approach. Another research [35] implements 5 machine learning approaches i.e. NN, SVM, fuzzy rules, CART (Classification and Regression Tree) and random forest. Best results are produced by CART and random forests.

3.6 kNN

For the sake of experimentation, in [15] the entire set of clinically important features is first ranked and then fed to numerous

classifiers including k-nearest neighbor (kNN), DT and SVM. Where for the given dataset SVM achieves the highest accuracy. While in [19], non-linear features are input to these two classifiers kNN, and SVM. The respective accuracies of the stated classifiers are 92.11% and 96.68%.

3.7 Random Forest

Zerina and Abdulhamit [20] examines five different classifiers namely C4.5, DT, kNN, SVM, ANN and Random Forest, where results reveal that random forest achieves the highest accuracy. ECG signals are obtained from BIDMC CHF and PTB Diagnostic ECG databases.

A random forest inspired technique called improved random forest has been proposed by F, Miao et al., [30]. They introduce a split rule and stopping threshold in order to recognize more accurate predictors which are used to differentiate between survivors and non-survivors.

3.8 Ensembles

Saba, Usman, and Farhan [13] presents an ensemble technique with Multilayer classification by making use of improved bagging, and weighting. The proposed model which is called HM-BagMoov is empowered to overcome the cons of conventional performance bottlenecks by using a seven heterogeneous-classifiers ensemble.

C.-H. Weng et al., [16] investigates the performance of various classifiers, including classifiers involved in an ensemble and solo classifiers. They conclude that NN ensembles can improve the ability of generalization of learning systems significantly by training a limited number of NNs and then integrating their results.

Raid et al., makes use of three classifiers i.e. ANN, LS-SVM, and NB to develop an ensemble framework. The proposed ensemble is implemented on a real series telehealth data of chronic heart disease patients, and the results depict that the ensemble yields incredible accuracy and can be used to reduce the risk of incorrect recommendations [24].

A two-level neural attention model over an RNN (Recurrent Neural Network) has been formed in the shape of RETAIN model in [26]. The ensemble gets a patient's history as an input and produces a binary prediction on whether the patient is going to have an onset heart failure or not.

4. ANALYSIS

In order to thoroughly analyze the research studies that have been selected, some parameters have been defined and all the selected papers are assessed against these parameters. These are the parameters that are used to evaluate and analyze the selected papers.

4.1 Research Problem

All the selected papers must implement its technique on a heart disease dataset, however, there is no restriction on the dataset's nature. The dataset can be signal images, numerical, or categorical.

4.2 Proposed Approach

The approaches proposed are of different natures, papers that implement individual classifiers are included and so are those that apply feature selection or reduction techniques before feeding it to the classifier. Moreover, ensemble frameworks are also included in this survey. Their proposed techniques have been depicted in Table 2.

4.3 Dataset

Heart disease datasets including UCI, PHP, BIDMC CHF dataset, and PTB Diagnostic ECG datasets, have been implemented in the selected papers. The proposed algorithms have been mainly applied to two types of datasets i.e. categorical or integer or real and ECG signals. The most common attributes in the former type are age, sex, chest pain, resting blood pressure, cholesterol, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, and slope of peak exercise etc. As far as the ECG signals are concerned they consist of features of two major domains i.e. time and frequency. The former includes mean i.e. mean of all the RR intervals, standard deviation RR, square of the mean of the sum of the difference between adjacent intervals etc., and the latter includes the power of the low and high-frequency bands etc. Moreover, nonlinear dynamics measures are varied according to the applied or proposed techniques.

4.4 Validation Results

This is one of the most significant points in the analysis of the papers because the domain of the papers revolves around saving life. Validation results i.e. accuracies have been given in Table 2, the accuracies given in the stated table are average accuracies and are given in percentages.

4.5 Future Work/Limitation

This section is going to play a vital role when it comes to assisting researchers in finding gaps in the existing literature. Due to the fact that HF is a severely life-threatening disease, the need of diagnosing it dire. This section of Table 2 contains adequate information about the possible future work and/or the limitation in each individual paper.

Table 2. ML techniques applied to Heart Disease Datasets

Source	Research Problem	Proposed Approach	Dataset	Nature of Data	Accuracy in %	Future Work/Limitations
[1]	A method for congestive heart failure recognition.	Genetic Algorithm with Support Vector Machine.	CHF2DB and Normal Sinus Rhythm	ECG signals	98.79	-
[2]	A technique for automatically diagnosing normal and Coronary Artery Disease conditions.	It extracts features from Heart Rate Variability signals and applies PCA and SVM on it.	Long term ST DB	ECG signals	99.2	The study uses very little data for its training..
[3]	A hybrid method for diagnosing Coronary Artery Disease has been presented.	Genetic Algorithm and neural network.	Z-Alizadeh Sani Dataset	Integer/Real	93.85	The number of attributes is way too many which may result in higher computation time.

[4]	A model for predicting coronary heart disease.	Decision trees.	A dataset of 2346 patients' records has been used in this study.	Integer/Real	94	The algorithm has not been applied to any other relevant dataset.
[5]	An optimization algorithm for the prediction of paroxysmal atrial fibrillation.	SVM and Heart Rate Variability.	Atrial Fibrillation Prediction Database	ECG signals	87.7	Future works include stretching and implementing the presented algorithm on other HRV research gaps, for instance, detecting sleep apnea.
[6]	Assisting unspecialized doctors in making a decision regarding heart disease risk level.	Rules-based classifier.	Cleveland	Categorical, Integer, Real	86.7	Various combinations of confidence, MinItemSets, and the threshold for the algorithm can be attempted.
[7]	A screening system for the early detection of Coronary Artery Disease.	Rules and fuzzy experts approach.	Advanced Medical Research Institute	Categorical, Integer, Real	84.2	The current system has been designed for one disease, however, the rule organization supports for an extension of multiple disease expert system.
[8]	The study classifies five types of ECG signals for predicting abnormal cardiac activity.	PCA.	MIT-BIH	ECG signals	98.11	-
[9]	A system for diagnosing heart disease has been proposed in this study.	An integration of rough sets based attribute reduction and interval type-2 fuzzy logic.	UCI and SPECTF	Categorical, Integer, Real	82.6	When it comes to training time and dealing with large records, the technique can be improved by utilizing Levy flights in moving strategy of Firefly.
[10]	This research works to diagnose the risk of heart failure.	Fuzzy analytic hierarchy process for computing global and an ANN classifier.	Cleveland Heart Disease Dataset	Categorical, Integer, Real	91.1	Automatic determination of the best number of hidden nodes and their links can still be seen as a challenge, it is evident from the fact that the ANN classifiers are trained numerously.
[11]	This paper aims to diagnose Congestive Heart Failure for evading life-threatening events.	A Probabilistic Symbol Pattern Recognition method has been deployed for detecting subjects in CHF, while an ensemble of bagged decision trees is used for classification.	PhysioNet database	ECG signals	98.8	Patient traits such as genetic risk factors, demographics, and co-morbidities may further increase the classification.
[12]	A hybrid technique which predicts disease and recommends medicines to cardiac patients.	It uses SVM, Random Forest (RF) and Multi-layer Perceptron for prediction and for the recommendation it uses a knowledge base.	POF Hospital	Categorical, Integer, Real	97.8	The model can be protracted to analyze the influence of features like age, gender and weight etc., moreover, smartphone or web-based interface can also be developed, for the recommender, in future.
[13]	An ensemble has been proposed for a medical decision support system.	HM-BagMoov: an ensemble of seven heterogeneous classifiers.	Cleveland	Categorical, Integer, Real	86.2	A publicly available is deemed to be developed for people who can make use of the application via the internet.
[14]	The research develops and validates	Contrast Pattern Aided Logistic Regression	EHR data Mayo Clinic	Electronic Health	91.4	-

	prognostic risk models for predicting 1, 2, and 5-year survival in HF.	integrated with loss functions.		Records		
[15]	The paper automatically classifies HRV signals to predict sudden cardiac deaths.	The student's t-test has been used for feature ranking while DT, kNN, and SVM are used for classification.	MIT-BIH and NormalSinus Rhythm	ECG signals	94.7	It is intended that the HRV signal analysis is incorporated in a novel ECG and the clinicians are warned 24 hours prior to the SCD.
[16]	This study proposes a method for disease prediction.	Ensemble mainly consisting of ANN.	UCI	Categorical, Integer, Real	85.31	An issue of overfitting might occur if the training dataset is high.
[17]	A technique has been proposed that perform clinical risk prediction i.e. it estimates the possibility of disease-risk faced by a patient.	A novel, brain-inspired, an algorithm called Artificial Neural Cell System for classification or ANCSc.	Honolulu Heart Program	ECG signals	73.6	The first limitation is that the technique i.e. ANCSc has been evaluated using a single risk prediction task, whereas, the second one is that there is a huge room of improvement in its accuracy.
[18]	This research is focused on the prediction of congestive heart failure and coronary artery disease.	Multilayer perceptron neural network.	Normal Sinus Rhythm and Long-term ST.	ECG signals	97.83	Future works include a real-time model for the early prediction of CHF and algorithms based on a genetic search for enhancing accuracy.
[19]	A newly integrated index for the diagnosis of Sudden Cardiac Death using ECG signals has been proposed and evaluated in this paper.	Discrete Wavelet Transform with SVM has been deployed to predict SCD two minutes before death.	MIT-BIH SCD Holter and Normal Sinus Rhythm	ECG signals	98.68	The study calculates an index for SCD which can be constituted into a tool for clinicians and physicians to empower them in SCD diagnosis.
[20]	Long-term ECG time series has been classified as normal and congestive heart failure.	Autoregressive Burg for feature extraction and for classification five methods have been implemented; amongst which random forest classifies with maximum accuracy.	BIDMC CHF and PTB Diagnostic ECG databases	ECG signals	100	-
[21]	Automatic prediction of Congestive Heart Failure.	Dual-tree complex wavelets transform.	MIT-BIH NSR, Fantasia, and BIDMC CHF	ECG signals	99.86	The method is planned to be evaluated on data with bigger subject pool size.
[22]	Computer-aided chronic heart failure diagnosis.	Least-Square SVM has been used in this technique.	A dataset of 152 heart sound samples has been compiled for this study.	ECG signals	95.39	The dataset is very limited, moreover, it does not contains heart murmurs samples.
[23]	Computer-aided classification of clinical datasets.	The ensemble of ANN, PSO, and Differential Evolution.	UCI	Categorical, Integer, Real	86.66	-
[24]	Supporting recommendations for heart patients	Least-Squares SVM, ANN, and Naïve Bayes.	Tunstall	ECG signals	94.83	Future works include the application of this technique on a more reliable and appropriate dataset.
[25]	The study works on an adaptive and context-aware decision-making system for Intensive Health Care provision.	The proposed algorithm is based on RBF SVM and LKF SVM.	Compiled their own dataset.	ECG signals	87.9	The articles do not say much about its dataset, and it should be evaluated on a bigger and reliable dataset.
[26]	Heart risk prediction.	This study proposes and evaluates RETAIN on an	Cerner Health Facts	Electronic Health	82	The study does not take into account the factors liable for

		enormous dataset for risk prediction.		Records		the varying accuracies among the datasets, which is stated to be their future work.
[27]	This article aims to recognize cardiac health on the basis of ECG signals.	SVM with Genetic Algorithm and 10 fold cross-validation.	MIT-BIH Arrhythmia	ECG signals	98.85	Future works include the development of a prototype for fetching ECG signals and this algorithm to diagnose heart problems.
[28]	The study aims to automatically classify heart sound for pathology detection.	Convolutional Neural Network.	UoC-murmur and PhysioNet-2016	PCG and ECG signals	84.6	This study lacks an insight analysis as far as the sub-band filtering processes are concerned.
[29]	Identification of coronary artery diseases ECG signals.	Long Short-Term Memory network and CNN.	PhysioNet database	ECG signals	99.85	Future works include the installation of the algorithm in portable devices so that healthcare personnel can diagnose CAD at the earliest.
[30]	Prediction of heart failure patients' mortality.	The algorithm is based on improved random survival forest.	Multi-parameter Intelligent Monitoring in Intensive Care	Categorical, Integer, Real	82.1	ECG, PPG, and BP variation may be merged with the input data to improve the performance of the algorithm.
[31]	Heart failure risk prediction.	Neural networks and Long Short-term Memory network.	-	Electronic Health Records	66.55	In future, expert knowledge can be incorporated into the system.
[32]	Recognition and classification of heart murmurs.	A novel convolutional neural network.	PhysioNet/CinC Challenge	EEG signals	97	-
[33]	Diagnosis of coronary artery diseases.	The presented model is founded on Random Forest, C5.0, and fuzzy modeling.	UCI	Categorical, Integer, Real	90.50	Future works include containing more features and generalizing the algorithm for other datasets.
[34]	Heart disease prediction	Fuzzy rules-based method.	Personal Health Record	Categorical, Integer, Real	69.22	In the future, an analysis of the proposed technique on a large scale dataset can be performed.
[35]	Analysis of heart patients data for severity evaluation and type prediction	Regression models called: Classification And Regression Tree	Cardiology Department at the St. Maria Nuova Hospital in Florence, Italy	Categorical, Integer, Real	84.7	The findings can be generalized with a larger sample size.

5. DISCUSSION

HF has been proven one of the leading causes of deaths worldwide, this is the main reason why accurate prediction of HF risks is extremely vital in order to prevent and treat it [10]. A timely diagnosis of CHF is critical for evading a life-endangering event [11]. Apart from timeliness, accuracy plays an extremely significant role in the medical domain as it is related to the life of a person. A great and extensive amount of research has already been carried out on disease classification and prediction using ML techniques. Conversely, an agreement is yet to be made about which technique or classifier is best suited for which datasets. However, it is proven that feature selection and reduction technique increases the accuracy and reliability of classifiers. Moreover, it is also clear that classifier ensembles have been proven to have improved classification accuracy [13].

There are certain aspects in healthcare problems which cannot be overlooked, e.g. time taken to execute the technique and computational complexity which is depends on the number of features, accuracy, and generalization. According to our analysis, some of the studies have been able to focus on avoiding the inclusion of redundant features. The use of state of the art techniques e.g. GA in [1], [17] and [27], weights by SVM in [2], information gain in [12], F-score in [13], for feature selection has helped in negating the problem of too many features and hence decreasing the time of execution and the computational complexity.

When we carry out an insight analysis of the papers, we realize that the curse of dimensionality has been ameliorated by some studies using advanced techniques of feature reduction. For example, PCA in [2], [3], [8], chaos firefly in [9], and minimum Redundancy maximum Relevance in [21] are extremely efficient for dimensionality reduction. On the contrary, we also realize that the question that whether feature reduction causes information loss or not has not been answered in most of the studies.

Furthermore, when it is a question of life, accuracy is the most important aspect and since machines are used to do that prediction, the margin of error ought to tend to zero. Some studies have focused on achieving the highest accuracies by proposing extremely reliable and efficient techniques, e.g. GA with SVM to achieve over 98% accuracy in [1], PCA and SVM to gain over 99% accuracy in [2], PCA for getting over 98% accuracy in [8], and an ensemble of bagged decision trees for attaining around 99% accuracy in [11], more such techniques and their accuracies are depicted in Table 2. What is alarming is that most of the techniques have been applied to limited datasets and a dire need of evaluating them on a large dataset still exists.

The last aspect too is a significant one, sometimes researchers commit the mistake of over-training their algorithm on one particular dataset. Which causes the problem of overfitting, where an algorithm performs outstandingly well on a particular dataset and fails when tested on an unseen dataset. In this research we find some studies which have tackled this problem, these studies include [13] and [21], where the algorithms are evaluated on 5 datasets each. The accuracies in these techniques can still be improved because the mean accuracy that has been achieved by [13] is 85.75% and the same achieved by [3] is 88.28%.

Figure 3 shows the use of different algorithms and techniques in the most recent researches.

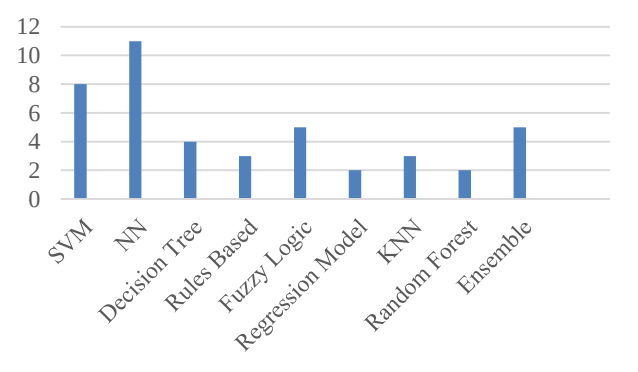

Figure 3. Techniques used in recent researches

6. CONCLUSION AND FUTURE WORK

This survey aims to explore, summarize, and critically analyze the most recent and state of the art research papers in order to find research gaps for future studies. This research has been conducted systematically, to help readers gain the knowledge of previous researches conducted in the domain of heart failure and risk detection. The limitations and future work provided in Table 2 can assist researchers in fulfilling the need for future research and gap in research. Whereas, the discussion section can help them direct their research.

In future, we intend to implement various classification techniques on heart datasets and compare and assess their performance in terms of accuracy, specificity, and sensitivity. Since, the use of ensemble increases generalization in many cases, we would choose the best techniques and present and implement a novel ensemble and test it on numerous datasets in order to evaluate its performance.

7. REFERENCES

[1] Yu, S. and Lee, M. 2012. Bispectral analysis and genetic algorithm for congestive heart failure recognition based on heart rate variability. *Computers in Biology and Medicine*. 42, 8 (2012), 816-825.

[2] Davari, D. A. et al. 2017. Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM. *Computer Methods and Programs in Biomedicine*. 138, (2017), 117-126.

[3] Arabasadi, Z. et al. 2017. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. *Computer Methods and Programs in Biomedicine*. 141, (2017), 19-26.

[4] Tayefi, M. et al. 2017. hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm. *Computer Methods and Programs in Biomedicine*. 141, (2017), 105-109.

[5] Boon, K. et al. 2018. Paroxysmal atrial fibrillation prediction based on HRV analysis and non-dominated sorting genetic algorithm III. *Computer Methods and Programs in Biomedicine*. 153, (2018), 171-184.

[6] Purushottam et al. 2016. Efficient Heart Disease Prediction System. *Procedia Computer Science*. 85, (2016), 962-969.

[7] Pal, D. et al. 2012. Fuzzy expert system approach for coronary artery disease screening using clinical parameters. *Knowledge-Based Systems*. 36, (2012), 162-174.

[8] Martis, R. et al. 2012. Application of principal component analysis to ECG signals for automated diagnosis of cardiac health. *Expert Systems with Applications*. 39, 14 (2012), 11792-11800.

[9] Long, N. et al. 2015. A highly accurate firefly based algorithm for heart disease prediction. Expert Systems with Applications. 42, 21 (2015), 8221-8231.

[10] Samuel, O. et al. 2016. An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. (2016), 163-172.

[11] Mahajan, R. et al. 2017. Improved detection of congestive heart failure via probabilistic symbolic pattern recognition and heart rate variability metrics. *International Journal of Medical Informatics*. 108, (2017), 55-63.

[12] Mustaqeem, A. et al. 2017. A statistical analysis based recommender model for heart disease patients. *International Journal of Medical Informatics*. 108, (2017), 134-145.

[13] Bashir, S. et al. 2016. IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework. *Journal of Biomedical Informatics*. 59, (2016), 185-200.

[14] Taslimitehrani, V. et al. 2016. Developing EHR-driven heart failure risk prediction models using CPXR(Log) with the probabilistic loss function. *Journal of Biomedical Informatics*. 60, (2016), 260-269.

[15] Fujita, H. et al. 2016. Sudden cardiac death (SCD) prediction based on nonlinear heart rate variability features and SCD index. *Applied Soft Computing*. 43, (2016), 510-519.

[16] Weng, C. et al. 2016. Disease prediction with different types of neural network classifiers. *Telematics and Informatics*. 33, 2 (2016), 277-292.

[17] Tay, D. et al. 2015. A novel neural-inspired learning algorithm with application to clinical risk prediction. *Journal of Biomedical Informatics*. 54, (2015), 305-314.

[18] Altan, G. et al. 2016. A new approach to early diagnosis of congestive heart failure disease by using Hilbert–Huang transform. *Computer Methods and Programs in Biomedicine*. 137, (2016), 23-34.

[19] Acharya, U. et al. 2015. An integrated index for detection of Sudden Cardiac Death using Discrete Wavelet Transform and nonlinear features. *Knowledge-Based Systems*. 83, (2015), 149-158.

[20] Masetic, Z. and Subasi, A. 2016. Congestive heart failure detection using random forest classifier. *Computer Methods and Programs in Biomedicine.* 130, (2016), 54-64.

[21] Sudarshan, V. et al. 2017. Automated diagnosis of congestive heart failure using dual tree complex wavelet transform and statistical features extracted from 2 s of ECG signals. *Computers in Biology and Medicine*. 83, (2017), 48-58.

[22] Y. Zheng, X. Guo, J. Qin and S. Xiao. 2018. Computer-assisted diagnosis for chronic heart failure by the analysis of their cardiac reserve and heart sound characteristics. 2018.

[23] Leema, N. et al. 2016. Neural network classifier optimization using Differential Evolution with Global Information and Back Propagation algorithm for clinical datasets. *Applied Soft Computing*. 49, (2016), 834-844.

[24] J. Zhang, R. Lafta, X. Tao, Y. Li, F. Chen, Y. Luo and X. Zhu. 2017. Coupling a Fast Fourier Transformation With a Machine Learning Ensemble Model to Support Recommendations for Heart Disease Patients in a Telehealth Environment. *IEEE Access*, vol. 5, pp. 10674-10685, 2017.

[25] M. Aborokbah, S. Al-Mutairi, A. Sangaiah and O. Samuel. 2018. Adaptive context aware decision computing paradigm for intensive health care delivery in smart cities—A case analysis. 2018.

[26] L. Rasmy, Y. Wu, N. Wang, X. Geng, W. Zheng, F. Wang, H. Wu, H. Xu and D. Zhi. 2018. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. *Journal of Biomedical Informatics*, vol. 84, pp. 11-16, 2018.

[27] P. Pławiak. 2018. Novel methodology of cardiac health recognition based on ECG signals and evolutionary-neural system. *Expert Systems with Applications*, vol. 92, pp. 334-349, 2018.

[28] B. Bozkurt, I. Germanakis and Y. Stylianou. 2018. A study of time-frequency features for CNN-based automatic heart sound classification for pathology detection. *Computers in Biology and Medicine*, vol. 100, pp. 132-143, 2018.

[29] J. Tan, Y. Hagiwara, W. Pang, I. Lim, S. Oh, M. Adam, R. Tan, M. Chen and U. Acharya. 2018. Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. *Computers in Biology and Medicine*, vol. 94, pp. 19-26, 2018.

[30] F. Miao, Y. Cai, Y. Zhang, X. Fan and Y. Li. Predictive Modeling of Hospital Mortality for Patients With Heart Failure by Using an Improved Random Survival Forest. *IEEE Access*, vol. 6, pp. 7244-7253, 2018.

[31] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin and X. Wei, "Predicting the Risk of Heart Failure With EHR Sequential Data Modeling. IEEE Access, vol. 6, pp. 9256-9261, 2018.

[32] J. Dominguez-Morales, A. Jimenez-Fernandez, M. Dominguez-Morales and G. Jimenez-Moreno, "Deep Neural Networks for the Recognition and Classification of Heart Murmurs Using Neuromorphic Auditory Sensors. *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 1, pp. 24-34, 2018.

[33] S. Mokeddem. 2017. A fuzzy classification model for myocardial infarction risk assessment. *Applied Intelligence*, 2017.

[34] J. Kim, J. Lee, D. Park, Y. Lim, Y. Lee and E. Jung. 2013. Adaptive mining prediction model for content recommendation to coronary heart disease patients. *Cluster Computing*, vol. 17, no. 3, pp. 881-891, 2013.

[35] G. Guidi, M. Pettenati, P. Melillo and E. Iadanza. 2014. A Machine Learning System to Improve Heart Failure Patient Assistance. *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 6, pp. 1750-1756, 2014.

[36] K. Raza. 2018. Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. *U-Healthcare Monitoring Systems*, vol. 1, pp. 179-196, 2018.

[37] J. Wassan, H. Wang and H. Zheng. 2018. Machine Learning in Bioinformatics. *Encyclopedia of Bioinformatics and Computational Biology*, vol. 1, pp. 300-308, 2018.

[38] M. Amin, Y. Chiam and K. Varathan. 2018. Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, 2018.

[39] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez. 2018. A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. *IEEE Computer Society*, 2018.

[40] S. Layeghian Javan, M. Sepehri and H. Aghajani. 2018. Toward analyzing and synthesizing previous research in early prediction of cardiac arrest using machine learning based on a multi-layered integrative framework. *Journal of Biomedical Informatics*, vol. 88, pp. 70-89, 2018.