2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395785

A Research Survey on State of the art Heart Disease Prediction Systems

Dr. Lakshmi Prasad Koyi
Professor
Department of CSIT
CIET,Gunture.
prasad.koyi@gmail.com

Tejaswi.Borra
Assistant Professor
Department of CSE
IAER,Hydrabad.
tej.519@gmail.com

Dr. G. Lakshmi Vara Prasad
Associate Professor
Department of CSE
CMRTC,Hydrabad.
glv.prasad19@gmail.com

Abstract— **Disease prediction systems are the better alternatives, to avoid the human errors in disease diagnosis and also assist in disease prevention with early detections. High-demand in preventing the rapidly increasing heart disease death tolls expanded the horizons of the former research scholars for introducing the intelligent heart disease prediction systems. Prediction of heart disease from patient's health record attributes is, a proven multi-dimensional decision-making system, which merely depends on mining attribute correlations too. Patient Health Records (PHR's) with structured categorical data and unstructured text/image data are the major input resources for heart disease prediction. Heart disease dataset preparation, prediction system's process flow design, process execution and results evaluation are the most common life cycle modules of any heart disease prediction system. Although many former research were introduced various heart disease prediction models, but they are still suffering from some common set of problems. Input dataset attributes modeling, attribute risk factor calculation, correlations mining; threshold determination and achieving the high accuracy in disease prediction are the major limitations of the existing heart disease prediction systems. As part of my research on designing intelligent heart disease prediction models, several research papers are analyzed and narrated that knowledge in a proper manner with detailed description. The main objective of this study is to represent the current scenario of heart disease prediction systems and their associated modules in brief.**

Keywords— Disease prediction systems, Machine Learning algorithms, Cleveland Heart disease dataset, Digitalt Health Records, Data mining Platforms and prediction metrics

I. INTRODUCTION

Heart Disease is the highly considerable human health problem and the statistics revealed that, this is the main reason behind more than 60% of the adult mortalities every year noticed by WHO[1]. Any mature obstacles in heart blood pumping process to the body organs can causes the high blood pressure, which further leads to flare-up the heart disease and may also to sudden death. Early prediction [7, 10] or sensing of this disease occurrence can help in preventing the people from its negative impacts. Some Bio-Medical tests, Electromagnetic sensors, digital X-ray scanners and Cardiac catheterization methods are the popular heart disease diagnosis methodologies in medical field today. According to the former medical researches [2, 3], different health conditions and symptoms can hint in early about the flare up of heart disease levels in patients. Age, Gender, Blood Pressure(BP), Chest Pain, Electrocardiogram (ECG), Blood Glucose Levels, Cholesterol Levels (both LDL-C & HDL-C), HbA1c, disease hereditary, stress, smoking and drinking are the considerable attributes while heart disease diagnosis process. Coronary Heart Disease, Cardiac Arrest, Hypertension, Myocardial Infarction (Heart Failure) are the frequently occurring prominent heart disease types.

Today the health care industry is almost dependent on physicians, to determine the disease prediction and severity analysis using different clinical reports information. As it is already known that, the manual decision-making systems in disease prediction are inaccurate and unreliable as they were generated from limited sources of human knowledge. In order to catch up the advantages of digitalization process, hospitals were started managing their patient's information in the format of Electronic Health Records (EHR's) [4]. To address the challenges of manual disease prediction systems, EHR based many automated and semi-automated approaches have been employed in the area disease prediction. Due to the need of research attention, recently heart disease occurrence probability calculation from electronic health records using intelligent disease prediction systems became a pioneered research area. Former research papers [10 - 14] introduced several architectures, methodologies, technologies and algorithms to diagnosis the heart disease occurrences. As part of my research on designing the accurate and reliable heart disease prediction systems, we are decided to explore the detailed information about the heart disease prediction systems in this paper. This comprehensive study on heart disease prediction systems will discuss about all the possible dimensions involved in disease prediction. Dataset description, former prediction models, machine learning algorithms, data mining tools, metrics and research challenges were described in this paper. The knowledge from this study is very useful for the beginners and the research scholars of this domain in understanding about the disease prediction systems and its relevant modules in a single attempt.

In the remaining part of this paper, Section-II describes the Celaveland Dataset information along with the machine learning algorithms, section-III presents the information about former disease prediction systems, Section-IV discusses popular machine learning tools & metrics, and finally the

Section - V is dedicated for highlighting the present research challenges in heart disease prediction models.

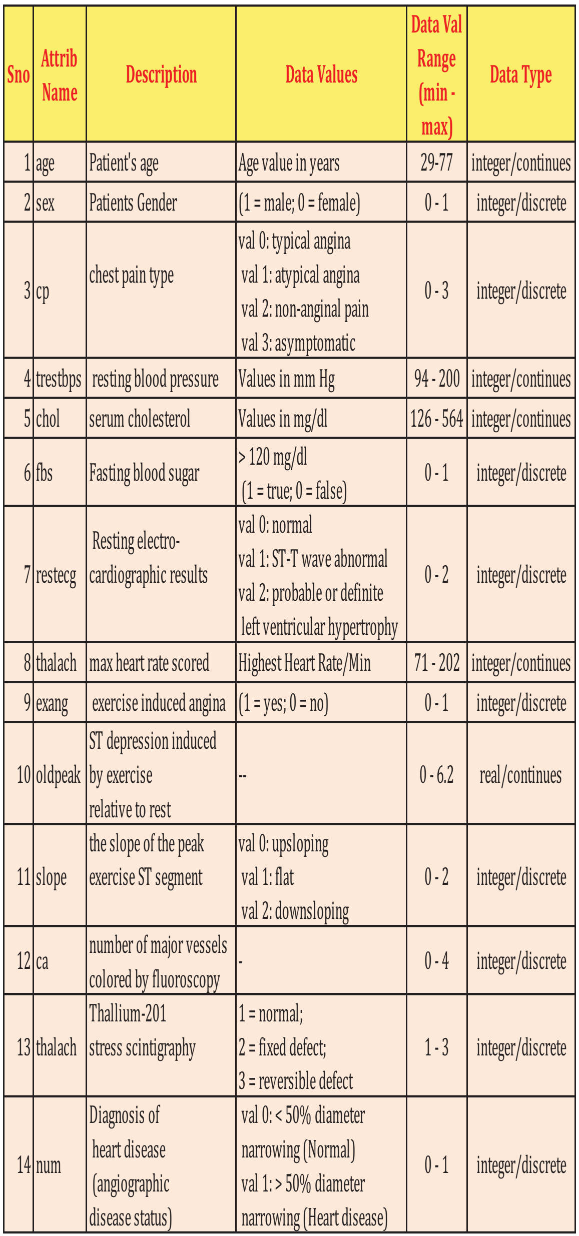
II. DATASET AND LEARNING MODELS

A. Dataset

First, Dataset is the collection of homogeneous data records. In this section, we explore the information about heart disease data sets and its compatible machine learning models for disease prediction.

The most popular and reliable heart disease data set is Cleveland heart disease data set [5] which is from the UCI machine learning repository. Most of the former prediction systems [7, 10, 13, and 16] either used their proprietary heart disease data sets or else utilized this Cleveland data set. This data set [5] is composed with total 303 heart disease data records and each record are having 76 disease relevant attributes. These multi-variant data set attributes are populated with integers, real numbers and categorical values. Among the 76 attributes, only the 14 heart disease relevance attributes (as shown in Table.1) have been chosen by the former heart disease prediction systems[13, 16, 17] are: age, sex, cp, trestbps, chol, fbs, restecg,thalach, exang, oldpeak, slope, ca, thal and num. As this dataset carries the desired output values (num=0 is normal & num=1 is heart disease) at each record level using the 'num' attribute, the same dataset can be partitioned for training and testing subsets, in case of supervised learning model.

However, the selected 14 attributes from the Cleveland dataset are participating in heart disease prediction, these attributes contained associative relations also playing a vital role. For example, Diabetes, blood pressure (BP) and age are the three independent variables with their respective (individual) weights in disease prediction, but when these three attributes became a set in disease prediction, than certainly this set or sub group impact will be useful in achieving the more result accuracy and system reliability. From the medical knowledge [6] also it is true that, the elder diabetic patient with high blood pressure is having the high chances of heart disease occurrence when compared to an elder patient with only high blood pressure. The reason behind this is, in diabetic the unprocessed glucose frequently stocks in blood vessels, which makes the blood vessels stiff and narrow to disrupt the normal blood flow. If we extracted such kind of attribute sets and their correlations with respective impact values that will improve the result accuracy of heart disease prediction system at great extent. Majority of the former prediction systems were not implemented, this efficient and comprehensive manner of attribute pattern mining method with relations extraction while disease prediction.



Sno	Attrib Name	Description	Data Values	Data Val Range (min - max)	Data Type
1	age	Patient's age	Age value in years	29-77	integer/continues
2	sex	Patients Gender	(1 = male; 0 = female)	0 - 1	integer/discrete
3	cp	chest pain type	val 0: typical angina val 1: atypical angina val 2: non-anginal pain val 3: asymptomatic	0 - 3	integer/discrete
4	trestbps	resting blood pressure	Values in mm Hg	94 - 200	integer/continues
5	chol	serum cholesterol	Values in mg/dl	126 - 564	integer/continues
6	fbs	Fasting blood sugar	> 120 mg/dl (1 = true; 0 = false)	0 - 1	integer/discrete
7	restecg	Resting electro-cardiographic results	val 0: normal val 1: ST-T wave abnormal val 2: probable or definite left ventricular hypertrophy	0 - 2	integer/discrete
8	thalach	max heart rate scored	Highest Heart Rate/Min	71 - 202	integer/continues
9	exang	exercise induced angina	(1 = yes; 0 = no)	0 - 1	integer/discrete
10	oldpeak	ST depression induced by exercise relative to rest	--	0 - 6.2	real/continues
11	slope	the slope of the peak exercise ST segment	val 0: upsloping val 1: flat val 2: downsloping	0 - 2	integer/discrete
12	ca	number of major vessels colored by fluoroscopy	-	0 - 4	integer/discrete

13	thalach	Thallium-201 stress scintigraphy	1 = normal;
			2 = fixed defect;
			3 = reversible defect
				1 - 3	integer/discrete

14	num	Diagnosis of heart disease (angiographic disease status)	val 0: < 50% diameter narrowing (Normal) val 1: > 50% diameter narrowing (Heart disease)	0 - 1	integer/discrete

Table-1 Cleveland Heart Disease Dataset Attributes Information

B. Learning Models

In order to predict the heart disease occurrence from medical datasets, relevant machine learning algorithms [8, 9] must be employed in main process of disease prediction. Supervised learning (means learning from training data) models and Unsupervised learning (learning from discovered patterns) models are the two different categories of machine learning. Support Vector Machine (SVM), Naïve Bayes and Nearest Neighbor (NN) are belonging to the classification type subcategory of supervised learning models, whereas Decision Trees, Linear Regression (LR) and some others are belongs to the regression type sub category of supervised learning. K-Means, Neural Networks, Gaussian Models and Hidden Markov models are belonging to the clustering category of unsupervised learning. The former heart disease prediction systems [11 - 15] widely used machine learning models are: Decision Trees, Naïve Bayes, K-Means, Artificial Neural Networks (ANN), Support Vector Machines (SVM) and logistic Regression (LR).To achieve the high accuracy from results, some hybrid prediction systems [17, 18] were introduced recently, by incorporating the two or more machine learning models together. In this section we describe the popular machine learning models, which are widely using in the domain of disease prediction in brief.

Naïve Bayes: This is a supervised learning algorithm [8, 9], which comes under the classification category of machine learning, to create the predictive models. This predictive model learns and creates the classifiers from the underlying datasets with desired output values. This is the fastest classification model, and simplifies the probabilities calculation using Bayes theorem [2 and 13]. This is an extremely suitable for the high volume datasets like heart disease set. Naïve Bayes considers each attribute of the heart disease dataset as an independent attribute, although the relations among the attributes were considered while decision making. Each attribute impact will be assigned as weight value and all attributes independent support is mainly considered while disease occurrence probability calculation. According to the Bayes theorem the posterior probability is calculated as illustrated:

$$P(c|x) = \frac{P(x|c) * P(c)}{P(x)} \qquad \text{----- (1)}$$

Here the P(c/x) is the expected posterior probability of target class ‘c’ and the attribute predictor ‘x’. The main advantage of the Naïve Bayes algorithm is, speedy in processing and well suitable for large datasets with limited training data. The most considerable limitation of Naïve Bayes (NB) in the view of HDPS is, NB assumes that each attribute of dataset is independent and considers only the attribute independent impact score while probability calculation. A.N Rapaka et al (2019) [35], proposed the “Smart Heart Disease Prediction” model using the naïve bayes to assess the prominent risk factors, which can cause to heart diseases.

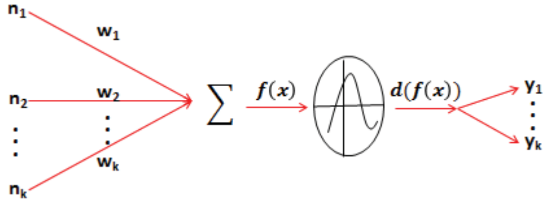
Decision Trees: This is another classification variant model of supervised learning [9], which can deal with both continuous and binary data attributes in an efficient manner. To perform the knowledge learning and data classification, it creates multiple identical subsets with different attribute combinations to analyze the attribute (or set) significance in decision making. Based on the target variable data type it employs either binary variable decision trees or continuous variable decision trees for classification. Gini Index, ID3, C4.5 and chi-square are the popular algorithms used to construct the decision trees. The ID3 algorithm based decision tree entropy calculation [7] with single attribute and multiple attributes are illustrated as shown:

$$E(s) = \textstyle\sum_{k=0}^{n} -P_k \log_2 P_k \quad \text{and}$$

$$E(T,X) = \sum_{k\in X} P(k)E(k) \qquad \text{----- (2)}$$

Irrespective of attribute data types, the decision trees can identify the high Impact attributes of dataset and supports the correlations mining among the attributes. P Ghosh et al (2021) [37] upgraded the decisions trees and the other machine learning algorithms using the bagging methods to achieve the high accuracy in disease predictions. Over-fitting problem [19] and weak in processing of continuous variables are the considerable limitations of decision trees.

Neural Networks: This machine learning model mimics the behavior of human brain and helps in learning from multi attribute data set. Neural networks [16] are the supervised regression model implementations with back propagation and forward propagation facilities. Single layer perception, multi-layer perception, Hopfield Networks and Radial-Basis function networks are the examples for different ANN subtypes. In order to process the multi attribute heart disease dataset, the multi-layered perception networks (subtype of ANN [20]) are useful. In this model the risk factor attributes of data set are given as input nodes from input layer, which are further mapped through the hidden layer with possible relations and finally the expected outputs are obtained from the output layer. The basic processing model of ANN’s single neuron mapping is illustrated below.



In above single neuron model, input node set N={n_1,n_2...n_k} is a set of risk factor nodes as input attributes with their associated weights from {w_1,w_2...w_k}. All these input nodes are mapped to the hidden layer sigmoid function f(x) and then forwarded to the propagated function d(f(x)) for regression model processing. y_1 to y_k are the expected outcomes from the neural networks.

S. Ambedkar et al (2018) [36], proposed convolutional neural network related CNN-UDRP algorithm to assess the risk with each patient record of the heart disease dataset. Their auto answering model even helps the patients with respective answers about the heart disease. V. S. Dehnavi et al (2020) [38], proposed the artificial neural networks-based classification techniques, which are trained based on genetic algorithm. Fuzzy sets were used in this model to optimize the non-linear parameters of the heart disease dataset. The combination of these models helped in obtaining considerable accuracy over the other machine learning models.

ANN’s are capable enough to deal with complex relations and can they support generalization especially while dealing with high volume of datasets without any restrictions. Apart from the benefits of ANN’s, they have some slight amount of disadvantages like slow convergence rate [21] of back propagation algorithm and blind allocation function for initial weights allocation.

Support Vector Machine: SVM [9, 22] is the prominent supervised classification model of machine learning today. SVM considers each attribute as an independent point of Ndimensional plot and the respective frontier values of each attribute helps in drawing the hyper planes for efficient classification. After trying to draw all possible hyper-planes the right plane will be selected to achieve the high accuracy from results. These hyper-planes are classified into two types are linear and non-linear. In case of non-linear plotting and classification tasks are more complex, when compared to the linear. The linear and non-linear (kernel trick) discriminant function models are illustrates as shown below:

$$f(x) = sgn\left(\sum_{i=1}^{n} \alpha_i y_i x_i^T x + b\right)$$

➔ Linear Discrimination ----- (3)

$$f(x) = sgn(\sum_{i=1}^{n} \alpha_i y_i K(x_i ,x) + b)$$

➔ Non-Linear Discrimination ----- (4)

In the above SVM discrimination process function, b is the offset value, x is the dataset value and K is the kernel trick sub function for non-linear discrimination. Finally, the sigmoid function will be executed for the count of support vectors n with target labels x and y_i. SVM supports regularization with L2 feature and efficiently tackle with non-linear datasets with stability in hyper-plane (decision making) design. SVM requires long time for training, while handling the high-volume data like our heart disease dataset [23] is the considerable limitation. J.S. Raj et al (2019) [39] applied the SVM learning model with RNN to obtain the high accuracy in data sequence predictions. Chandy Abraham et al (2019) [40], proposed the machine learning based resource leverage models, which even helps in leveraging the resources while processing the huge heart disease datasets on cloud.

All the above machine learning models are having their own advantages and limitations in processing of the real time datasets with different types of data. The perceived knowledge from analysis specifies that, the selection of learning model is dependent on various properties of dataset like volume, type, linearity, domain, relations and dependencies etc.

III. HDPS And Research Directions

In this section we describe about the intelligent heart disease prediction systems (HDPS) and their respective process models in detail.

Heart disease prediction systems are the subset of machine learning models, which are implemented in either supervised or unsupervised manner. Heart disease prediction models, initially selects a disease positive data set to train the classification model and this knowledge is used to predict the disease occurrence values from test data set. To accomplish this training and classification processes, different machine learning techniques like decision trees, Bayesian classifiers, K-Nearest Neighbor's, artificial neural networks, support vector machines and some other clustering and classification methods have been used. Dataset knowledge extraction, hidden patterns identification, attribute correlation mining, attribute impact values (weights) calculation and threshold determination are the main goals and activities of the heart disease prediction systems. Business requirements selection, raw data understating, data construction, process modeling, process evaluation and empirical studies are the common HDPS development lifecycle phases. Result accuracy, Precision, Recall, F-Measure, sensitivity and specificity are the important metrics to evaluate while experiments

As part of the research on HDPS, this section describes the simulations happened in the past by various authors [24, 25, 26 and 27] in brief with results accuracy values. To conduct the experiments with HDPS, the authors were used the Cleveland heart disease dataset, which is authenticated and available under open-source community license. Most of the algorithms were coded in python, MATLAB and java only. AMD 64 and Intel i5 to i7 range processors with maximum 8 GB ram was used for executing these operations. As the main expectation is the accuracy in disease prediction, the prediction systems were not much bothered about the execution speed. The dataset with 300 plus records were divided into the training dataset (70-75%) and the testing dataset (25-30%).

In 2008, Sellappan.P et al proposed the CRISP-DM methodology [24] for an intelligent heart disease prediction system, to classify the heart disease patients from clinical dataset. He had chosen the popular Cleveland Heart Disease database [5], which contains more than 600 heart disease data records with 15 disease relevant attributes. Among them half of the records were randomly selected for training model preparation and the others were used as test dataset for classification. CRISP-DM employed the Decision trees, Naïve Bayes and Neural networks as disease diagnosis group and for the data classification activity on Cleveland's test dataset. Empirical results of CRISP-DM methodology presenting that, the Decision trees recorded accuracy value is 80.4%, Neural networks recoded accuracy value is 85.68% and the Naïve Bayes recorded accuracy value is 86.12%. Medical profile based disease occurrence probability calculation, attribute impact in disease prediction and attributes correlations mining are the main goals to be achieved by this heart disease prediction system.

Like other prediction systems, this CRISP-DM also predicting the disease occurrence, but limited only to consider the categorical data. Another limitation is that, the CRISP-DM is using only the three machine learning models as part of disease diagnosis group, joining the other machine learning models may help in achieve more accuracy in results. As this dataset contains limited number records for training, the classification accuracy will be reduced while applying this training knowledge on real time data records.

Syed Umar Amin et al (2013) [25], proposed a hybrid neural network model with genetic algorithm weights for designing the heart disease prediction model. While the neural networks processing the underlying dataset, genetic algorithms were used to optimize the neural network weights instead of the common back propagation techniques. Using the genetic algorithm-based weights allocation process, the blind initial allocation problem and convergence problems of neural networks were addressed efficiently. Heart disease influencing or impacting attributes from dataset are extracted and presented as risk factors with risk value, which helps in specifying the importance of each attribute in disease prediction.

He setup the MATLAB environment for executing the proposed hybrid model with genetic algorithm and neural networks for disease prediction. To evaluate this hybrid, model the author chosen his own proprietary dataset with 50 records and each record with 12 attributes. A high-speed back propagation algorithm "Trainlm" is used for weights optimization based on Levenberg-Marquardt [21] proposed technique. The weights have been allocated to all participating attributes between -1.0 to +1.0. At input layer 12 neurons, at hidden layer 10 neurons and output layer 2 neurons total 152 process combinations were created for efficient processing. Among the 50 patient records 70% were used for training and learning, whereas remaining in 30% of records were shared equally for testing and validation process. Experimental results evaluated that the training dataset accuracy is 96.2%, test dataset accuracy is 92% and the validation set accuracy is 89%. Although this hybrid disease prediction method recorded high accuracy when compared to standalone neural networks, there are few limitations to be addressed.

The main issue with the genetic algorithm is allocating the unguided random weights initially, which leads to low convergence rate of this algorithm. As this research considered a small dataset with only 50 records, this leads to allocate unreliable weight values. Attributes correlations mining and the hidden pattern identifications were not implemented by this work. This research is considering only the structured categorical heart disease input data, but not any unstructured radio imaging or digital scan data.

In 2015, Ordonez et al [26] implemented the heart disease prediction system with constrained association rule mining to process their proprietary heart disease dataset. They strongly believe that mining the association rules and interdependency among the attributes of dataset, helps in allocating the suitable weights and classifies the data records with more accuracy. They had the proprietary medical data from a local hospital with 655 records and 25 attributes. To apply the association rule mining process on this medical dataset, the patient's medical records have been transformed into "the transaction like structures" using data transformation algorithms. Dissimilar to the previous machine learning approaches like Decision trees, Naïve Bayes and Neural Networks, this technique identifies the hidden relations among the attributes in the form of association rules. To control the high volume of association rules generation and to avoid the bottlenecks in, Ordonez et al [26] defined a default threshold based on association set size.

Using association rule mining process, each attribute level impact value and its relevant subset level impact value both can be calculated, but the attribute standalone impact value had the higher priority than the attribute subset level impact value. As part of experimental analysis they defined many rules with different attributes combination. Although this model can extract the hidden relations (in the form of association rules) among the attributes efficiently, it is also facing some limitations in association rules mining. Identifying the important rules for classification became a complex operation as many attributes with different impact values presented in dataset. Only association rules cannot result high accuracy in training as well as in classification, hence the hybrid approach with the coordination of machine learning techniques and association rule mining need to be implemented. Justifications should be implemented for increasing the reliability of this prediction model, while identifying the useful constraints and their impact calculation.

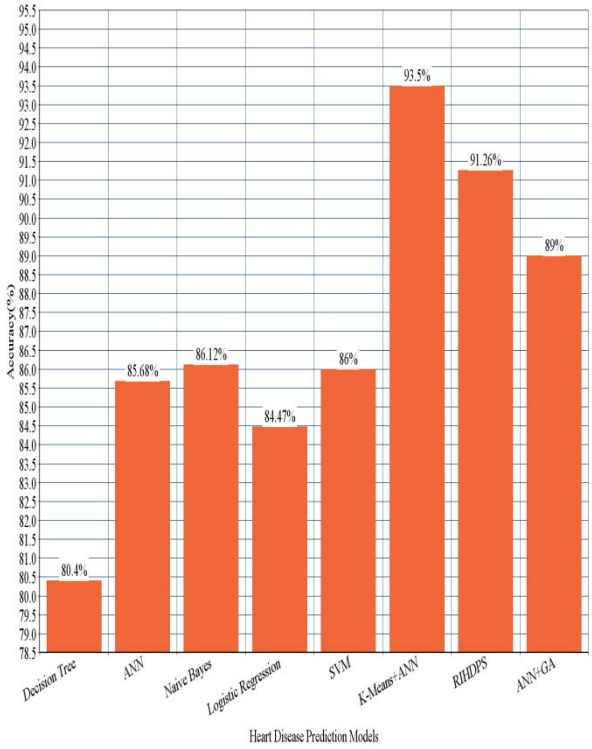
Rahman et al[27](2019) proposed a "Robust Intelligent Heart Disease Prediction System (RIHDPS)" and evaluated the experiments on Cleveland dataset by incorporating Neural Networks, Logistic Regression and Naïve Bayes like machine learning classification models. This research used the same Cleveland heart disease dataset with 14 attributes and 303 patient's health records. In these 303 records 66% records (with 42.5% disease [+ve] and 57.5% disease [-ve]) used as training dataset and the leftover 34% records (with 52.4% disease [+ve] and 47.6% disease [-ve]) are used as testing data. The empirical study on proposed RIHDPS calculated several metrics like precision, recall, F-Measure, True-False discovery matrix and the result accuracy rate to describe the performance. The proposed RIHDPS model accuracy rate is 91.26%, whereas the other individual machine learning models like Naïve Bayes accuracy value is 89.32%, Logistic Regression accuracy value is 84.47% and the neural networks accuracy value is 90.2%.

Due to the design of RIHDPS with three different machine learning processing models, it scored the high accuracy value is 91.26%, but in reality the time complexity and space complexity values of the proposed RIHDPS experiments is too high. In a pre-known disease result dataset like Cleveland [5], it is possible to record the high accuracy by designing the prediction system with a set of processing models, but the same prediction system will face the ambiguity issues while confirming the result accuracy as multiple process models were involved. This question will raise, because of three models are different in their nature of classification, hence one model may result a record as disease +ve record, whereas the others cannot. In this case "which classification model result is to be trusted?" will be the main research challenge.

In 2018 Amit Malav et al [28], designed a hybrid heart disease prediction model with K-Means clustering and Artificial Neural Network (ANN) classification. This research assumed that, the single machine learning model contained prediction systems are recorded less results accuracy compared to the hybrid machine learning models contained prediction systems. In this paper, the research scholars presented the detailed architecture of disease prediction systems with modulus like data understanding, data preparation, data processing and data evaluation. After preprocessing of in medical data set, K-Means algorithm is applied on Cleveland dataset for attribute vertical clustering. After the clustering process is completed, than each attribute vector from a cluster is extracted and the relevant weight is computed based on the range and impact of the attribute value from data set.

Later the Artificial Neural Network (ANN) classifier is used to identify the similarities and dependencies among the data set cluster attributes to create the inter-relevant subgroups. These subgroups are further classified with ANN in a recursive manner to train the learning models. To overcome the slow convergence rate of ANN's while propagation, the K-Means speed convergence mechanism is adjoined. With this hybrid approach they recorded the 93.5% of accuracy, which is higher than the former studies.

The below accuracy graph is the fairly simple model, to represent the accuracy values obtained from different machine learning models and from the aforementioned literature review statistics.



Graph-1 Heart Disease Prediction Models Accuracy Comparison

IV. TOOLS AND METRICS

In this section we planned to describe the prominent machine learning implementation tools with the respective coding technologies and the necessary metrics to be followed while conducting the experimental analysis.

A. WEKA

Waikato Environment for Knowledge Analysis (Weka) [29] is a GNU licensed java GUI software tool for machine learning models implementation. Weka is equipped with most of the machine learning algorithms and required graphical representation modules. The complete life cycle (SDLC) activities of a data mining application like data preparation, data preprocessing, main processing (clustering & classification), testing and evaluation can be implemented by this tool. Due to the java technology used in design of Weka, it became portable and supports the execution on all major operating systems like windows, ioS, Linux and Solaris etc. In 2016, Serdar et al [12] implemented different machine learning algorithms with weka tool, to predict the heart disease from health records.

B. Rapid Miner

It is another popular data mining platform [30], to run the predictive analysis on datasets efficiently. Today major data analytic organizations are using this Rapid Miner platform widely for text mining and deep learning activities. All popular data preprocessing techniques, mining algorithms, clustering techniques, classifiers and test simulations are integrated in the standalone Rapid Miner application. Moderated result optimization techniques and process statistics visualization environments are expanded the services of this tool to the great extent. The core libraries of this data mining tool also, written in java programming language only.

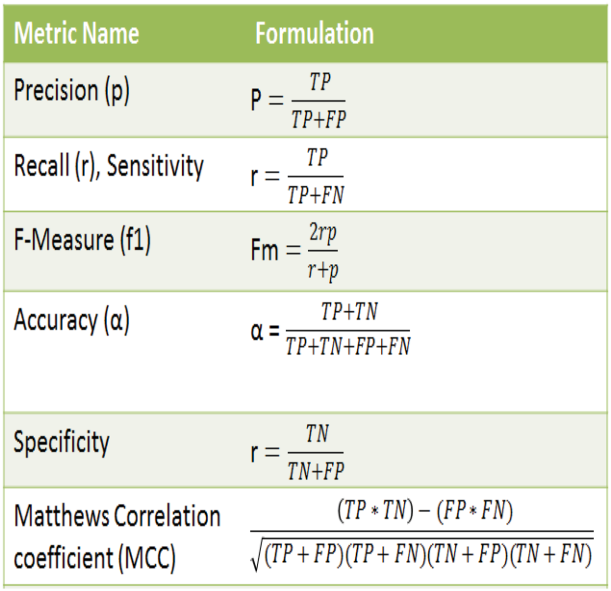
C. Apache Mahout

Apache Software Foundation designed this open-source machine learning platform with map reduce capabilities to support big data processing. From mathematicians to data scientists anybody can use this platform in a flexible manner to implement their proprietary algorithms with machine learning models. In order to deal with the high volumes of disease datasets, this mahout tool [31] from apache is a fine-grained and reliable solution as it builds on top of Hadoop libraries. Due to the advanced light weight processing models used in Mahout Tool, it reduces the application's data processing time and storage space requirements dramatically. But the main disadvantage of Mahout is, it needs a dataset with thousands of records for processing to record better performance, hence this is not compatible for small scale less volume datasets. K. Zolfaghar et al [32], used the Mahout 0.7 on Linux platform to predict the risk of readmission values from the real time heart disease dataset.

D. MATLAB

The "Math Works" organization designed Multi-Purpose data mining applications development environment is "MATLAB" [33]. Today the engineering research scholars are widely using this MATLAB tool to accomplish their research needs in machine learning. For all supervised classification & regression models development and for all unsupervised clustering models development sake this tool is most convenient. For coding and the custom application development with MATLAB developers use the compatible C, C#, Java and Python programming languages.

Measuring the performance with all possible metrics on machine learning experiments, helps in understanding the result progression in a statistical manner. As we mentioned in the beginning of this section, we would like to discuss about the metrics, which are relevant to calculate the performance of the heart disease prediction systems. In order to measure the progression of the experimental results of heart disease dataset, several research metrics have been proposed by many former research scholars [24, 27, 32 and 34]. The most prominent research metrics with respective formulas are presented with table.2.



Metric Name	Formulation
Precision (p)	$P = \frac{TP}{TP+FP}$
Recall (r), Sensitivity	$r = \frac{TP}{TP+FN}$
F-Measure (f1)	$Fm = \frac{2rp}{r+p}$
Accuracy (α)	$\alpha = \frac{TP+TN}{TP+TN+FP+FN}$
Specificity	$r = \frac{TN}{TN+FP}$
Matthews Correlation coefficient (MCC)	$\frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Table.2 HDPS Results Testing Associated Metrics

V. Research Challanges

Although many former research worked on designing of HDPS, they are still suffering from most considerable research challenges, which have to be addressed in future to make the HDPS more efficient and reliable. We are presenting those prominent research challenges in this section, which were discovered from our research analysis and literature review on HDPS.

We knew the fundamental slogan from healthcare industry is "prevention is better than cure". Most of the present HDPS systems [24, 27] are just implemented with the disease diagnosis process based on supervised training knowledge. As per our knowledge none of the former systems were tried to forecast or predict the disease occurrence chances. This means identifying the patients, whose data record risk attribute classification values are very close to the frontiers or thresholds of classification algorithms. If our HDPS is able to extract those most possible disease occurrence records from the heart disease negative (num=0) data set, it helps to identify the people, who may soon be affected by the heart disease. This forecasting process helps in suggesting such pre-heart disease profiles to doctors, to prevent them from future severe heart risks with early stage medications.

From the current research analysis on HDPS, we observed that the majority of HDPS [20, 26 and 27] are considering the risk factor attribute independent weight values while classification. We knew that the heart disease dataset attributes are logically dependent and having the correlations among them. For example, a person with high BP and smoking habit is having more chances to become heart patient in future. Like this many hidden relations and patterns among the attributes must be identified to classify the data records with more confidence and to achieve the high accuracy from results.

The hybrid disease prediction approaches [17, 18 and 28] are recording high accuracy in result prediction when compare to the standalone disease prediction models [13, 16 and 23] (as shown in Graph.1), by employing the different clustering and classification techniques. Finding the feasible machine learning models with compatibility, flexibility and interoperability to design the hybrid prediction systems is another major research challenge. Training from limited data records, Attribute weights allocation and adjustments, Data process model selection, Result optimization were the other limitations observed from past research works.

VI. Conclusion and Future Work

In this survey paper, we presented the detailed information about the existing heart disease prediction systems. As part of the survey, we discussed about the Cleveland dataset, machine learning algorithms, former heart disease prediction models, supportive tools, considerable metrics and research challenges etc. Former authors proposed disease prediction models are explored by emphasizing the advantages and limitations of them. Most reliable machine learning application development platforms like Weka, Rapid Miner, Mahout and MATLAB also covered in brief. The present heart disease prediction systems facing limitations are narrated as research challenges. The main objective of this study is to present the current scenario of heart disease prediction systems and their associated modules. In future, it is required to design the robust machine learning models to achieve the high prediction accuracy. For this the training, processing and classification models of a standard machine learning technology (i.e., ANN, SVM) must be upgraded with other efficient algorithms (i.e. Genetic Algorithm and Fuzzy logic). While processing the heart disease data for predictions, the attributes impact on results must be calculated individually to avoid the involvement of less impact attributes in processing to save time and to increase the accuracy.

References

[1] World Health Organization, Department of Health Statistics and Information Systems. Global Health Estimates: key figures and tables [Internet] Geneva: World Health Organization; 2016. http://www.who.int/healthinfo/global_burden_disease/en

[2] Miranda E, Irwansyah E, Amelga AY, Maribondang MM, Salim M. Detection of Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier. Healthc Inform Res. 2016;22(3):196–205. doi:10.4258/hir.2016.22.3.196.

[3] P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou et al., "Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association," Circulation, vol. 123, no. 8, pp. 933–944, 2011.

[4] Kalra et al "Electronic Health Records and Outpatient Cardiovascular Disease Care Delivery: Insights from the American College of Cardiology's PINNACLE India Quality Improvement Program (PIQIP)". Indian Heart Journal-2018.

[5] Blake, C.L., Mertz, C.J.: "UCI Machine Learning Databases", http://mlearn.ics.uci.edu/databases/heart-disease/, 2004.

[6] American Heart Association, http://www.heart.org/HEARTORG/Conditions

[7] Vikas Chaurasia, and Saurabh Pal, 2013, "Early Prediction of Heart Diseases Using Data Mining Techniques", Caribbean Journal of Science and Technology, ISSN: 0799-3757, Vol.1, pp. 208-218.

[8] Mehmed, K.: "Data mining: Concepts, Models, Methods and Algorithms", New Jersey: John Wiley, 2003.

[9] X. Wu and V. Kumar, Top 10 Algorithms in Data Mining, Springer, Berlin, Germany, 2007.

[10] A. Methaila, P. Kansal, H. Arya, and P. Kumar, "Early heart disease prediction using data mining techniques," in Proceedings of Computer Science & Information Technology (CCSIT-2014),vol. 24, pp. 53–59, Sydney, NSW, Australia, 2014.

[11] P. W. Wagacha, "Induction of decision trees," Foundations of Learning and Adaptive Systems, vol. 12, pp. 1–14, 2003.

[12] Das, R., Turkoglu, I., & Sengur, A., 2009, "Effective diagnosis of heart disease through neural networks ensembles", Expert systems with applications, 36(4), 7675-7680.

[13] Shadab Adam and Asma Parveen "Prediction System For Heart Disease Using Naive Bayes" International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp. 290-294.

[14] N.Cristianini and J.Shawe-Taylor "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods", Cambridge University Press, Cambridge, UK, 2000.

[15] K. Zolfaghar, N. Meadem, A. Teredesai, S. B. Roy, S. Chin and B. Muckian, "Big data solutions for predicting risk-of-readmission for congestive heart failure patients," 2013 IEEE International Conference on Big Data, Silicon Valley, CA, 2013, pp. 64-71.

[16] K. Vanisree and J. Singaraju, ''Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks,'' Int. J. Comput. Appl., vol. 19, no. 6, pp. 6–12, 2011.

[17] H. Yang and J. M. Garibaldi, ''A hybrid model for automatic identification of risk factors for heart disease,'' J. Biomed. Informatics, vol. 58, pp. S171–S182, Dec. 2015.

[18] G. Manogaran, R. Varatharajan, and M. K. Priyan, ''Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system,'' Multimedia Tools Appl., vol. 77, no. 4, pp. 4379–4399, 2018.

[19] J. Loughrey and P. Cunningham, ''Overfitting in wrapper-based feature subset selection: The harder you try the worse it gets,'' in Research and Development in Intelligent Systems XXI. London, U.K.: Springer, 2005, pp. 33–43.

[20] H. Yan, Y. Jiang, J. Zheng, C. Peng, and Q. Li, ''A multilayer perceptron-based medical decision support system for heart disease diagnosis,''Expert Syst. Appl., vol. 30, no. 2, pp. 272–281, 2006.

[21] Yu, Hao & Wilamowski, Bogdan. (2011). Levenberg–Marquardt Training. 10.1201/b10604-15. http://www.eng.auburn.edu/~wilambm/pap/2011/K10149_C012.pdf

[22] H.-L. Chen, B. Yang, J. Liu, and D. Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," Expert Systems with Applications, vol. 38, no. 7, pp. 9014–9022, 2011.

[23] A. D. Dolatabadi, S. E. Z. Khadem, and B. M. Asl, ''Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM,'' Comput. Methods Programs Biomed., vol. 138, pp. 117–126, Jan. 2017.

[24] Sellappan Palaniappan and Rafiah Awang "Intelligent Heart Disease Prediction System Using Data Mining Techniques" IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008.

[25] Syed Umar Amin, Kavita Agarwal, and Dr. Rizwan Beg "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors" Proceedings of 2013 IEEE Conference on Information and Communication Technologies - ICT 2013.

[26] C. Ordonez et al., "Mining constrained association rules to predict heart disease," Proceedings 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, 2001, pp. 433-440.

[27] M. J. Rahman, R. I. Sultan, F. Mahmud, A. Shawon and A. Khan, "Ensemble of Multiple Models For Robust Intelligent Heart Disease Prediction System," 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), Dhaka, Bangladesh, 2018, pp. 58-63.

[28] Malav, A & Kadam, K. (2018). A hybrid approach for Heart Disease Prediction using Artificial Neural Network and K-means. International Journal of Pure and Applied Mathematics. 118. 103-109.

[29] Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G. & Cunningham, S.J. (1999). Weka: Practical machine learning tools and techniques with Java implementations. (Working paper 99/11). Hamilton, New Zealand: University of Waikato, Department of Computer Science.

[30] Felix Jungermann "Information Extraction with RapidMiner", Artificial Intelligence Group, TU Dortmund, https://pdfs.semanticscholar.org/6309/62cb17efbdb9a9904d75dabc17e7ed8e2bfe.pdf

[31] V. R. Eluri, M. Ramesh, A. S. M. Al-Jabri and M. Jane, "A comparative study of various clustering techniques on big data sets using Apache Mahout," 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC), Muscat, 2016, pp. 1-4.

[32] K. Zolfaghar, N. Meadem, A. Teredesai, S. B. Roy, S. Chin and B. Muckian, "Big data solutions for predicting risk-of-readmission for congestive heart failure patients," 2013 IEEE International Conference on Big Data, Silicon Valley, CA, 2013, pp. 64-71.

[33] Löfberg, Johan. (2004). A toolbox for modeling and optimization in MATLAB. Proceedings of the CACSD Conference. 284 - 289. 10.1109/CACSD.2004.1393890.

[34] Serdar AYDIN, Meysam Ahanpanjeh, and Sogol Mohabbatiyan, February 2016, "Comparison and Evaluation of Data Mining Techniques in the Diagnosis of Heart Disease", International Journal on Computational Science & Applications (IJCSA), Vol. 6, No.1, pp. 1-15.

[35] A. N. Repaka, S. D. Ravikanti and R. G. Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 292-297, doi: 10.1109/ICOEI.2019.8862604.

[36] S. Ambekar and R. Phalnikar, "Disease Risk Prediction by Using Convolutional Neural Network," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697423.

[37] P. Ghosh et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," in IEEE Access, vol. 9, pp. 19304-19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[38] V. S. Dehnavi and M. Shafiee, "The risk prediction of heart disease by using neuro-fuzzy and improved GOA," 2020 11th International Conference on Information and Knowledge Technology (IKT), Tehran, Iran, 2020, pp. 127-131, doi: 10.1109/IKT51791.2020.9345630.

[39] Raj, J. S., & Ananthi, J. V. (2019). RECURRENT NEURAL NETWORKS AND NONLINEAR PREDICTION IN SUPPORT VECTOR MACHINES. Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40.

[40] Chandy, Abraham. "SMART RESOURCE USAGE PREDICTION USING CLOUD COMPUTING FOR MASSIVE DATA PROCESSING SYSTEMS." Journal of Information Technology 1, no. 02 (2019): 108-118.