International Journal of Computer Science & Communication (ISSN: 0973-7391)
Special Issue, Page 48-53, December 2020 ------ available at www.csjournals.com

A Survey on Machine Learning Techniques to Predict Heart Disease

Kavitha B S[1]
III Sem MTech
Department of Computer Science & Engineering
Sri Siddhartha Institute of Technology, Tumakuru

Dr. M Siddappa[2]
Professor and HOD
Department of Computer Science & Engineering
Sri Siddhartha Institute of Technology, Tumakuru

Abstract: Heart disease is one of the fatal diseases in the world. Every year 17.9 million people are died due to heart disease. It is a major cause of death. Early diagnosis of heart disease can save lives of many people. Early prediction of heart disease helps to get the treatment at proper time. We can also increase the life span of people and decrease the mortality rate. Heart disease can be predicted early by using machine
learning techniques. The machine learning algorithms such as Support vector machine, Decision tree, Naïve bayes, K-nearest neighbour, K means clustering, Logistic regression, Linear regression are commonly used to predict heart disease. In this paper, a survey of some research works identified with utilization of machine learning technique in prediction of heart related disease is anticipated.

Keywords- Machine learning, Supervised learning, Deep learning.

I. INTRODUCTION

The heart pumps blood through the bloodstream to all the cells in the body, carrying oxygen and nutrients. The abnormality of the coronary artery can cause decreased flow of oxygen and nutrients to the heart, leading to a heart attack and likely death.

The main objective of this paper is to summarize the different machine learning techniques used to predict heart disease and its percentage of accuracy in predicting the heart disease.

The symptoms of heart disease are fatigue, bloating, swollen legs, shortness of breath, chest pain, nausea, irregular heart rhythm etc. The risk factors that cause heart disease are age, unhealthy diet, alcohol, overweight, smoking etc. Healthy a known dataset to make predictions. The inferences from datasets consisting of input data without labelled responses are derived from the unsupervised model.

The logistic regression (LR), linear regression (LR), decision tree (DT), support vector machine (SVM), k-means clustering, k nearest neighbor, Naïve bayes (NB) are the different algorithms of machine learning.

The ML algorithms can handle large quantity of data which is produced by the real world. Its processing speed is high and makes real time predictions. By using machine learning methods, it is possible to predict heart disease. It helps clinicians/doctors to make early predictions of heart disease and also helps to save the lives of people.

The rest of the paper is organized as follows: Section II on literature survey, section III on comparative study of machine learning technique, section IV on general framework to predict heart disease, section V on conclusion, section VI on future work.

life style can help to prevent heart disease such as proper diet, doing regular exercise, stress free life.

The health department produces a large amount of information daily related to heart diseases and patients. The doctors and researchers do not use this data efficiently. Machine learning can efficiently analyse the data and can make decision.

Machine learning plays a vital role in predicting the heart disease. Machine learning is a subfield of artificial intelligence that focuses on the design of a model that can learn from experience and make decisions and predictions. Here the system is not programmed to carry out a particular task instead it learns by itself with the input data. When exposed to new data, these systems are designed to learn and evolve over time.

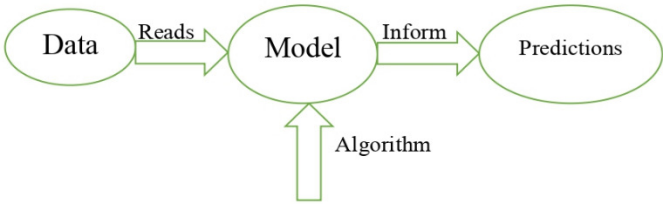

Fig. 1 Schematic diagram of basic ML model.

The ML algorithms learn from the data and classify between the normal and heart disease patient. The ML models can be categorized into supervised [10] and unsupervised models. The supervised model utilizes

II. LITERATURE SURVEY

Ankur Gupta et al. proposed machine intelligence framework (MIFH) to detect heart disease. The factor analysis of mixed data (FAMD) is used to select the features from UCI Cleveland dataset.

The data collected from Cleveland dataset contains some missing values. The data imputation was used to fill the new value with all the missing values of the features. The data standardization was used to convert the data into an appropriate format. The data is partitioned into training data and testing data. The derived features are given as input to themachine learning classifiers LR, KNN, RF, SVM and DT. The machine learning algorithm efficiency is checked by using performance metrics such as sensitivity, precision, accuracy, f-measure.

The results show that RF machine learning classifier along with FAMD gives an accuracy of 93.44%. The proposed MIFH can effectively classifies between the normal and heart patients.

Chieh-Chen Wu et al. proposed a scoring model that identifies major adverse cardiac events (MACE) in patients with chest pains. The two models were built by using invasive and non-invasive variables. The full risk stratification model developed by using both the variables. The reduced risk stratification model developed by using only non-invasive variables.

The data was taken from cohort of 1175 patients having chest pain. The ML models such as SVM, RF, ANN are used to derive the features. The LR was used to verify whether the selected features detect the development of MACE within 90 days. If the output of LR has a value greater than 1 indicates positive result and if the value is less than or equal to one indicates negative result.

The results show that both the full model and reduced model gives high performance and identify the development of MACE within 90 days. The score 0-3 indicates low risk of getting MACE over next 90 days and the score 4-6 indicates medium risk of getting MACE over next 90 days and the score 7-10 indicates high risk of getting MACE over next 90 days

Dakun lai et al. proposed a method to identify sudden cardiac death (SCD) [10] via arrhythmic risk markers. Sudden cardiac arrest (SCA) happens when heart experiences an arrhythmia that causes it to stop beating. This paper proposed a method to identify sudden cardiac death via arrhythmic risk markers which are taken directly from ECG signals. The risk markers are given as input to the machine learning classifiers KNN, DT, NB, SVM, RF which classifies into SCD and normal. It can notice SCD half hour earlier before the incidence of SCD.

In the first step, ECG data are collected and pre-processed. They are collected from three databases AHA, MIT-BIH, NSRDB. In the second step, vital ECG parameters are identified. In the third step, 5 arrhythmic markers are derived and calculated. Finally, these 5 arrhythmic markers are given as input to each classifier. The classifiers are validated by using performance metrics such as sensitivity, accuracy, specificity.

The results show that the classifier RF gives an accuracy of 99.49% compared to other models. It can effectively classify between SCD and normal.

body. This method uses both machine learning and deep learning technique. The heart sounds are recorded by using

Devansh Shah et al. studied various machine learning algorithms that accurately and effectively predicts the development of heart disease or not in a person.

The data is taken from the Cleveland database. The data is pre-processed and features are selected. For the training of ML models such as KNN, SVM, NB, DT and RF, these features are used. The ML model which gives highest accuracy is selected to predict the heart disease.

The result shows that K-nearest neighbour has got highest accuracy at k=7.

Avinash Golande et al. discussed about different machine learning algorithms and its efficiency in predicting the heart disease. The proposed model consists of data pre-processing, splitting and classification.

In the first stage, the data is pre-processed and the data which is not a number is converted into a number. The data is partitioned into training data and testing data. The training data is given as input to the 4 machine learning models such as KNN, K-mean, Adaboost, DT. The ML models learn from the training data and it effectively classify the test data into normal and heart disease. The algorithm with highest accuracy rate and lesser error rate can be used to classify between the normal and heart disease patient.

Jian Ping Li et al. proposed a new system to detect heart disease. The system is built by using machine learning algorithms SVM, KNN, NB, DT, LR, ANN. To choose appropriate features, feature selection algorithms such as least absolute shrinkage selection operator (LASSO) [12], relief [12], minimal redundancy maximal relevance (mRMR) [12] were used. They also proposed a new fast conditional mutual information (FCMIM) feature selection algorithm to select key features.

The dataset is taken from the Cleveland repository. The data is loaded for pre-processing [11]. The features are selected by using common feature selection algorithms. These features are given as input to the machine learning classifiers. The classifiers learn from these features. The output obtained from each classifier is recorded.

The FCMIM is used to select the features. These features are given as input to the machine learning classifiers. The performance of proposed FCMIM is compared with the standard feature selection algorithm. The machine learning classifier performance is validated by using performance metrics such as sensitivity, specificity, precision, accuracy, f-measure.

The results show that the machine learning classifier SVM with proposed FCMIM gives an accuracy of 92.37% as compared with other classifiers.

Martin Gjoreski et al. proposed a method to identify chronic heart failure (CHF) based on heart sounds. CHF is a disorder when heart is unable to circulate the blood throughout the phonocardiogram (PCG). Two heart sounds are normally recorded in a healthy person, additional heart sounds are also

recorded which is not considered as normal. This additional sounds indicates presence of heart disease.

The machine learning module extracts features from the raw PCG signal. The deep learning module works directly with the raw PCG signal. The ML module studies from the features defined by experts and uses opensmile tool to extract features. The sounds which are not related to heart were also present in the recording. These irrelevant sounds need to be removed. The DL consists of many layers. The output of one layer is given as input to the next layer. The output of both the modules DL and ML are combined by using recording-based ML. The RF algorithm is used to train the recording-based ML. The output of recording-based ML predicts whether the sound is from normal person or from heart patients and also it identifies different stages of CHF.

The results show that an accuracy of 93.2% is obtained by using above method.

Norma Latif Fitriyani et al**.** proposed Heart disease prediction model (HDPM) developed by using statlog and Cleveland datasets. The HDPM model consists of density-based spatial clustering of noise applications (DBSCAN), synthetic minority over sampling technique edited nearest neighbor (SMOTE-ENN), XGBOOST. The performance of HDPM model is evaluated by comparing with other machine learning models LR, NB, RF, DT and SVM. The results show that proposed HDPM model achieved an accuracy of 98.40 percent and 95.90 percent for Cleveland and statlog compared with other machine learning models.

The HDPM model is combined with Clinical decision support system (CDSS) to generate new model HDCDSS that helps doctors to identify patients with heart disease based on their current condition.

The heart disease datasets are collected from the freely available datasets Cleveland and statlog. The data are pre-processed and important features are extracted by using feature selection process.

The data are grouped by using DBSCAN. The data which does not belong to the group called outlier data are identified. The outlier data is excluded from the dataset for training. The training datasets are balanced by using SMOTE-ENN. The XGBoost machine learning algorithm is used to develop the HDPM. Using performance measures such as accuracy, precision, sensitivity, f-measure, false negative rate (FNR), false positive rate (FPR), true negative rate (TNR), MCC, HDPM performance is evaluated. The performance of HDPM is compared with all the ML models. The results show that HDPM achieved an accuracy of 98.40% and 95.90 % for Cleveland and statlog dataset repository.

The HDPM is integrated into CDSS and new model HDCDSS is created. The doctor sends patients data such as id, age etc along with diagnosis data to the webserver. The HDPM predicts the status of heart disease based on the input data and sends results back to doctors. The doctors can easily diagnose the heart disease status and take precautionary measures to save patient life.

Senthil Kumar Mohan et al. proposed a new prediction model to predict heart disease. The prediction model is developed by combining one or more techniques. This new technique termed as hybrid method. In this paper, they combined both Random forest and Linear method and proposed a new HRFLM method.

The main aim of this paper is to increase the accuracy of cardiac disease prediction. All the features can be selected by the suggested hybrid random forest and linear method (HRFLM). The range of features does not have any limitations.

The data for heart disease prediction is taken from Cleveland UCI repository. It contains 297 patient instances of which 45 records are used for testing and 252 records are used for training. The HRFLM method selects 13 features as input.

The pre-processing [11] of data is done after collecting 297 patients records. The best features are selected by using Decision tree entropy. The 13 features are given as input to the machine learning classifiers Random forest, Linear regression, Linear method, SVM, Decision trees, Naïve bayes, Neural network. The results are recorded for each classifier. Each machine learning classifier is validated by using confusion matrix. The accuracy, f-measure, precision, sensitivity and specificity are calculated for each classifier.

The results show that RF and LM method are giving high accuracy in predicting the heart disease.

The attributes age and sex are used for personal identification of patient. The remaining attributes contains a vital information related to heart disease prediction. If the value of num attribute contains a value ZERO indicates absence of heart disease. The value from ONE to FOUR indicates presence of heart disease. The value four indicates high risk of heart disease.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Special Issue, Page 48-53, December 2020 ------ available at www.csjournals.com

III. Comparative Study Of Different Machine Learning Techniques

Author [Reference]	Year	Proposed Method	Dataset	Technique	Accuracy	Attributes	Limitations
Ankur Gupta et al. [1]	2020	→Proposed machine intelligence framework to detect heart disease. →Uses factor analysis of mixed data (FAMD) to select the features.	UCI Cleveland	FAMD+ RF	93.44	14	→Unable to handle imbalanced class of data. →Unable to work with real-life situations in hospitals.
Chieh-Chenwu et al. [2]	2020	→Proposed a scoring model that identifies major adverse cardiac events (MACE) in patients. →Full risk stratification and reduced risk stratification model were built.	Cohort	LR	98.2	5	→Unable to deal with complex QTc-MACE relationships.
Dakun Lai et al. [3]	2019	→Proposed a method to identify a SCD via arrhythmic risk markers taken directly from ECG signals.	AHA, MIT-BIH, NSRDB	RF	99.49	5	→It works only with small data. →The risk markers used in this paper are in epidemiological state and are not used frequently.
Devansh shah et al. [4]	2020	→Studied various machine learning algorithms that predicts the development of heart disease.	UCI Cleveland	SVM, KNN DT NB RF	K-Nearest neighbour has got highest accuracy at K=7	14	→Association and clustering rules are not considered.
Avinash Golande et al. [5]	2019	→Discussed about different machine learning algorithms to detect heart disease.	-	KNN K-MEAN Adaboost DT	The algorithm with highest accuracy rate and lesser error rate can be used.	-	-
Jian Ping li et al. [6]	2020	→Proposed a new system FCMIM-SVM to detect cardiac disease. →The system is built by using SVM, KNN, NB, DT, LR, ANN. → The features are selected by using LASSO, relief and mRMR algorithms. →New feature selection algorithm FCMIM used to select key features.	UCI Cleveland	FCMIM-SVM	92.37	14	→Simple feature selection algorithm is used to select the features →more complex feature selection algorithm to be used to select features and to reduce the computation time.
Martin Gjoreski et al. [7]	2020	→Proposed a method to identify chronic heart failure (CHF) based on heart sounds. →The heart sounds are recorded by using PCG.	Physio-net UKC-JSI	RF	89-93.2	15	-
Norma Latif Fitriyani et al. [8]	2020	→Proposed HDPM model to predict heart disease. →The HDPM is integrated into CDSS and new model HDCDSS is developed that helps doctors to identify patients with heart disease based on their current condition.	Cleveland Statlog	DBSCAN SMOTEEN N XG BOOST	98.40 95.90	13+1 output class attribute.	→They have not taken feedback from cardiac experts. →It works only with small amount data.
Senthil Kumar Mohan et al. [9]	2019	→New prediction model HRFLM to predict heart diseases. →If the value of num attribute contains a value zero indicates absence of HD. The value from 1-4 indicates presence of HD.	UCI Cleveland repository	RF+LM	88.4	13	→Unable to work with real life data.

IV. General Framework to Predict Heart Disease.

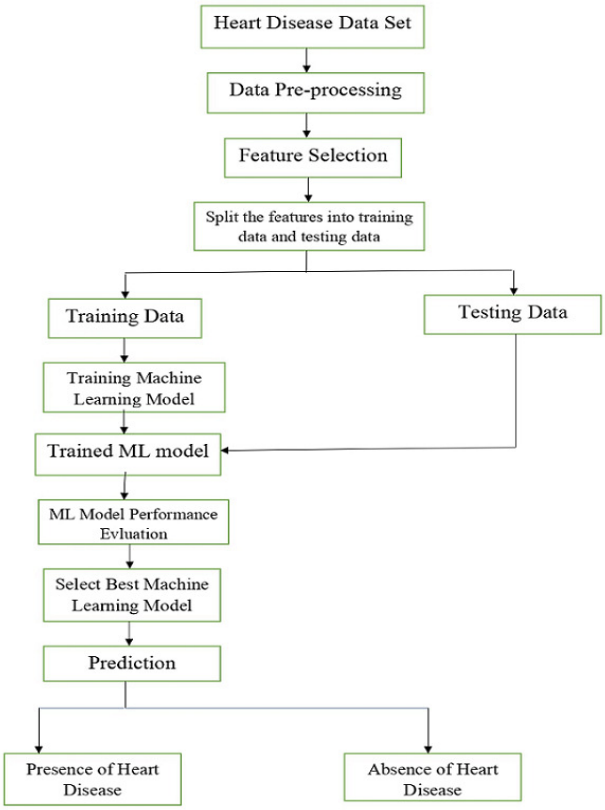

Fig. 2 The general framework to predict heart disease using machine learning technique.

A. *Dataset*

The dataset is a collection of data.

B. Data Pre-Processing

The raw data contains noises, missing values and in an incorrect format that cannot be used by the ML model. The real-world data must be pre-processed to remove noises and to replace the missing values with a new value and to convert the data into an appropriate format.

C. Feature Selection

The feature selection technique is a method used to select the key features from the dataset. The FS technique is used to reduce the number of input data, to reduce the computation time and to improve the performance of the ML model. Then the key features are extracted from the dataset. The features play a very important role in the prediction of heart disease.

D. Training data and Testing Data

The features are split into training data and testing data. The 80% of the features are used to train the ML model. The remaining 20% of the features are used to test the ML model. The ML models learn from the features and classifies the input data into normal and heart disease data.

E. Performance Evaluation

The ML model performance is evaluated using performance metrics such as accuracy, sensitivity, specificity etc. The best ML model which gives highest accuracy and lesser error rate is selected to classify between normal and heart disease patient.

V. Conclusion

From the study of the various research papers, it was found that most of the data taken from Cleveland repository which contains 303 examples and 14 attributes. Various machine learning classifiers like NB, SVM, KNN, DT, LR and K-nearest neighbour are used to build the heart disease prediction model. The dataset is classified into training data and test data. The 80% of the information is utilized for training purpose and the remaining 20% for testing purpose. The heart disease prediction model which is built by using machine learning techniques can efficiently predict the presence of heart disease and help the doctors to diagnose the disease accurately. It can also reduce the mortality rate by identifying the disease in the early stage.

Based on the survey, it was found that the RF algorithm showed highest accuracy as compared with other models. The RF algorithm reduces the overfitting problem and replaces the missing values in the dataset with the new values. The other algorithms such as SVM and LR performed well in some cases and have shown poor performances in some cases.

VI. Future Work

It was found that more focus was given on classification. In the future work both regression and association rules must be used to get highest accuracy in heart disease prediction and complex hybrid model must be built by utilizing both machine learning and data mining techniques to achieve high accuracy in heart disease prediction.

References

[1] G. Ankur , K. Rahul , Harkirat Singh Arora And Balasubramanian Raman, "*MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis," IEEE Access,* vol. 8, pp. 14659-14674, 2020.

[2] Chieh-Chen Wu, Wen-Ding Hsu2, Yao-Chin Wang, Woon-Man Kung, I-Shiang Tzeng, Chih-Wei Huang,

Chu-Ya Huang And Yu-Chuan Li1, *"An Innovative Scoring System for Predicting Major Adverse Cardiac Events in Patients With Chest Pain Based on Machine Learning," IEEE Access,* vol. 8, pp. 124076 -124083, 2020.

[3] L. Dakun , Z. Yifei , . Z. Xinshu, S. Ye And Md Belal Bin Heyat, *"An Automated Strategy for Early Risk Identification of Sudden Cardiac Death by Using Machine Learning Approach on Measurable Arrhythmic Risk Markers," IEEE Access,* vol. 7, pp. 94701-94716, 2019

[4] Devansh Shah, Samir Pate and Santosh Kumar Bharti, *"Heart Disease Prediction using Machine Learning Techniques," SN Computer Science,* pp. 1-6, 2020.

[5] A. Golande and P. K. T, *"Heart Disease Prediction Using Effective Machine Learning Techniques,"* IJRTE, vol. 8, no. 1S4, pp. 944-950, 2019.

[6] Jian Ping L, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan And Abdus Saboor, *"Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare,"* IEEE Access, vol. 8, pp. 107562-107582, 2020.

[7] Martin Gjoreski, Anton Gradišek, Borut Budna, Matjaž Gams And Gregor Poglajen, *"Machine Learning and End-to-End Deep Learning for the Detection of Chronic Heart Failure From Heart Sounds,"* IEEE Access, vol. 8, pp. 20313-20324, 2020.

[8] Norma Latif Fitriyani, Muhammad Syafrudin, Ganjar Alfian And Jongtae Rhee, *"HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System,"* IEEE Access, vol. 8, pp. 133034 -133050, 2020.

[9] M. Senthilkumar , T. Chandrasegar And S. Gautam , *"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,"* IEEE Access, vol. 7, pp. 81542-81554, 2019.

[10] Shakti Chourasiya and Suvrat Jain, *"A Study Review On Supervised Machine Learning Algorithms," (SSRG-IJCSE),* vol. 6, no. 8, 2019.

[11] Rajesh N, T Maneesha, Shaik Hafeez and Hari Krishna, *"Prediction of Heart Disease Using Machine Learning Algorithms,"* International Journal of Engineering & Technology, vol. 7, pp. 364-366, 2018.

[12] Amin Ul Haq , Jian Ping Li , Muhammad Hammad Memon, Shah Nazir and Ruinan Sun, *"A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms,"* pp. 2-21, 2018.