A Survey on Machine learning and Mining Techniques for heart disease prediction

[1]B.Vani , [2]B.Nagasri, [3]D.Priyanka

ABSTRACT--*It is important to save lives by detecting the heart disease earlier. Machine Learning then Data Mining is used as a aid contraptions by using providing the essential data and classification in accordance with diagnose a heart disease, primarily based on concerning the given input data. This survey paper analyse a* ***systematic*** *literature review based on journal articles published since 2012. This study significantly analyse the chosen papers and finds gaps between the current literature yet is helpful because researchers anybody want in accordance with apply machine learning algorithms among clinical domains, especially concerning heart disease datasets. This survey finds oversee that prediction exactness beyond most popular machine learning algorithms permanency like Random Forest, Decision Trees, and K-Nearest Neighbours.*

Keywords-- *Heart disease, Classification, Decision tree, Random Forest, Naive Bayes, K-Nearest Neighbours, Super Vector Machine.*

I. INTRODUCTION

In numerous nations like China, chronic diseases are the most important reason over loss of life then in the US; the regimen spends yearly around 2.7 trillion USD because of the remedy concerning chronic diseases. The continuous technological enhancement between Artificial intelligence then Machine learning helped the researchers in accordance with discover a modern methodology. The increase among health problems bear additionally leading to an extend between the technology over big data. Then because of using that big records to increase an automated computer-based system that perform stand used in imitation to predict heart diseases by implementing machine learning algorithms as choice conduct in accordance with work effectively because quite a number challenges as stability occur among the datasets.

In this paper, we analyse then estimate the makes use of of specific machine learning algorithms among the prediction about coronary heart disease through combining whole the attributes between the dataset after enhance the classification models. Random Forest, K-Nearest Neighbours, then Decision Tree classification models for coronary heart disease risk prediction are developed. These models execute effectively predict the danger over heart disease, according to complete whether or not an individual is rear under in accordance with suffer out of Coronary heart disease.

II. LITERATURE REVIEW

[1] *Assistant Professor(s), Department of Information TechnologySaveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai,b.vanirajan2004@gmail.com*
[2] *Assistant Professor(s), Department of Information TechnologySaveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai,* ***nagasrib1995@gmail.com***
[3] *Assistant Professor(s), Department of Information Technology*
Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai, ***priyavidya227@gmail.com***

Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020

Divya Krishnani, Anjali Kumari et.alPropose a pre-processing method to predict Coronary Heart Diseases (CHD). The method entails changing void values, resampling, standardization, normalization, classification, or prediction the usage of machine learning algorithms kind of Random Forest, Decision Trees, then K-Nearest Neighbours. It additionally compares this algorithms about the groundwork on prediction exactness attained. KNN 92.81% Random Forest is the almost combat adversary model because predicting yet gives the perfect performance measure. The accuracy, recall, precision, specificity yet F1 rating on RF concerning the proposed work are 96.71%, 98.74%, 94.4%, 99%, 96.61% respectively, beneath execution period over 1.3969 seconds.

Dr. Anooj P.K recommend a weighted fuzzy rule-based clinical decision support system (CDSS) The stability risk prediction over heart sufferers using it guide system consists about two phases, (1) computerized approach because technology about weighted fuzzy rules, then (2) increasing a fuzzy rule-based selection guide system. Utilizing accuracy, sensitivity then specificity the overall performance regarding the system is in contrast along the neural network. To discover the overall performance metrics, preceding such has according to calculate together with terms like, True positive, True negative, false negative then false high-quality stability

Nan Liu, Zhiping Lin et.al proposed an intelligent system according to give a lively risk measure as include three primary factors changeable selection, initial score calculation, then classification-based score updating. The calculation measurement is formed based totally over the estimate on characteristic vectors present out of the information on more than one patient.

Method	Sensitivity	Specificity	PPV	NPV
PCA	78.8%	80.8%	80.4%	79.2%
LDA	76.9%	80.8%	80.0%	77.8%
KPCA	75.0%	80.8%	79.6%	76.4%

It deploy three characteristic predicting methods, known as namely durability principal component analysis (PCA), linear discriminant analysis (LDA), then kernel PCA (KPCA) [26], then evaluated their rating metrics in accordance with the prediction performance.

Rong Tao , Shulin Zhang , Xiao Huang et.al. developed a ischemic heart disease detection methodology. T wave used to be separated out of Magnetocardiography (MCG) recordings then it services have been break up between three groups: period domain features, frequency domain features, then data concept features. It evaluate the precision about overall performance metrics about distinctive machine learning classifiers as k-nearest neighbour, decision tree, support vector machine (SVM), then XGBoost.

Decision Tree	88.31%
KNN	86.08%
SVM	88.84%
XGBoost	93.23%
SVM-XGBoostmixed model	94.03%

The SVM-XGBoost mix model achieves best performance of accuracy = 94.03%.

D. P. Shukla.et.al proposed a dimension that finds the association among the a number of attributes within a dataset. By the usage of the probability dimension this model offers a valid association rules. The dimension is utilized after each common and rare itemset in accordance with the dataset. This technique implies as Coronary Vascular Disease (CVD) risk is greater among male gender who are within the range over age within 55 & 65

Similarly in this technique because of prediction over heart attack measures are utilized to compute the data of every attributes frequency.

Buettner Aalen machine learning technique used in this paper is the Random Forests. Cross-validation was once used in accordance with attain an perfect result using the Random Forests algorithm. Cross validation divides the information employ into a unique range of subsets. In this approach every subset is used through repeating each as a training document then as a test record[15]. The random forest reached an average exactness of 84.448%. The approach is additionally examined without the 10 time cross-validation. The end result was exactness over 82.895%, which is 1.553% much less than ensuing algorithm together with a 10 times cross validation included.

Abderrahmane Ed-daoudy.et.al proposed a real-time heart disease prediction system primarily based concerning apache Spark. It is a significant scale platform because of distributed computing used for streaming information towards machine learning algorithms via in-memory computations. Using machine learning library MLlib, observations is made completely among individual node cluster primarily based concerning the computer-aided classification system deploying scala.This approach focuses about making use of real-time classification model because of constantly monitoring the patient's health using heart disease attributes.

MA.JABBAR.et.al focuses concerning analyses the usage of Association rule mining for Heart disease Prediction. It proposed recent algorithm known as Cluster Based Association Rule Mining Based on Sequence Number (CBARBSN) in imitation of mine association rules out of medical records based totally on number sequence then clustering[20]. The complete records base is divided in even sized partitions or known as namely cluster. Clusters are viewed only one at a time the first cluster is loaded in memory because calculating common records object sets. Then the second cluster is viewed in a similar way for calculating the common records item sets[22]. This technique reduces major memory capability due to the fact it considers small cluster at a period consequently that is scalable then efficient.

Amanda H. Gonsalves.et.al proposes ML methods engaged for the prediction concerning CHD are:Decision Tree – C4.5 (J48), Naive Bayes Algorithm, Support Vector Machine (SVM) DT is primarily based regarding the C4.5 algorithm [17]. The techniques break up the attribute by means of imposing the data acquire ratio impurity technique then organize the data into order at every node. The overall performance used to be analysed because the obtained methods using evaluation measures concerning Accuracy, Sensitivity, Specificity, TPs, TNs, FPs then FNs. NB performed the perfect accuracy amongst the three models.

Amin Ul Haq.et.al classifies humans including heart disease or healthy humans using specific machine learning predictive algorithms. The selection algorithms kind of Relief, mRMR, and LASSO had been used in accordance with choose essential features yet the overall performance concerning the classifiers have been examined on the selected features. The method over the proposed algorithm divided within 5 levels such as (1) preprocessing concerning dataset, (2) characteristic selection, (3) cross validation method, (4) machine learning classifiers, then (5) classifiers.

Method	Sensitivity	Specificity	accuracy
ANN Classifiers	73%	74%	74%
Decision Tree	68%	76%	70%
Random Forest	94%	70%	83%
SVM	78%	86%	86%

Naïve Bayes	78%	84%	84%

SVM the use of linear kernel has the good specificity , sensitivity , and accuracy . NB used to be the 2nd good classifier.

SENTHILKUMAR MOHAN.et.al proposes a prediction model including specific combinations about features and classification techniques. Hybrid random forest together with a linear model (HRFLM) produces a beneficial overall performance stage together with an accuracy concerning 88.7%. ML techniques begin from a pre-processing information phase, feature selection primarily based about DecisionTree entropy, classification on modelling accompanied with the aid of overall performance evaluation. The RandomForest error rate for dataset is excessive (20.9%) in contrast together with the other datasets. The LinearModel approach for the dataset is good (9.1%) in contrast together with DT and RF methods. We combine the RF approach together with LM and recommend HRFLM approach to enhance the results.

N Satyanandam.et.al compares the present classification, clustering and prediction models or executes an increased method within that domain. Decision Tree primarily based classification yet Bayesian classifications are base comparable accuracy however each will take longer instances because of classification. Similarly the neural network primarily based methods processing instances are much less in contrast to other approaches. The preliminary time committed for constructing the neural network model is absolutely high therefore that produces poor outcomes of time complexity. So, this work proposes some other familiar predictive optimization method according to preserve time complexity trade off at some point of accuracy.The proposed method OMLR (Optimal Multinomial Logistic Regression) algorithm gives 93.2% accuracy together with lesser time complexity than in contrst to other models.

Jyoti Soni.et.al introduced an sensible and effective heart attack prediction system with the help of Weighted Associative Classifier. It evaluates among terms over accuracy the overall performance concerning WAC. The experimental outcomes show that WAC is an effective approach because of the expansion regarding widespread patterns from the dataset for heart disease. In the form regarding Prediction policies the patterns are stored. The effectively over accuracy is measured namely 81.51% for WAC. The effectively concerning most accuracy is executed the use of the value 25% and confidence is 80%. The biased about Heart disease because of the patient can be anticipated using a GUI designed in accordance with run up the patient's data and the patterns saved into the rule base.

Table 1. Machine Learning approaches used for heart disease prediction

Source	Research Work	Proposed Method	Dataset	Result	Limitations	Future Work
1	Supervised machine learning algorithms to predict coronary heart disease	Random Forest, K-nearest neighbours, Decision Tree	Framingham Heart Study (FHS) dataset with 4240 records	exactness is greater and the time committed is absolutely less RF has produced much higher accuracy	dataset with much less attributes are used as training model	refining the preprocessing in addition will produce veracious outcomes.
2	clinical decision	Decision Tree,	Cleveland, Hungarian	clinical decision	The range of attributes	Bigger and real time dataset can be offered

	support system (CDSS)	Neural network	and Switzerland data sets from UCI machine learning repository	support system was improvised accordingly in terms of accuracy, sensitivity and specificity.	result in computation time	to attain a higher training model
3	Intelligent predicting system using HRV parameters	SVM used as classifier for Geometric distance based score prediction	November 2006 to December 2007. 1386 patients are monitored 1025 ECG datas are used from Department of Emergency Medicine, Singapore General Hospital	validations between balanced and imbalanced dataset results in achieving satisfactory prediction.	Consistency is required for samples in training as well as testing datasets	Biased variable subsets using supervised machine learning Algorithms to be implemented in imbalanced data.
4	Magnetocardiography-Based Ischemic Heart Disease Detection	K-nearest neighbours, Decision Tree, Super vectormachine, SVM XGBoost	The data of 347 healthy people and 227 people with ill patient are used	magnetic field map is used to exactly the heart patients with healthy people	MCG is sensitive to NSTEMI patients	magnetic pole pattern can be designed further more relevant
5	rule-based decision support system	Defining Association StrengthAlgorithm for Attribute Association	Medical Data set	find the association between frequency of symptoms and diseases for a critically ill patient	Contains only less attributes	large real time datasets can be used to predict the heart attack and comparison of proposed algorithm with related algorithms.
6	based on the information of test result and clinical data identify heart disease	Random Forest with & without 10 crosss validation	data sets from Cleveland Clinic Foundation	Maximum Heart Rate achieved, and the Resting Electrocardiographic Measurement seem to be less important	database with more attributes will increase the achieved level of accuracy	ECG sensor data can be used along with physiological sensor data for prediction
7	Detection using big data approach	Spark and Cassandra frameworks and random forest machine learning algorithm	Cleveland data of heart disease dataset from UCI repository	multiple data streams are simulated by sending real-time data to the spark cluster and generate more	TCP messages of data set attributes are used for system performance	big data technologies can be integrated for more efficiency

				than 500000 data streams/sec		
8	CLUSTER BASED ASSOCIATION RULE MINING	Cluster Based Association Rule Mining Based on Sequence Number (CBARBSN)	Clevand Heart Disease data set	Only single cluster is considered at a time reduces main memory capacity to be scalable and efficient	one cluster is considered each at a time	-
9	Prediction of Coronary Heart Disease	DT- C4.5 algorithm Naïve Bayes, SVM	KEEL Waikato South African Heart Disease dataset	performance was analysed based on Accuracy, Sensitivity, Specificity NB achieved the more accuracy between the three models.	Classifiers like SVM, DT and NB algorithms are alone used.	Sensitivity and Specificity rates can be increased for NB to predict effectively
10	Hybrid Intelligent System Framework	Evaluation of performance for classifiers are measured by metrics using Matthews' correlation coefficient and time taken for execution.	Cleveland heart disease dataset 2016.	Relief FS classifier performance is higher in contrast to mRMR and LASSO.	The feature selection algorithms is used to improve the classifiers.	other feature selection algorithms and optimization techniques can be implemented to improve classification accuracy and reduce execution time
11	Hybrid machine learning techniques	apriori, predictive and Tertius association rule mining used to find the reasons for heart disease	R Studio Rattle processes UCI Cleveland dataset.	Hybrid HRFLM accuracy proved with higher performance prediction of heart disease.	The accuracy is calculated based on number of feature selection algorithm	for broader perception a new model feature selection can be aopted
12	Predictive Optimization Techniques	Optimal Multi Nominal Logistics Regression algorithm	UCI heart disease datasets	OMLR algorithm produce 93.2% accuracy with lesser time complexity to increase the efficiency.	In OMLR algorithm only 15 attributes are considered for analysis and prediction of	to detect diseases like brain tumours and diabetes prediction the OMLR algorithm can be implemented

					heart disease	
13	Weighted Associative Classifiers	GUI based Interface using Weighted Association rule based Classifier	Cleveland Heart Disease database	using rule base classifier design a GUI which detect the presence of heart disease in patients	It uses only 303 patients records and 13 attributes from the data set	more number of patient records increased attributes is used to find new rule for the existing mining techniques

III. FUTURE WORK AND CONCLUSION

This Survey summarizes and aims to analyse critically about different machine learning and mining techniques used to predict heart disease at early stage. In this survey discussed about different machine learning approaches proposed by different authors with the accuracy of the algorithm for prediction. As a result in most of these proposed work Random Forest Tree as attained maximum accuracy compared other machine learning Algorithm.

In future some advanced machine learning algorithm like survival Random Forest, Bagging Algorithm and reinforcement algorithms can be used to predict the occurrence of heart diseases in terms of increased accuracy, Sensitivity, Specificity compared to traditional machine learning algorithms. In terms of performance evaluation some new methods can be implemented to show the accurate result.

REFERENCES

1. Divya Krishnani, Anjali Kumari, Akash Dewangan, Aditya Singh, Nenavath Srinivas Naik “Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms” 978-1-7281-1895-6/19/$31.00_c 2019 IEEE Trascations.
2. Dr. Anooj P.K. “ Clinical Decision Support System: Risk Level Prediction Of Heart Disease Using Decision Tree Fuzzy Rules” September 2012 ATC-60203031©Asian-Transactions.
3. Nan Liu, Zhiping Lin, Jiuwen Cao, Zhixiong Koh, Tongtong Zhang, Guang-Bin Huang, Marcus Eng Hock Ong “An Intelligent Scoring System and Its Application to Cardiac Arrest Prediction” IEEE Transactions On Information Technology In Biomedicine, Vol. 16, No. 6, November 2012
4. Rong Tao , Shulin Zhang , Xiao Huang, Minfang Tao, Jian Ma, Shixin Ma, Chaoxiang Zhang, Tongxin Zhang, Fakuan Tang, Jianping Lu, Chenxing Shen, and Xiaoming Xie “Magnetocardiography-Based Ischemic Heart Disease Detection and Localization Using Machine Learning Methods” IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 66, NO. 6, JUNE 2019.
5. Ricardo Buettner “Efficient machine learning based detection of heart disease” Proceedings: IEEE International Conference on E-health Networking, Application & Services, October 14-19, 2019
6. Abderrahmane Ed-daoudy, Khalil Maalmi “Real-time machine learning for early detection of heart disease using big data approach” 978-1-5386-7850-3/19/$31.00 ©2019 IEEE
7. Ma.Jabbar,2 Dr.Priti Chandra, 3b.L.Deekshatulu” Cluster Based Association Rule Mining For Heart Attack Prediction” Journal of Theoretical and Applied Information Technology 31st October 2011. Vol. 32No.2
8. Amanda H. Gonsalves “Prediction of Coronary Heart Disease using Machine Learning: An Experimental Analysis” ICDLT 2019, July 5–7, 2019, Xiamen, China © 2019 Association for Computing Machinery.
9. Amin Ul Haq “A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms” Hindawi Mobile Information Systems Volume 2018, Article ID 3860146, 21 pages.

10. Senthilkumar Mohan, Chandrasegar Thirumalai , And Gautam Srivastava “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” Ieee Access Special Section On Smart Caching, Communications, Computing And Cybersecurity For Information-Centric Internet Of Things.
11. N Satyanandam, Dr. Ch Satyanarayana, “Heart Disease Detection Using Predictive Optimization Techniques” I.J. Image, Graphics and Signal Processing, 2019, 9, 18-24 Published Online September 2019 in MECS.

12. Jyoti Soni, Uzma Ansari, Dipesh Sharma “Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers” International Journal on Computer Science and Engineering (IJCSE) Vol. 3 No. 6 June 2011.

13. 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.
14. 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.
15. 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.
16. 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.
17. Boon, K. et al. 2018. Paroxysmal atrial fibrillation prediction based on HRV analysis and non-dominated sorting genetic algorithm III. Computer Methods and Programs inBiomedicine. 153, (2018), 171-184.
18. Purushottam et al. 2016. Efficient Heart Disease Prediction System. Procedia Computer Science. 85, (2016), 962-969.
19. Pal, D. et al. 2012. Fuzzy expert system approach for coronary artery disease screening using clinical parameters. Knowledge-Based Systems. 36, (2012), 162-174.
20. Dr.M.G.Gireeshan,,”VEHICLE CONTROLLED BY MIND. EEG (ELECTROENCEPHALOGRAM)” International Journal of Pharmacy & Technology [ISSN: 0975-766X], IJPT| July-2015 | Vol. 7 | Issue No.1 | 8486-8489
21. 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.
22. Dr.M.G.Gireeshan,”FILM ANTIPIRACY SYSTEM” International Journal of Pharmacy & Technology [ISSN: 0975-766X], IJPT| July-2015 | Vol. 7 | Issue No.1 | 8468-8471
23. Long, N. et al. 2015. A highly accurate firefly based algorithm for heart disease prediction. Expert Systems with Applications. 42, 21 (2015), 8221-8231.
24. Samuel, O. et al. 2016. An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. (2016), 163-172.
25. 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.
26. Mustaqeem, A. et al. 2017. A statistical analysis based recommender model for heart disease patients. International Journal of Medical Informatics. 108, (2017), 134-145.
27. 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.