A Survey on Cardiovascular Prediction using Variant Machine learning

*L.*Chandrika[1,*], *Dr. K.*Madhavi[2], *B.*Sindhuja[3], *M.*Arshi[4]

[1]PG Student, Computer Science and Engineering, GRIET, Hyderabad, Telangana, India.
[2]Professor, Computer Science and Engineering, GRIET, Hyderabad, Telangana,India.
[3]Assistant Professor, Computer Science and Engineering, GRIET, Hyderabad, Telangana, India.
[4]Assistant Professor, Computer Science and Engineering, GRIET, Hyderabad, Telangana, India.

Abstract. Prediction of a cardiovascular diseases has always a tedious challenge for doctors and medical practitioners. Most of the practitioners and hospital staff offers expensive medication, care and surgeries to treat the cardiovascular patients. At early-stage of prediction of heart-oriented problems will be giving a chance of survival by taking necessary precautions. Over the years there are different types of methodologies were proposed to predict the cardiovascular diseases one of the best methodologies is a Machine learning approach. These years many scientific advancements take place in the Artificial Intelligence, Machine learning, and Deep learning which gives an extra push up to help and implement the path in the field of medical image processing and medical data analysis. By using the enormous dataset from various medical experts used to help the researchers to predict the coronary problems prior to happening. Many researchers have tried and implemented different machine learning algorithms to automate the prediction analysis using the enormous number of datasets. There are numerous algorithms and procedures to predict the cardiovascular diseases and accessible to be specific Classification methods including Artificial Neural Networks (AI), Decision tree (DT), Support vector machine (SVM), Genetic algorithm (GA), Neural network (NN), Naive Bayes (NB) and Clustering algorithms like K-NN. A few examinations have been done for creating expectation models utilizing singular procedures and additionally concatenating at least two strategies. This paper gives a speedy and simple survey and knowledge of approachable prediction models using different researchers work from 2004 to 2019. The examination indicates the precision of individual experiments done by various researchers.

1 Introduction

The heart is a significant organ which plays crucial part of every living being especially in humans. It pumps blood to all parts of our life systems. In case of failing to pumpor doesn't work accurately, the mind and different organs will quit working, and inside couple of moments, the individual will pass on. Changes in way of life, business related pressure, and terrible food propensities add to the expansion in the pace of a few heart-related sicknesses.

Heart sicknesses have arisen as perhaps the most unmistakable reasons for death all around the planet. The expanding populace in heftiness and smoking, the mortality from coronary sicknesses is slowly on the ascent, which has gotten the "best executioner" that compromises human wellbeing contrasted with malignant growth, Helps, and different illnesses, whatever age, character or area. As per WHO Coronary-related issues are responsible for the death of 17.7 million people every year, around 31% of worldwide death mortality.Especially in India, coronary illness became primary cause of mortality. Coronary problems causes the death of 1.7 million in 2016, In 2016 Global Burden of Disease Report, issued on 15th September 2017.

Coronary illness analytics states that the expenditure on hospitalization and treatment is gradually increased compared to the previous years and also diminishesthe survival rate of a person. Evaluations of WHO, explains that in India cardiac patients has spent around $237 billion, in the span of 10 years from 2005-2015. In this way, attainable and precise expectation of coronary related illness is vital.

The health care industry and health care research organisations collects the abundant data of patients,in which DM techniques are not used. The clinical data of the patients have covered up designs that are fundamental for data examination in the location of coronary illness. Coronary illness is a main source of death worldwide for as far back as 15 years. As the heart pumps the blood there is a possibility of on and off condition comes when the blood circulates inside the body which is inadequate, the rest of the organs inside the patient body especially brain and heart stops working, which causes the sudden death in few seconds. The importantelements are distinguished as age, hypertension, diabetes, family ancestry, tobacco, smoking, very high levels of cholesterol, daily routine of alcohol, actual dormancy, stoutness, chest torment, and eating junk food routine[1].

Data gathered by inspecting the verified records of the inpatient.it is feasible to separate the details and

* Corresponding author: lingalachandrika97@gmail.com

provide a report on CVD even though it is negative or positive. These segments guide us to distinguish CVD. Determination of the reports is basically taken on the patient's "Echocardiography (ECHO)" and "Electrocardiogram (ECG)" tested results done by experienced doctors.

ECG uses synthetic plugs which fix to the patient's body to record the heartbeat based on electrical movement. This heartbeat electrical waves extracted by depolarizing, for every heartbeat is represented as the electrical action which can also explains as a binary form 1 and 0. ECHOproduces ultrasonic-waves to make a structural imitation of the heart and utilizes to differentiate cardiovascular problems (CVP). This is major upcoming undertakings to analyse, which requires future data including great abilities. DM is the process of investigation to find out the hidden patterns of data and multiple perceptions to categorize the raw data to use[2].Using the data provided by the healthcare industry by applying complex algorithms. Big data strategies are suggested as a gigantic record of the dataset. DM and big data are 2-distinct strategies. The completed task of these two methodologies is tantamount to focusing on get-together the colossal proportion of information, to study, explore, and setting up a analysis of the data which segregate the data. DM is basically a research and exposition of analyzing the acquired information which is appropriate and with explicit information by utilizing bigdata. The accommodating models with concealed plans, obscure connections are consistently dealt with for making capable decisions through this Bigdata assessment measure.

2 Literature survey

An audit is accomplished on various procedures redone by researchers to predict CVD. Vast improvements of DM are plighted with in the design of CVD forecast model.

In 2004, C. Ordonez, brought up an enhanced analysis by using the associative rule to get the prediction of CVD. This examination includes detection of CVD. The evaluated data collection secure healthcare records of patients having CVD with properties like Pain in Chest, BP, diabetes, and cholesterol for health hazard factors. The heart perfusion estimated and corridor exhaustion was observed by the Author/researcher. Later stages the research ermitigated the restrictions of "Associative-rule'. Clinical data collected by radiology department of Emory University. The planned algorithm look for attributes and constraints whichunderstudy diminished set of rules. This framework looks for associative-rules by training, Validating and testing on dataset independently[3].

Advantages:Compared with traditional supervised ML or statistical approaches (DT, LR,SVM) associative rules have major advantages.
Associative rules have a direct understanding depends on the probability of an event of pattern and the contingent probability between two patterns.
These rules can connect combinations of anticipated attributes
These rules simultaneously deal with several anticipated attributes without the involvement of separate information subsets or individual runs.
These rules are able to discover patterns that exist in little subsets of properties.
At last, every rule able to allude with the dataset which is overlapped subsets of it based on other rules.

In 2006, P. Leijdekkers et al.developed a "customized Heart rate monitoring application with the help of sensors in a cell phone used as a telemetric in other words remote sensor. The whole analysis and experiment are depending upon sensors embedded in the smart gadgets those can collect patient's information also from the environment and givescautionto the helpline to get anambulanceso suffering person can get a quick care by the specialists and rescue the existence of suffering person and help him away from peril. The alarm message is shipped off a far-off medical care helpline to get to know the patient's location. Based on that appropriate considerations are suggested by cardiologists. The algorithm likewise obliges repair activities which help for the quick healing of the CVD patient by remote monitoring. This incorporates a legitimate eating regimen followed by work out. The framework was planned utilizing "2- techniques Windows Mobile PC and 'DotNetdensemodel reached out to Open NETCF' sections used while executing the model. The data is benefited from SQL server whose structure is intricate for storing data on cellular phones[4]
Advantage: This mobile based monitoring system monitors 24-hour patient heart rate and alarms both ambulance and cardiologist in life threatening situations also the log helps the doctor to understand the cause behind the blood pressure.

In 2006, Jiang, Yingtao et al.developeda "Multilayer Perceptron" (MLP) which has 40 nodesin the Input layer and 5 neurons in the output layer. The enhanced back propagation (BP) framework was utilized to train the model. 356 clinical datasets gathered to train and test the model. Evaluation done using holdout, bootstrapping and cross-validation techniques applied to survey the model. Multi-Layer Perceptron is a very reputed framework which makes the model generally significant in Neural Network frame works also they are basic and simple algorithms for better understanding. Multi-Layer Perceptron consists of three layers in specific Output Layer, Hidden layer, and Input layers.

The hidden-layers has gotten before Drop rate measure. The exploratory outcomes accomplished with 90.0% of precision[5].

Advantage: MLP's fast learning ability which helps to learn how to do experiment depends on heart data for training purpose.

In 2007 Y. Xing et al. designed DM algorithm for prediction for endurance of Cardio Vascular Disease (CVD) patients depends on 999 cases. Since the CVD prediction methodology is a prerequisite to address it as

a challenge to health care industry. The progress conveyed tremendous perception on the clinical dataset for a half year of 999CVD data. Data on the endurance evaluations was documented. The 3-significant DM procedures were utilized upon 501 cases. "10-overlay cross- validation" gauges and measures. The execution and precision of above procedures. The attributes are "specificity, accuracy, and sensitivity". In-order to calculate those three measures Confusion Matrix is used. The précised accuracy acquired92.1% using Support Vector Machines, 91.0% using Artificial Neural Networks, and 89.6% using DM Techniques. An examination completed by contrasting diverse expectation frameworks of CVD inpatients data with cross-validation which was of 10-overlap enables a variety of information. Endurance "1" with chance of hold out and non-endurance as "0"is called as a Boolean or parallel representation with respect to crude data. Among 999CVDreports, 797reports of endurance and 201reports of passing [1].

In 2012, D.M. Chitrali proposed a framework to anticipate different cardiac problems utilizing the DNFS procedure. DNFS implies a DT based deep neural network. Here researcher proposed a framework that can anticipate coronary problems with the assistance of DM methods, just as ML classifiers. for example, DT, k-NN, NB, ANN and SVM for the classification and prediction. The researcher utilized the dataset gathered by UCI for the forecast with Thirteen highlights. In authors proposed work they used KNN algorithm to extract and structure the datasets. In the very next step they approximated the nearest neighbor of fuzzy memberships by updating a linear combination. The Neural networks are used to predict the heart disease. Later they have used neuro fuzzy system for classification purpose and genetic algorithm to improve the neuro fuzzy model learning. Additionally, they led a relative report on different algorithms likewise, lastly, they discovered that NB and DT gives the best precision of 100%, 99.62%, 90.74%[6].

In 2013, Bahadur, S et al. Reduced fourteen attributes to six by utilizing Genetic Algorithm.

At that point Classifiers NB, clustering & classification and DT were utilized to anticipate the determination of cardio vascular problems. The Genetic algorithm is used on fourteen types and which was mitigated to six. The 6 types were old-peak(OLDPK) Resting-BP(RBP), chest-pain-type(CPTYPE),max heart-rate-achieved(THAL), Exercise-induced-angina(EIA),major-vessel-colored (VSL). This mitigated dataset applied to three classifier frameworks. Planning the model 4 computing assessment estimates utilized which in particular False-positive (FP), True-positive (TP) and False-Negative(FN), True-Negative(TN) and WEKA apparatus was utilized for usage. The accuracy acquired by NB, DT, Classification clustering was 96.5%, 99.2%, and 88.3% separately [7].

In 2014, Masethe, Mosimaet al. experimented and published using an algorithms J48, Naive Bayes, "REPTREE CART" and –Bayes Network sare utilized to detect "Heart failures". The informational index was gathered from clinics and specialists who were having good experience in South Africa region. Eleven ascribes showed restraint Id, Age, SEX, heart torment, BP, cardiogram, heartbeat rating, cholesterol, Alcohol consumption, tobacco utilization and diabetes rating. The apparatus called WEKA was utilized for the forecast of CVD problems. WEKA instrument was critical in finding dissecting and anticipating designs. The precision got was 99.0741 for J48, 99.222 for REPTREE, 98.148 for Naïve Bayes, 99.0741 for Bayes Net, and straightforward CART individually; Bayes Net calculation beat the Naive Bayes algorithm[8].

In 2015, Gnanapandithan, etal Contrasted diverse classification procedures with plan of danger prediction framework to find CVD. The 2- sorts of frameworks looked at first one is to test dataset and second one is a consolidated framework which is a hybrid framework to test dataset. These two frameworks are utilized to analyze the data. The researchers in this investigation have considered just the grouping strategies on account of a solitary model and a consolidated model. The precision got for Association Rule, Decision Technique, ANN,K-NN, Hybrid approach, Naive Bayes was 58%, 55%,86%, 76%, 85%, 96%, and 69% separately. It was presumed that to apply hybrid information mining methods got promising outcomes were gotten in the conclusion of coronary illness[9].

In 2016, Seemaet al. implemented prescient examination to forestall and control the Cardio Vascular illness (CVD) with the assistance of ML strategies like NB, DT, ANN and SVM uses the information acquired from the CMLIS at UCI they have utilized UCI AI database to ascertain the precision.

In all these above algorithms SVM provided the exact precision percentage of 95.55% and rest ANN was 94.27%, Naïve Bayes was 93.85% and Decision tree was 92.59%[10].

In 2016, Purushottamet al. built up an information mining model to foresee coronary illness productively. It chiefly encourages the clinical experts to settle on productive choices route subject to the given limits. The researcher has utilized the Cleveland dataset, and they used parameters likesex, age, chest torment, resting circulatory strain serum cholesterol, fasting glucose, and so on as traits. Moreover, they have partitioned the datasets into two sections one is for trying, and the other one is for preparing. They have utilized a 10-overlap technique to discover exactness[11].

In 2017 R G Saboji implemented a Random Forest algorithm which gives a metric solution and provides 98.01% accuracy is using 600dataset records. The researcher analyzed the accuracy achieved in the range of 200-600 records, accuracy also rise 88.0%-98.0%. It is very good observation to show that from 200 records

to 400 records accuracies expanded by 8%, when 400 records raised to 600 records the elevated accuracy is 2%, seemingly shows the degradable accuracy returns. Also, this model explains how useful the big data analytics in healthcare industry to provide faster and easy methodology for disease classification and prediction[12].

In 2018, A. Rahmanet al. suggested a system which is comprises of 2 strategies one is $(X)^2$ measurable and DNN. Highlight refinement is finished by $(X)^2$ measurable model and arrangement is finished by Deep-NN. while investigation, they took thecleveland collected dataset. There are three hundred and three(303) occurrences are there in that dataset, in which, 297 occurances have no loss of data, and the leftover 6 instances have loss ofdata. Out of 297occurances, 207 cases are seperated for training, and the rest 90 are kept in a side for testing puropose. This systemprovidesbest outcomes contrasted with customary ANN models which are available prior. Because of this utilizing this proposed model, they have 93.33% characterization precision utilizing DNN. Compared with traditional ANN framework this model gives 3.33% more[13].

In 2019L. Ali et al. proposed a specialist framework dependent on two SVM models to predictCVD effectively. These two SVM's have individua tasks to handle which isthe 1st SVM model is utilized to eliminate the features which are unnecessary, and 2nd SVM model is used for classification and prediction. Besides, they have utilized the HGSA (crossover brace search calculation) to streamline the two strategies.

By utilizing this framework, they have accomplished 3.30% preferable precision over the ordinary SVM algorithmswhich are available prior[14].

In 2019 S. Zhouet al. built up a framework to enhance the prediction of coronary illness and end up with a problem calledoverfitting. The problem overfitting implies the suggestedframework gives better accuracywhile testing on dataset and gives lamentable accuracy result while training the coronary illness. To take care of this issue, they have come up with a system whichprovies the betterprecision on both testing and training dataset. This system comprises of 2-algorithms1)Random-search-Algorithm(RAS)2) RFalgorithm that is utilized to foresee the problem. This proposed systemproved with 93.33% accurady which is a better outcome in the whole dataset.[15].

In 2020 Jiabing Zhang et al., proposed a system which gathers patients medical information using Internet of Medical Things (IoMT) platform[16]. Toanalyse the gathered data they have used a ML frameworks RERF-ILM for CVD detection. They have used UCI cleaveland CVD dataset to train the framework and tested it with using IOMT medical device. Using RERF-ILM[16] With a large dataset they have achieved 95.20% accuracy[16].

3 Conclusion

In view of the above research, considered from the year of 2004 to 2020 of which provides the feasibilityforvarious models within reach and the distinctive machine learning models used. The precision acquired from these architectures is likewise referenced. In this survey it tends to be expect as there is a colossal degree for Machine Learning models while anticipating cardio-vascular problems. All the researcherswe referred in this survey has performed amazingly well sometimes yet ineffectively in fewother scenarios. DT when utilized with PCA, have worked amazingly well. However, decision trees have performed inadéquate in some different cases which could be expected to over fiting.

Random Forest and Ensemble models have performed very well since they take care of the issue of over fitting by utilizing different calculations (various Decision Trees on account of Random Forest). Models dependent on the Naïve Bayes classifier were computationally quick and have likewise performed well. SVM performed incredibly well for the vast majority of the cases. The frameworks depend upon machine learning models and strategies have been exceptionally precise in anticipating cardiovascular problems also there is a huge ton of research is needed to find the best approach to overcome the problems of using high/multi-dimensional data and over fitting. A ton of examination should likewise be possible on the right outfit of calculations to use for a specific kind of information.

4 Futurework

The primary objective of this survey is to distinguish the key examples and highlights from the clinical data of CVD patient's by analyzing the data using different machine learning procedures to predict and detect the CVD problems prior it effects the patient, which also help the clinical experts to provide the right treatment at right time. The other techniques will be using a very small dataset in-order to achieve very précised predicting model.

Table 1: Comparison Table

Year	Author	Algorithm/ System	Precision/ Accuracy	Ref.
2004	C. Ordonez,	Association Rule(AR)	--	3
2006	Jiang, Yingtao et al	Multilayer-perceptron	90%	5
2006	Y. Xing et al.	ANN, Support Vector Machines,	91.0%, 92.1%, 89.6%	1
2012	D.M.Chitrali et al	NN, NB DT,	100%, 90.74%, 99.62%	6
2013	Bahadur,	DT, NB,	99.2%,	7

	S et al	Classification Clustering	96.5%, 88.3%	
2014	Masethe, Mosima et al.	J48, NB, REPTREE, CART	99.0741% , 99.222%, 98.148%, 99.0741%	8
2016	Purushotta m et al	Decision tree	Testing- 86.3% & Training- 87.3%	11
2015	Gnanapan dithan, et al.	Association Rule(AR), DT, ANN, K-NN, Hybrid approach, NB	55%, 76%, 58%, 86%, 85%, 96% 69%	9
2016	Seema et al.	NB, ANN, DT SVM	93.85% 94.27% 92.59% 95.55%	10
2017	G Saboji	Random Forest(RF)	98%	12
2018	A. Rahman et al	χ 2 Statistical Model along with Optimally Configured DNN	93.33%	13
2019	S. Zhou et al.	(RSA), RF	93.33%	15
2020	Jiabing Zhang et al.,	ReRF-ILM	95.20%	16

Table 2: Dataset Table

Year	Author	Dataset	Source	Ref
2004	C. Ordonez,	Medical Dataset	Emory University radiology department	3
2006	P. Leijdekkers et al.	ECG files	IT/BIH Arrhythmia Database	4
2006	Jiang, Yingtao et al	Heart diseases database	Self-collectionSouthwest Hospital and Dajiang Hospital, Chongqing, China356 Medical records	5
2007	Y. Xing et al.	Clinical Dataset(Heart)	knowledge discovery in databases (KDD)	1
2012	D.M.Chitrali et al	Cleveland Dataset	UCI	6
2013	Bahadur, S et al	Heart Disease Dataset	WEKA Dataset	7
2014	Masethe, Mosima et al.	Patient Dataset(11 Attributes)	Data collected from medical practitioners in South Africa	8
2015	Purushottam et al	Cleveland Dataset	UCI	11
2016	Seema et al.	Cleveland Dataset	UCI	10
2017	G Saboji	Cleveland Dataset	UCI	12
2018	A. Rahman et al	Cleveland Dataset	UCI	13
2019	S. Zhou et al.	Cleveland Dataset	UCI	15
2020	Jiabing Zhang et al.,	Cleveland Dataset	UCI	16

References

1. Y. Xing, J. Wang, Z. Zhao, and A. Gao, in *2007 International Conference on Convergence Information Technology (ICCIT 2007)*, pp. 868–872(2007)
2. "C. Ordonez, IEEE Trans. Inf. Technol. Biomed. **10**, 334 (2006)
3. P. Leijdekkers and V. Gay, in *2006 International Conference on Mobile Business*, pp. 29–29,(2006)
4. H. Yan, Y. Jiang, J. Zheng, C. Peng, and Q. Li, Expert Syst. Appl. **30**, 272 (2006)
5. S. Bahadur, Research Scholar, Department of Computer Science & Mathematics, Govt. P. G. Science College Rewa (M. P.), and India, IOSR Journal of Agriculture and Veterinary Science **4**, 60 (2013)
6. H. D. Masethe and M. A. Masethe, Proceedings of the World Congress on (2014)
7. G. Purusothaman and P. Krishnakumari, Indian J. Sci. Technol. **8**, (2015)
8. K. Deepika and S. Seema, in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, pp. 381–386(2016)
9. Purushottam, K. Saxena, and R. Sharma, Procedia Comput. Sci. **85**, 962 (2016)
10. R. G. Saboji, in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pp. 1780–1785(2017)
11. L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, IEEE Access **7**, 34938 (2019)
12. L. Ali, A. Niamat, J. A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, R. Nour, and S. A. C. Bukhari, IEEE Access **7**, 54007 (2019)
13. A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, IEEE Access **7**, 180235 (2019)
14. C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han, and J. Yu, IEEE Access **8**, 59247 (2020)

Ç. Srinath Reddy, B. Shah, C. Varghese, and A.
15. Ramadoss, Lancet **366**, 1744 (2005)

16. M. C. O. Gómez and W. G. Pérez, (2020)
17. N. K. S. Banu and S. Swamy, in *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)* (2016)
18. M. Nilugonda and K. Madhavi, E3S Web of Conferences **184**, 01053 (2020)
19. M. Arshi, M. D. Nasreen, and K. Madhavi, E3S Web of Conferences **184**, 01052 (2020)