Analysis & Prediction of Heart Attack using Machine Learning

Kumar Saurav, Hritwiz Yash, Affan
Computer Science and Engineering,
Galgotias University,
Greater Noida, India
Saurav2011ara@gmail.com, hritwizsri101@gmail.com, affans6557@gmail.com

Abstract- Heart-related sicknesses or Cardiovascular Diseases (CVDs) are the fundamental justification behind countless demise on the planet throughout recent many years and have arisen as the most perilous infection, in India as well as in the entire world. In this way, there is a need for a solid, precise, and practical framework to analyze such infections in time for legitimate treatment. AI calculations and strategies have been applied to different clinical datasets to computerize the examination of enormous and complex information. Numerous scientists, lately, have been utilizing a few AI strategies to assist the well-being with the caring industry and the experts in the determination of heat-related sicknesses. The heart is the following significant organ contrasting with the mind which has a greater need in the Human body. It siphons the blood and supplies it to all organs of the entire body. The expectation of events of heart illnesses in the clinical field is huge work. Information examination is valuable for forecasting from more data and it assists the clinical focus with anticipating different illnesses. An enormous measure of patient- related information is kept up with on a month-to-month premise. Put-away information can be helpful for the wellspring of foreseeing the event of future infections. A portion of the information mining and AI procedures are utilized to anticipate heart infections, like Artificial Neural Network (ANN), Random Forest, and Support Vector Machine (SVM). Prediction and diagnosing of coronary illness become a difficult variable looked by specialists and clinics both in India and abroad. To decrease the enormous size of passing from heart illnesses, a speedy and proficient recognition strategy is to be found. Information mining strategies and AI calculations assume a vital part around here. The scientists speeding up their examination attempts to foster programming with the help of AI calculations which can assist specialists with choosing both expectations and diagnosing coronary illness. The fundamental goal of this examination project is to foresee the coronary illness of a patient utilizing AI calculations.

Keywords- Neural Network, Machine Learning, Supervised learning, Support vector machine, random forest, ANN.

I. INTRODUCTION

The heart is a significant organ of the human body. It siphons blood to all aspects of our life systems. In the event that it neglects to work accurately, the mind and different organs will quit working, and in somewhere around couple of moment's minutes, the individual will kick the bucket.

Change in way of life, business-related pressure and awful food propensities add into the expansion in the pace of a few heart-related sicknesses. Heart sicknesses have arisen as one of the most noticeable reasons for death from one side of the planet to the other.

As indicated by World Health Organization, heart-related illnesses are answerable for requiring 17.7 million lives consistently, 31% of every single worldwide demise. In India as well, heart-related infections have turned into the main source of mortality. Heart illnesses have killed 1.7 million Indians in 2016, as indicated by the 2016 Global Burden of Disease Report, delivered on September 15, 2017.

Heart-related illnesses increment the spending on medical care and furthermore diminish the efficiency of a person. Gauges made by the World Organization (WHO), propose that India has lost up to $237 billion, from 2005-to 2015, because of heart- related or cardiovascular illnesses. Hence, the practical and precise expectation of heart-related sicknesses is vital.

Clinical associations, from one side of the planet to the other, gather information on different wellbeing-related issues. This information can be taken advantage of by utilizing different AI procedures to acquire helpful bits of knowledge. However, the information gathered is extremely huge and, ordinarily, this information can be exceptionally uproarious.

These datasets, which are excessively overpowering for human personalities to fathom, can be handily investigated utilizing different AI strategies. Hence, these calculations have become extremely helpful, as of late, to anticipate the presence or nonattendance of heart-related illnesses precisely.

The utilization of data innovation in the medical services industry is expanding step by step to help specialists in dynamic exercises. It helps specialists and doctors in infection the executives, meds, and revelation of examples and connections among determination information. Current ways to deal with anticipate cardiovascular gamble neglect to recognize many individuals who might profit from preventive treatment, while others get pointless intercession. AI offers an amazing chance to further develop exactness by taking advantage of mind-boggling collaborations between risk factors. We evaluated whether AI can further develop cardiovascular disease forecasts.

II. LITERATURE SURVEY

Chala Beyene et all[1], recommended Prediction and Analysis of the occasion of Heart Disease Using Data Mining Techniques. The guideline objective is to expect the occasion of coronary disease for early modified assurance of the disorder inside achieving a short period of time. The proposed approach is in likemanner fundamental in a clinical benefits relationship with experts that have no more data and aptitude. It uses different clinical qualities, for instance, glucose and heartbeat, age, and sex are a piece of the attributes are integrated to perceive if the individual has a coronary ailment or not. Examinations of the dataset are enrolled using WEKA programming.

Senthil kumar Mohan et all[2], implemented cream AI for coronary sickness assumption. The enlightening list used is Cleveland instructive file. The underlying advance is the data pre- taking care of step. In this, the tuples are taken out from the instructive record which has missed the characteristics. Attributes age and sex from educational assortment are moreover not used as the makers envision that it's own special information and no influences predication. The extra 11 credits are seen as huge as they contain central clinical records. They have proposed their own Hybrid Random Forest Linear Method (HRFLM) which is a mix of Random Forest (RF) and linear procedure (LM). In the HRFLM computation, the makers have used four estimations.

The first computation oversees separating the information dataset. It relies upon a decision tree which is executed for every illustration of the dataset. Directly following perceiving the component space, the dataset is separated into theleaf center points. A consequence of the first estimation is the Partition of instructive records. After that in the second computation, they apply rules to the instructive file and the result here is the portrayal of data with those guidelines. In the third estimation, features are taken out using Less Error Classifier. This computation oversees noticing the base and most outrageous misstep rate from the classifier. An aftereffect of this computation is the features with gathered attributes. In forward computation, they apply Classifier which is a mutt strategy considering the misstep rate on the Extracted Features. Finally, they have pondered the results procured ensuing in applying HRFLM with other game plan computations such as a decision tree and support vector machine. In a result as RF and LM are giving better results than others, both the estimations are gathered and a new unique computation HRFLM is made. The makers suggest further improvement in precision by using a mixof various AI estimations.

Ali, Liaqat, et all[3], propose a system containing two models taking into account straight Support Vector Machine (SVM). The first is called L1 regularized and the resulting one is called L2 regularized. The first model is used for killing pointless features by making the coefficient of those components zero. The ensuing model is used for the figure. Predication of disease is done in this part. To improve the two models they proposed a cream grid search estimation.

This computation improves two models considering estimations: precision, responsiveness, simplicity, the Matthews relationship coefficient, ROC outline and locale under the curve. They used Cleveland's enlightening list. Data parts into 70% readiness and 30% testing used holdout endorsement. There are two investigations done and every assessment is finished for various potential gains of C1, C2 and k where C1 is the hyper parameter of L1 regularized model, C2 is hyper parameter of L2 regularized model and k is the size of picked subset of features. First examination is the L1-direct SVM model stacked with L2- straight SVM model which is giving most outrageous testing accuracy of 91.11% and planning precision of 84.05%. The ensuing examination is L1-straight SVM model streamed with L2-direct SVM model with RBF part. This is giving most outrageous testing accuracy of 92.22% and getting ready precision of 85.02.They have gotten an improvement in precision over standard SVM models by 3.3%

Singh, Yeshvendra K. et all[4], deal with various directed AI estimations like Random Forest, Support Vector Machine, Logistic Regression, Linear Regression, Decision Tree with 3 cross-over, 5 wrinkle and 10 overlay cross- endorsement methodologies. They have used Cleveland educational assortment having 303 tuples, with some tuples havingmissing credits.

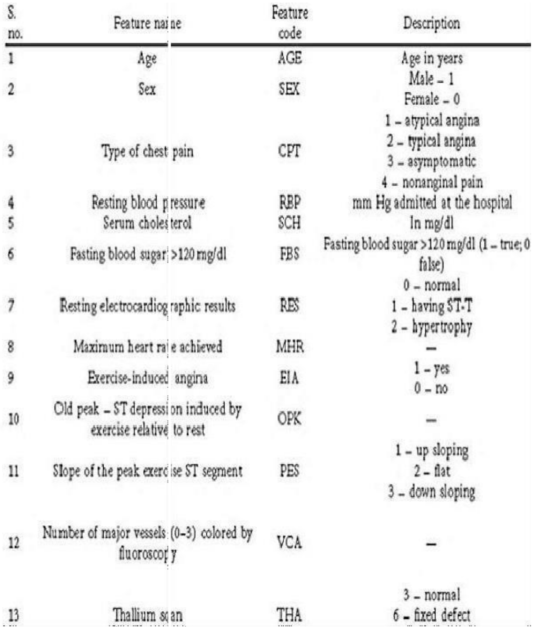
In the preprocessing of data they just killed the missing worth tuple from the instructive assortment which are six in number and a short time later from the overabundance 297 tuples, they apportioned the data as planning 70% and testing 30%.First estimation applied is Linear Regression. In this, they have described the dependence of one property over others which can be sprightly disengaged from each other. Basically the game plan occurs with the help of the social event of attributes used for twofold request. They have obtained best results in 10 wrinkles which is 83.82%. Determined backslide request is done using a sigmoid limit. This estimation applied for coronary ailment assumption shows most outrageous precision with 3 and 5 wrinkle cross-endorsement and it is 83.83%. Support Vector Machine is the request computation in managed AI. In this the request is done by hyper plane.

The most extreme precision accomplished by S M 3 overlay cross-approval is 83.17%.For Decision Tree in this paper, the creators have utilized different number parts and different number of leaf hubs to track down the most extreme precision. With 37 number parts and 6 leaf hubs most extreme precision is accomplished which is 79.12%. When utilized with cross- approval, precision accomplished by the choice tree 79.54% with 5 overlay. Irregular woodland calculation utilized on nonlinear informational index gives improved outcomes when contrasted with the choice tree.

Arbitrary woodland is the gathering of choice tree made by the different root hubs. From this gathering of choice tree, Casting a ballot should be possible first and afterward order should be possible from the one getting greatest votes. Creators have utilized different number parts, different number of tree different number of folds for cross-approval. For arbitrary woods, 85.81% precision is accomplished by 75 Number of trees and 10 numbers of folds.

II. DATASET

Table 1. Data Set.



S. no.	Feature name	Feature code	Description
1	Age	AGE	Age in years
2	Sex	SEX	Male – 1 Female – 0
3	Type of chest pain	CPT	1 – atypical angina 2 – typical angina 3 – asymptomatic 4 – nonanginal pain
4	Resting blood pressure	RBP	mm Hg admitted at the hospital
5	Serum cholesterol	SCH	In mg/dl
6	Fasting blood sugar >120 mg/dl	FBS	Fasting blood sugar >120 mg/dl (1 – true; 0 false)
7	Resting electrocardiographic results	RES	0 – normal 1 – having ST-T 2 – hypertrophy
8	Maximum heart rate achieved	MHR	–
9	Exercise-induced angina	EIA	1 – yes 0 – no
10	Old peak – ST depression induced by exercise relative to rest	OPK	–
11	Slope of the peak exercise ST segment	PES	1 – up sloping 2 – flat 3 – down sloping
12	Number of major vessels (0–3) colored by fluoroscopy	VCA	–
13	Thallium scan	THA	3 – normal 6 – fixed defect

We performed virtual experience on one dataset. Dataset is a Heart dataset. The dataset is accessible in UCI Machine Learning Repository [10]. Dataset contains 303 examples and 14 info highlights as well as 1 result include. The elements depict monetary, individual, and social component of credit candidates. The result include is the choice class which has esteem 1 for Good credit and 2 for Bad credit. The dataset-1 contains 700 examples displayed as Good acknowledge while 300 occasions as awful credit. The data set contains highlights communicated on ostensible, ordinal, or span scales. A rundown of that large number of highlights is given in Table.

III. PROPOSED SYSTEM

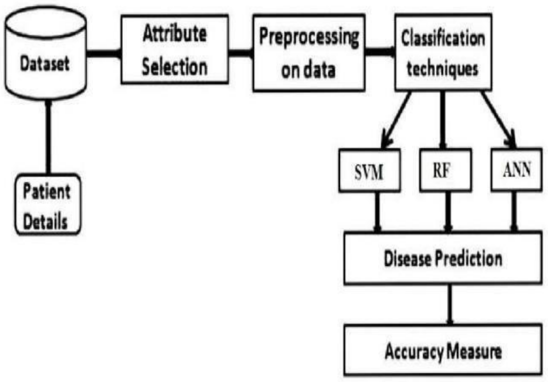


Fig 1. Proposed System.

- **1. Random Forest:**

Random Forest is a managed AI calculation. This method can be utilized for both relapse and order undertakings yet by and large performs better in arrangement assignments. As the name proposes, the Random Forest procedure considers numerous choice trees prior to giving a result. In this way, it is essentially a troupe of choice trees. This procedure depends on the conviction that more trees would unite to be the best choice. For grouping, it utilizes a democratic framework and afterward chooses the class while in relapse it takes the mean of the multitude of results of every one of the choice trees. It functions admirably with enormous datasets with high dimensionality.

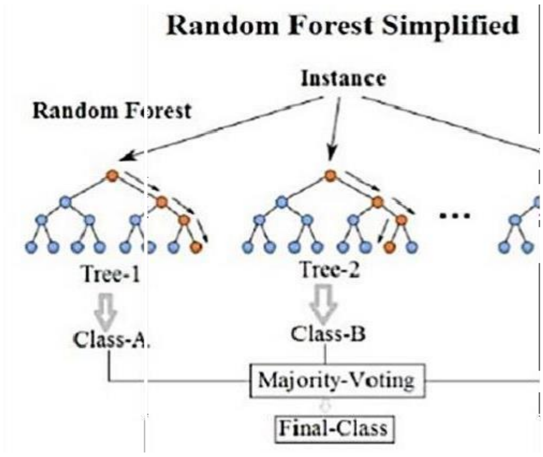


Fig 2. Support vector machines (SVMS).

2. Support Vector Machine (SVMS):
Support vector machines exist in various structures, straight and non-direct. A help vector machine is an administered classifier. What is common in this unique situation, two distinct datasets are engaged with SVM, preparing and a test set. Experiencing the same thing the classes are straightly distinguishable. Experiencing the same thing a line can be found, what divides the two classes impeccably? Anyway, one-line parts the dataset impeccably, yet an entire pack of lines does. From these lines, the best is chosen as the "isolating line".

An SVM can make a few blunders to abstain from over-fitting. It attempts to limit the number of blunders that will be made. Support vector machines classifiers are applied in numerous applications. They are extremely famous in late exploration. Thisnotoriety is because of the great in general observational presentation. Contrasting the credulous Bayes and the SVM classifier, the SVM has been applied the most

3. Artificial Neural Network:
These are utilized to show/reenact the conveyance, capacities or mappings among factors as modules of a powerful framework related with a learning rule or a learning calculation. The modules here mimic neurons in sensory system and subsequently ANN by and large alludes to the neuron test systems and their synapsis recreating interconnections between these modules in various layers.

Brain Network is worked by stacking together numerous neuronsin layers to deliver a last result. First layer is the info layer and the latter is the result layer. Every one of the layers in the middle is called secret layers. Every neuron has an initiation work. A portion of the well-known Activation capacities are Sigmoid, ReLU, tanh and so on. The boundaries of the organization are the loads and predispositions of each layer. The objective of the brain network is to gain proficiency with the organization boundaries to such an extent that the anticipated result is equivalent to the ground truth. Back-spread along misfortune work is utilized to become familiar with the organization boundaries.

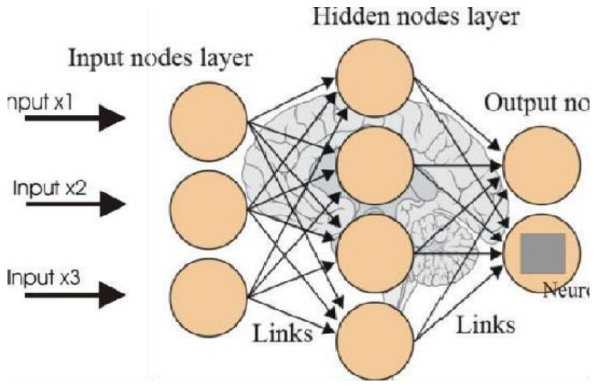


Fig 3. Artificial Neural Network.

IV. SOFTWARE USED

1. Python*:*
To gather information a web scraper modified in Python was utilized. As per Wikipedia Python's linguistic structure permits developers to communicate ideas in fewer lines of code. Guido van Rossum at CWI in the Netherlands began Python's execution in December 1989. Python 2.0 was delivered on October sixteenth, 2000, and Python 3.0 was delivered on December third, 2008.

Why use Python for web scratching and not something else? Python offers a module called 'urllib2', which has appropriate capacities to effectively open sites and concentrate data. Python is utilized to program the web scrubber that is responsible for gathering the climate information for the model.

2. MS Excel:
Microsoft Succeed is a calculation sheet application created by Microsoft for Windows and Macintosh operating system X. It highlights estimation, diagramming devices, turntables, and a large-scale programming language. The main rendition was delivered in 1987. Why pick MS Succeed versus one more comparable kind of programming? MS Succeed is an extremely complete bookkeeping sheet application instrument, which upholds practically any sort of document augmentation, and it has a ton of highlights. Its easy-to-understand interface helps you more often than not. Nonetheless, on the off chance that this doesn't appear to be sufficient, I will say that, aside from the average things an ordinary client would do in Succeed (Outlines, Estimation…), it empowers you to utilize the VBA language tomake capacities to use on the calculation sheets you've made.

Succeed can likewise be utilized as though it were a SQL information base as was made sense of in a past part. Having said this, for me it is the ideal program.MS Succeed is utilized a ton all through the undertaking, to envision the information and perform cleaning assignments on it.

V. RESULT AND DISCUSSION

In the wake of preprocessing the information. The information grouping procedure specifically support vector machine, fake brain organization, irregular woods were applied. The task included investigation of the coronary illness patient dataset with appropriate information handling. Then, at that point, 3 models were prepared and tried with greatest scores as follow:
Machine, artificial neural network, and random forest were applied. The project involved analysis of the heart disease patient dataset with proper data processing.

Then, 3 models were trained and tested with maximum scores as follows:

- Support Vector Classifier: 84.0 %
- Neural Network: 90.16 %
- Random Forest Classifier: 80.0 %

VI. CONCLUSION

This project provides the deep insight into machine learning techniques for classification of heart diseases. The role of classifier is crucial in healthcare industry so that the results can be used for predicting the treatment which can be provided to patients. The existing techniques are studied and compared for finding the efficient and accurate systems.

Machine learning techniques significantly improves accuracy of cardiovascular risk prediction through which patients can be identified during an early stage of disease and can be benefitted by preventive treatment. It can be concluded that there is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart related diseases. Each of the above-mentioned algorithms has performed extremely well in some cases but poorly in some other cases.

VII. ACKNOWLEDGEMENTS

We would like to express special thanks of gratitude to our guide Ms Gitika Sharma as well as our project Reviewer Ms. N.suvarna, who gave us the golden opportunity to do this wonderful project on the topic Heart Disease Prediction Using Machine Learning, which also helped us in doing a lot of research and we came to know about so many new things. We would also like to thank our Dean Of CSE Munish Sabharwal for providing us the opportunity to implement our project. We are thankful to both of them. Finally, we would also like to thank our department faculty members and our parents & friends who helped us a lot in finalizing this project within the limited time frame.

REFERENCES

[1] Mr. ChalaBeyene, Prof. Pooja Kamat, “Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Technique”, International Journal of Pure and Applied Mathematics, 2018.

[2] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava, “Effective heart disease prediction using hybrid machine learning techniques” IEEE Access 7 (2019): 81542-81554.

[3] Ali, Liaqat, et al, “An optimized stacked support vector machines based expert system for the effective prediction of heart failure” IEEE Access 7 (2019): 54007-54014.

[4] Singh Yeshvendra K., Nikhil Sinha, and Sanjay K. Singh, “Heart Disease Prediction System Using Random Forest”, International Conference on Advances in Computing and Data Sciences. Springer, Singapore, 2016.

[5] Prerana T H M1, Shivaprakash N C2 , Swetha N3 “Prediction of Heart Disease Using Machine Learning ,Algorithms- Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS” International Journal of Science and Engineering Volume 3, Number 2 – 2015 PP: 90-99

[6] B.L DeekshatuluaPriti Chandra “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm” International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.

[7] Michael W.Berryet.al, Lecture notes in data mining, WorldScientific(2006)

[8] S. Shilaskar and A.Ghatol, “Feature selection for medical diagnosis: Evaluation for cardiovascular diseases, ExpertSyst.Appl” vol. 40, no. 10, pp. 4146–4153, Aug. 2013.

[9] C.-L. Chang and C.-H. Chen, “Applying decision tree and neural network to increase quality of dermatologic diagnosis,” Expert Syst. Appl., vol. 36, no. 2, Part 2, pp. 4035–4041, Mar. 2009.

[10] T. Azar and S. M. El-Metwally, “Decision tree classifiers for automated medical diagnosis,” Neural Comput. Appl.,vol. 23, no. 7–8, pp. 2387–2403, Dec. 2013.

[11] Y. C. T. Bo Jin, “Support vector machines with genetic fuzzy feature transformation for biomedical data classification.,” Inf Sci, vol. 177, no. 2, pp. 476–489, 2007.

[12] N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, “Knowledge discovery in medicine: Current issue and future trend,” Expert Syst. Appl., vol. 41, no. 9, pp. 4434– 4463, Jul. 2014.

[13] E. Hassanien and T. Kim, “Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neuralnetworks,” J. Appl. Log., vol. 10, no. 4, pp. 277–284, Dec. 2012.

[14] Sanjay Kumar Sen 1, Dr. Sujata Dash 21Asst. Prof., Orissa Engineering College, Bhubaneswar, Odisha – India.

[15] Domingos P and Pazzani M. “Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier”, in Proceedings of the 13th Conference on Machine Learning, Bari, Italy, pp 105-112, 1996.

[16] Elkan C. “Naive Bayesian Learning, Technical Report CS97-557”, Department of Computer Science and Engineering, University of California, San Diego, USA, 1997.

[17] B.L Deekshatulua Priti Chandra “Reader, PG Dept. Of Computer Application North Orissa University, Baripada, Odisha – India. Empirical Evaluation of Classifiers Performance Using Data Mining Algorithm”