ayUSHpLAG

by Preeti Garg

Heart Disease Prediction using various Machine Learning Techniques

Ayush Singh
KIET Group of Institutions
Delhi-NCR Ghaziabad
ayushvns3@gmail.com

Sahil Saxena
KIET Group of Institutions
Delhi-NCR Ghaziabad
sahilsaxenarbs@gmail.com

Yatika Nagar
KIET Group of Institutions
Delhi-NCR Ghaziabad
yatikanagar@gmail.com

Preeti Garg
KIET Group of Institutions
Delhi-NCR Ghaziabad
Preeti.itgarg@gmail.com

ABSTRACT- Technology and medical science have developed quite alot in last few decades and their combination has made a vital role in developing medicines and machines to help doctors in various diseases, sometimes these diseases are harder to predict and the delays could be fatal. If there is a way so that we can shortlist people who may have a particular disease it can significantly reduce the burden from the medical industry and can be a game changer. Currently we have a lot of machine learning algorithms that can help us in determining patterns that could lead to a disease. In Our findings we used different machine learning technics and algorithms like neuralnetwork, SVM, decision tree etc. to find whether an individual has Chronical Heart Disease or not. After comparison of results with other researches and correctly using the resources, methods , technique and technology we concluded everything in our results and findings where SVM had the best accuracy and others performed fairly.

Keywords: Chronical Heart Disease, Machine Learning, Neural network, SVM.

I. INTRODUCTION

Heart disease is an umbrella term that covers a variety of heart conditions. Another name for it is cardiovascular disease, which refers to disorders of the heart and blood vessels. Chronical heart disease is a major factor of cause of death in the world, but there are many methods to prevent and manage these types of heart diseases.

One of the main causes of people's dejection and death is heart disease. Heart disease alone claims the lives of about 610,000 Americans each year, accounting for 1 in 4 fatalities. Every year about 735,000 Americans get a heart attack out of which 525,000 of them were their first heart and attack and rest of them had already experienced heart attack before [1]. Projection of cardiovascular disease is considered as an essential subject in the field of medical science and data analysis. The amount of data that is currently available to the healthcare sector is enormous. The massive amount of data is transformed into insightful knowledge through data mining , which enables decision and prediction making that is well-informed.

Machine learning is a subfield of artificial intelligence that focuses on using data correctly and utilizing other methods to simulate human comprehension and gradually increase accuracy. They have received training on how to identify different patterns and ascertain the relationship between data. They use historic data and makes the predictions, cluster data points and even help generate new material, as shown by new AI running engines applications such as ChatGPT, Github copilot, Google Bard.

Several machine learning (ML) algorithms, including Naive Bayes, K-NN, Random forest, Decision Tree (DT), and SVM algorithms, can be used to create a system that can predict a variety of diseases. The other end of the coin is that these systems are usually not liked by the doctors as the disease predictions somewhat reduce the need for doctors, which makes the doctors panic of their livelihood. But these methods actually integrate technology and the doctor's knowledge to improve the doctor recommendations which solves the main issue making sure the people also trust the advice of doctors and also improves their business. Some of these systems and approaches have been implemented in our research and the results are also shown.

II. RELATED WORK

The UCI data repository is utilized for heart disease prediction in [1] through the application of K-Star, along with Multilayer Perception. SMO (89%) and Naïve Bayes (87%) exhibit optimal results, out-performing K-Star, Multilayer Perception. Despite these achievements, accuracy of these algorithms is deemed unsatisfactory.

Kaggle data is employed [2] to predict stroke patients using the knowledge discovery process with ANN and SVM. The results show 81.82% and 80.38% accuracy for Artificial neural network and SVM, in the training dataset.

Authors in [3] uses UCI repository data to assess various machine learning algorithms, including Naive Bayes, KNN. Among these, ANN attains the maximum accuracy.

In [4], the WEKA tool is employed to measure the performance of different ML algorithms. The application of PCA with ANN results in an accuracy of 94.5% before PCA and 97.7% after PCA. This substantial difference is observed. Here, Cardiovascular Disease is predicted using different machine learning techniques and algorithms which include Random Forest Classifier. The highest accuracy of 85% was the result of implementation of Random Forest classifier as the algorithm.

A different study [5] claims that when compared to other models, the artificial neural network has the best accuracy of 84.25%. It's interesting to note that this lower accuracy model is chosen as the final main model even though other algorithms demonstrate greater accuracy than ANN.

In [6], the Hidden Naïve Bayes algorithm achieves 100% accuracy in predicting heart disease, surpassing regular Naïve Bayes. Lastly, Authors in [7] suggests the use of Hidden Naïve Bayes algorithm for heart disease prediction, achieving 100% accuracy and outperforming regular Naïve Bayes.

Considerable efforts in diagnosing of chronical Heart disease through Machine Learning technics have spurred this study. The research paper includes a brief review of the literature and presents an efficient method of predicting chronic heart disease using multiple algorithms, such as Random Forest Classifier, KNN, and Logistic Regression. The Outcomes show that every algorithm possesses strengths in achieving defined objectives [8].

The model that incorporates IHDPS shows how deep learning models and both more advanced and more traditional machine learning techniques can be used to analyze the decision boundary. It makes important information and factors easier to access, like a family history of heart disease. When compared to the most recent emerging models, the IHDPS model's accuracy is noticeably lower, especially when it comes to identifying chronic heart disease through the use of artificial neural networks, other machine learning techniques, and deep learning algorithms. Using a built-in implementation algorithm that made use of neural network techniques, McPherson et al. [9] identified risk factors for atherosclerosis or coronary heart disease and accurately predicted the presence of the disease in test subjects.

Neural networks were used for the first time in the diagnosis and prediction of blood pressure and heart disease by R. Subramanian et al. [10]. They constructed a deep neural network with characteristics associated with the illness, resulting in an output that was handled by an output perceptron and contained nearly 120 hidden layers. This fundamental technique ensures accurate results when applied to a Dataset. A supervised network is recommended for diagnosing heart diseases [11]. During testing by a physician using unfamiliar data and unstructured data, the model utilizes prior learned data to precise results, thus calculating the accuracy of the given model.

III. IMPLEMENTATION

Treatment of heart disease can only be done by a Doctor who has greater knowledge of the type and the stage of the cancer. This is done by analyzing various symptoms and different factors such as cholesterol, age, gender, blood pressure, body mass index, et cetera. Analyzing different machine learning algorithms can assist in predicting the type of chronicle heart disease. Various Python libraries such as numpy, panda, and matplotlib will be utilized.

Additionally, the implementation made use of Keras and sklearn as well as the machine learning portion of the project. The University of California, Arvind Irvine machine learning repository makes the data set available. The data set comprises patient information about heart disease, diagnosis, and history that was gathered from multiple locations worldwide. Age, sex, resting blood pressure, cholesterol, echocardiogram, data, exercise, habits, and many more are among the 76 distinct attributes. We primarily concentrate on a subset of 14 attributes across all of these data, specifically, we will use the data collected at Cleveland clinic foundation.

- The KNN (K-nearest neighbour) algorithm, also known as KNN for short, is a supervised learning classifier that relies on proximity to make predictions and classifications about how individual data points will cluster. Although it is primarily used for classification algorithms, it can also be applied to regression and classification-type problems. A class label is created for classification issues based on a majority vote. It is discovered that K-NN has an accuracy of 85.06%.The K Neighbour classifier scores for various K values are displayed in the graph below. K neighbours' number is indicated on the x-axis.

- Support Vector Machines: SVM is an implementation of Vapnik's support vector machine, for the problems like regression, developing a ranking function and pattern recognition[12]. This algorithm can handle issues that have many thousand of support vector efficiently and it has quite scalable memory needs. From the gender distribution we conclude that male patients are more than female patients. Also the frequency of heart diseases patients in males are more than normal male patients. But for females its vice versa. From the age distribution plots we conclude that most density for normal patients is around age 50 and for heart diseases patients is around age 60. Accuracy in our case by using SVM algorithm is 94.41%.

- Naive Bayes: The Bayes theorem is crucial to many recently developed machine learning models as well as statistics. Bayesian reasoning plays a significant role in science because it updates the probability of a hypothesis whenever new evidence becomes available[13]. Mathematicians were able to find answers to questions with Bayesian analysis that were not previously found with frequentist statistical approaches. Actually, the frequentist paradigm has nothing to do with the idea behind giving a hypothesis its probability. Age, gender, test bps, cholesterol level, fbs, restecg, thalach, exang, oldpeak, slope, ca, and thal were the characteristics that were involved in preventing CHDs. The Naive Bayes algorithm has an accuracy of 85.25%.

- Logistic Regression: In the contemporary landscape, heart disease-related fatalities have escalated, with an alarming statistic of approximately one death per minute attributed to this health issue. The continuous influx of data, driven by rapid Information Technology growth, necessitates daily storage. Data analysis, employing diverse algorithmic combinations, transforms this collected data into actionable knowledge. However, within the field of heart disease, medical professionals face limitations in accurately predicting the likelihood of disease onset. This paper uses the Logistic Regression model in machine learning to improve the accuracy of Heart Disease prediction. The goal of the study is to overcome the predictive limitations posed by current healthcare practices by using a healthcare dataset to classify patients based on the presence or absence of heart disease [14].

- Fig 1 shows the confusion matrix of logistic regression. The performance parameters are shown in Table 1 which shows that the accuracy achieved here is 87%.

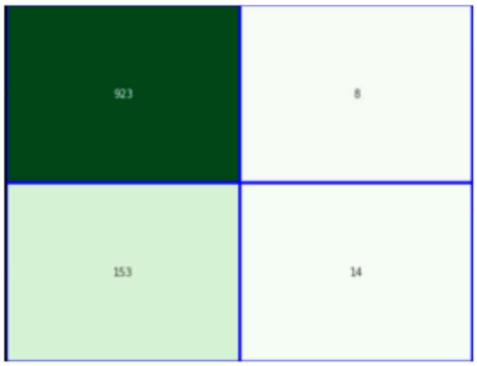

Fig 1: Confusion Matrix for Logistic Regression

Table 1: Performance Measures Result

	Precision	Recall	f1-score	Support
0	0.87	0.75	0.81	36
1	0.69	0.83	0.75	24

- *Neural Network*: McPherson et al. [2] successfully identified various critical factors for coronary heart disease or atherosclerosis through an inherent algorithmic implementation utilizing certain techniques within the Neural Network framework. Their approach accurately predicted the presence of the given disease in test patients. R. Subramanian et al. [3] pioneered the application of neural networks for diagnosing and predicting heart disease, blood pressure, and related attributes. They constructed a very complex and deep neural network incorporating disease-related attributes, featuring an output perceptron and an impressive 120 hidden layers. This innovative approach stands as a fundamental and highly effective technique, ensuring accurate results when applied to test datasets for the prediction of heart disease.

IV. RESULT AND DISCUSSION

We used a range of methods in our research on the prediction of chronicle heart disease and the application of various machine learning algorithms. Evaluating how well these algorithms performed in accurately categorizing patients according to their risk of heart disease was the main goal.

Our testing produced some intriguing findings about each algorithm's accuracy. With an accuracy of 85%, ANN was the best-performing algorithm; SVM came in second with an even more remarkable accuracy of 94.41%. With respectable accuracies of roughly 78%, KNN, Naive Bayes, and Decision Trees proved useful in the prediction of heart disease.

The models' complexity, the dataset's characteristics, and the hyperparameters that were employed during training are some of the reasons for the variations in accuracy between the algorithms. ANN, being a powerful and flexible model, can capture intricate relationships within the data, thereby achieving superior performance. On the other hand, simpler models like Naive Bayes and Decision Trees may struggle to capture complex patterns present in the dataset, leading to slightly lower accuracies.

Analyzing the distribution of attributes such as age and gender provided valuable insights into the risk factors associated with heart disease. Our findings revealed a higher prevalence of heart disease among male patients compared to females, with certain age groups exhibiting a higher density of heart disease occurrences. These observations align with existing medical literature, highlighting the importance of demographic factors in assessing cardiovascular risk.

The high accuracy achieved by SVM underscores its robustness in handling complex datasets and capturing nonlinear relationships. Additionally, SVM's ability to efficiently handle large-scale datasets with scalable memory requirements makes it well-suited for healthcare applications.

Nonetheless, we must admit the shortcomings of our models and work. The quality and quantity of supervised data, as well as the feature selection procedure, have a significant impact on the machine learning algorithms' performance and efficiency. Further research is necessary to determine whether our models can be applied to a variety of clinical settings and populations.

In summary, our research primarily shows how machine learning algorithms can be used to improve heart disease prediction systems in terms of both accuracy and potential. We can help medical professionals detect heart disease early and treat it individually by utilizing cutting-edge computational techniques. This will ultimately improve patient outcomes and save costs associated with healthcare.

V. COMPARISON OF DIFFERENT ALGORITHMS

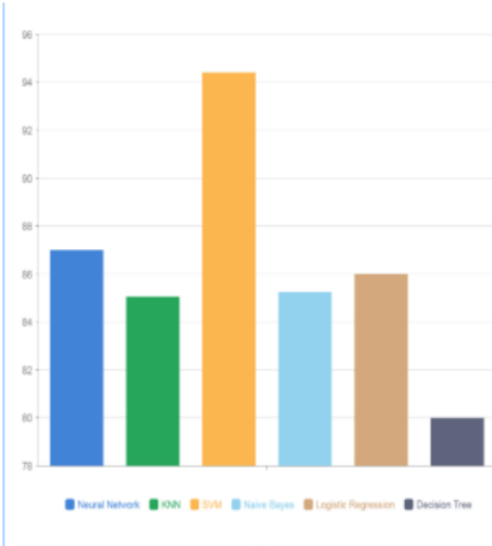

Fig 2: Accuracy comparison

Our study investigates the predictive power of different machine learning algorithms for heart disease. Based on their accuracy and applicability for this task, we evaluate the performance of Naive Bayes, K-Nearest Neighbors (KNN), Decision Trees, Support Vector Machines (SVM), and Artificial Neural Networks (ANN).

Naive Bayes, KNN, and Decision Trees are classic machine learning algorithms with distinct approaches to classification. Naive Bayes assumes feature independence, KNN classifies based on the majority class of its nearest neighbours, and Decision Trees partition the feature space using simple decision rules.

- Naive Bayes achieves an accuracy of approximately 85.25%, demonstrating its ability to capture basic patterns in the dataset.
- KNN also achieves an accuracy of approximately 85.06%, indicating its effectiveness in capturing local patterns within the data.
- Decision Trees achieve an accuracy of 80%, showcasing their ability to capture hierarchical relationships.

While Naive Bayes and KNN demonstrate similar accuracies, Decision Trees outperform both, suggesting its potential for heart disease prediction tasks where hierarchical relationships among features are essential. More sophisticated algorithms that can identify nonlinear relationships in the data include SVM and ANN.

- SVM achieves the highest accuracy of 94.41%, highlighting its robustness in handling complex datasets and capturing nonlinear relationships.

- ANN achieves an accuracy of 87%, demonstrating its ability to learn complex patterns within the dataset.

SVM performs better than ANN in terms of accuracy, but ANN strikes a compromise between interpretability and performance, which makes it a good choice for heart disease prediction systems where transparency is crucial.
Figure 2 shows the accuracy achieved by various techniques. Here the highest accuracy achieved is using SVM classification.

VI. CONCLUSION

The authors of this study used machine learning methods and algorithms, such as Naive Bayes, K-NN, Decision Trees, SVM, and Artificial Neural Networks (ANN), to create a prediction system for heart disease. Our goal was to use pertinent medical characteristics to accurately classify patients according to their risk of heart disease.
Through the experimentation, it has been found that SVM achieved the best accuracy of 94.41%, followed closely by Artificial Neural Network with an precision of 87%. KNN, Naive Bayes, and Decision Trees achieved accuracies of approximately 83%, indicating their effectiveness in predicting heart disease to a reasonable degree. Analyzing the distribution of attributes such as age, gender, and various physiological parameters revealed important insights. For instance, male patients showed a higher prevalence of heart disease compared to females, and certain age groups exhibited higher densities of heart disease occurrences.

Our results highlight the potential of machine learning algorithms to support health care providers in the early diagnosis and prognosis of heart disease. These predictive models can help with tailored treatment plans and preventive healthcare initiatives, which will lessen the number of deaths from heart disease.
Nonetheless, it's critical to recognize the shortcomings of our research. The quality and quantity of data that is available, as well as the feature selection procedure, have a significant impact on how well machine learning algorithms function.
Additionally, further validation of our models on diverse datasets and in clinical settings is warranted to assess their generalizability and real-world applicability.
To sum up, our study and work contribute to the continuing efforts to use machine learning to enhance chronic heart disease prediction systems. Through the utilization of data-driven methodologies, cardiology stands to be revolutionized, resulting in improved patient outcomes and decreased healthcare expenses.

References

1. Soni, Jyoti & Ansari, Ujma & Sharma, Dipesh & Soni, Sunita. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. International Journal of Computer Applications. 17. 43-48. 10.5120/2237-2860.

2. Shafique, Umair & Majeed, Fiaz & Qaiser, Haseeb & Mustafa, Irfan. (2015). Data Mining in Healthcare for Heart Diseases. International Journal of Innovation and Applied Studies. 10. 2028-9324.

3. Beyene, C. & Kamat, Pooja. (2018). Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics. 118. 165-173.

4. Awan, Shahid & Riaz, Muhammad & Khan, Abdul. (2018). Prediction of heart disease using artificial neural network. 13. 102-112.

5. Umair Shafique, Irfan Ul Mustafa, Haseeb Qaiser, Fiaz Majeed, "Data Mining in Healthcare for Heart Diseases",https://www.researchgate.net/ publication/ 274718934_Data_Mining_in_Healthcare_for_Heart_ Diseases. March 2015.

7. Akhil, Jabbar & Samreen, Shirina. (2016). Heart disease prediction system based on hidden naïve Bayes classifier. 10.1109/CIMCA.2016.8053261.

8. Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.

9. Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International Mutli-Conference on Automation, Computing,Communication, Control.

10. Kiyasu J Y (1982). U.S. Patent No. 4,338,396.Washington, DC: U.S. Patent and Trademark Office.

11. Raihan M, Mondal S, More A, Sagor M O F, SikderG, Majumder M A & Ghosh K (2016, December). Smartphone based ischemic heart disease(heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016.

12. Saw, Montu & Saxena, Tarun & Kaithwas, Sanjana & Yadav, Rahul & Lal, Nidhi. (2019). Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning. 10.1109/ICCCI48352.2020.9104210.

13. Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International Mutli-Conference on Automation, Computing,Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE.

6. Napa, Komal Kumar & Sindhu, G.Sarika & Prashanthi, D.Krishna & Sulthana, A.Shaeen. (2020). Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers. 15-21. 10.1109/ICACCS48705.2020.9074183. 340885231_Analysis_and_Prediction_of_Cardio_Vascular_Disease_using_Machine_Learning_Classifier s, April 2020 scular_Disease_using_Machine_Learning_Classifier s, April 20.

ayUSHpLAG

ORIGINALITY REPORT

10%	5%	7%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Kushagra Srivastava, Preeti Garg, Vineet Sharma, Nikhil Gupta. "Comparative Analysis of Various Machine Learning Techniques to Predict Breast Cancer", 2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2022 Publication	1%
2	www.ijraset.com Internet Source	1%
3	Submitted to Liverpool John Moores University Student Paper	1%
4	www.researchgate.net Internet Source	1%
5	www.iksadamerica.org Internet Source	1%
6	patentimages.storage.googleapis.com Internet Source	1%
7	Submitted to Coventry University	

Student Paper

<1%

8 Submitted to University of Wollongong
Student Paper

<1%

9 Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, Preeti Nagrath. "Heart disease prediction using machine learning algorithms", IOP Conference Series: Materials Science and Engineering, 2021
Publication

<1%

10 Maria Yousef, Prof. Khaled Batiha. "Heart Disease Prediction Model Using Naïve Bayes Algorithm and Machine Learning Techniques", International Journal of Engineering & Technology, 2021
Publication

<1%

11 www.ijcaonline.org
Internet Source

<1%

12 S.K.L. Sameer, P. Sriramya. "Improving the Accuracy for Prediction of Heart Disease by Novel Feature Selection Scheme using Decision tree comparing with Naive-Bayes Classifier Algorithms", 2022 International Conference on Business Analytics for Technology and Security (ICBATS), 2022
Publication

<1%

13	Submitted to The University of the West of Scotland Student Paper	<1%
14	kipdf.com Internet Source	<1%
15	www.irjmets.com Internet Source	<1%
16	"Application of Data Mining Techniques on Heart Disease Prediction: A Survey", Emerging Research in Computing Information Communication and Applications, 2016. Publication	<1%
17	Binju Saju, V Asha, Arpana Prasad, Harish Kumar P, Rakesh V, A.P. Nirmala. "Heart Disease Prediction Model using Machine Learning", 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS), 2022 Publication	<1%
18	"Artificial Intelligence and Data Science", Springer Science and Business Media LLC, 2022 Publication	<1%
19	Usman Ahmed, Jerry Chun-Wei Lin, Gautam Srivastava. "Multivariate time-series sensor vital sign forecasting of cardiovascular and	<1%

chronic respiratory diseases", Sustainable Computing: Informatics and Systems, 2023
Publication

20	academic-accelerator.com Internet Source	<1 %
21	www.atlantis-press.com Internet Source	<1 %

Exclude quotes Off
Exclude bibliography On
Exclude matches Off