Heart Disease Prediction using various Machine Learning Techniques

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

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

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

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 neural network, 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.

1. INTRODUCTION

Heart disease is a vague term which includes different types of heart problem. It's also commonly known as cardiovascular disease, meaning heart and blood vessel disease. 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.

Heart disease is a major cause of despair and mortality among the people. About 610,000 people in United Stated die from heart disease alone which means 1 in every 4 deaths. 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 branch of artificial intelligence which mainly aims on correct use of data and other techniques to mimic the way humans understand and improve the accuracy gradually. They are trained to determine the relationship between data and recognise various patterns. 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.

A system can be developed that can predict many disease by using some ML algorithms such as Naive Bayes, K-NN, Random forest, Decision Tree (DT), and SVM algorithms. 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.

2. 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.

According to a different study [5], the artificial neural network exhibits the best accuracy of 84.25% compared to other models. Interestingly, despite other algorithms showing greater accuracy than ANN, this model with lesser accuracy is selected as the final main model.

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 encompasses a concise literature survey review, presenting an effective prediction of chronical heart disease using various algorithms, including Logistic Regression, KNN, and Random Forest Classifier, among others. The Results illustrate that each algorithm possesses strengths in achieving defined objectives [8].

The model including IHDPS demonstrates the potential to analysis the decision boundary using both older and newer machine learning technics and deep learning models. It facilitates crucial factors/knowledge for example family history associated with heart disease. However, the accuracy achieved in the IHDPS model falls significantly when compared to latest emerging models, particularly in detecting chronical heart disease using artificial neural networks and other machine learning techniques and deep learning algorithms. McPherson et al. [9] identified risk factors for coronary heart disease or atherosclerosis using an inbuilt implementation algorithm employing techniques of Neural Network, accurately predicting the presence of the disease in test patients.

R. Subramanian et al. [10] introduced the diagnosis and prediction of Heart Disease and Blood Pressure, incorporating neural networks. They built a deep Neural Network with attributes related to the disease, producing an output processed by the output perceptron and including close to 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.

3. 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. The prediction of type of chronicle heart disease can be done with the help of analysis of various machine learning algorithms. We will be using different types of python libraries, for example numpy, panda and matplotlib.

Further more, the implementation used Machine Learning part of the project and utilizes sklearn and keras. the data set is available by the University of California Arvind Irvine machine learning repository. The data set includes patient's data regarding heart disease, diagnosis, history that was collected from several different location around the world. There are 76 different attributes including age, sex, resting blood pressure, cholesterol levels, echocardiogram, data, exercise, habits and many others. Throughout all these data we mainly focus on a subset of 14 attributes, most precisely, we will use the data collected at Cleveland clinic foundation.

- **KNN (K-nearest neighbour) Algorithm**: K-nearest algorithm also majorly called as KNN is a supervised learning classifier, which focuses on utilizes proximity for classifications and predictions related to clustering of an individual data points. It can also be utilized for both regression and classification type problems, but it is majorly used for classification algorithm. For classification issue, a class label is designed on the basis of majority vote.
  The accuracy of K-NN is found to be 85.06%.The graph below shows K Neighbour classifier scores for different K values.. Number of K neighbours is 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%.

- **Naïve Bayes:** Bayes' theorem has a very fundamental importance in statistics and many newly advanced machine learning models. Bayesian reasoning is an approach where we update the probability of hypothesis when we get a new evidence, thus factoring a major role in science[13]. Bayesian analysis allowed the mathematicians to answer the question which were not previously answered using frequentist statistical approach. In fact, the whole inspiration to assign a probability to a hypothesis is not a part of frequentist paradigm. The attributes that were involved for deterring the CHDs were age, gender, test bps, cholesterol level, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal. Accuracy using Naive bayes algorithm is 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 endeavours to enhance the precision of Heart Disease prediction using the Logistic Regression model in machine learning. The study focuses on a healthcare dataset, classifying patients based on their heart disease status, aiming to overcome predictive limitations inherent in current healthcare practices [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%.

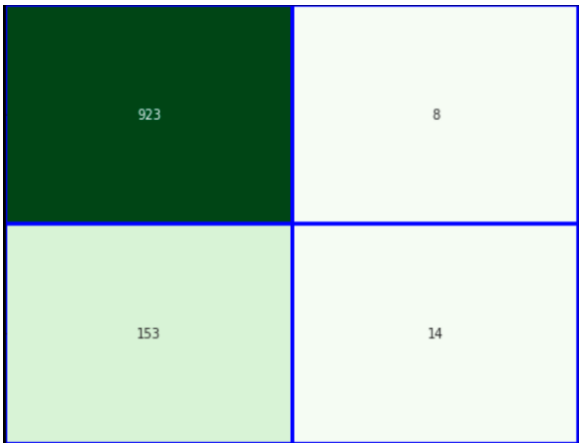

Fig 1: Confusion Matrix for Logistic Regression

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.

4. RESULT AND DISCUSSION

In our study on predicting chronicle heart disease and utilization of different machine learning algorithms, we employed a variety of techniques The primary objective was to evaluate the performance of these algorithms in accurately classifying patients based on their risk of heart disease. Our experimentation revealed interesting results regarding the accuracy of each algorithm. ANN emerged as the top-performing algorithm, achieving the highest accuracy of 85%, followed closely by SVM with an impressive accuracy of 94.41%. KNN, Naive Bayes, and Decision Trees demonstrated reasonable accuracies, approximately 78%, indicating their effectiveness in heart disease prediction.

The differences in accuracy among the algorithms can be attributed to various factors, including the complexity of the models, the nature of the dataset, and the hyper parameters used during training. 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. However, it is necessary to acknowledge the demerits of our work and models. The performance and efficiency of the machine learning algorithms greatly depends on the supervised data and its quality and quantity of availability, as well as the feature selection process. Furthermore, the generalizability of our models to diverse populations and clinical settings warrants further investigation.

As our final results, our research mainly demonstrates the potential and accuracy of machine learning algorithms in enhancing heart disease prediction systems. By leveraging advanced computational techniques, we can assist healthcare professionals in early detection and personalized management of heart disease, ultimately improving patient outcomes and reducing healthcare burdens.

5. COMPARISON OF DIFFERENT ALGORITHMS

Our research explores the efficiency of various machine learning algorithms for heart disease prediction. We compare the performance of Naive Bayes, K-Nearest Neighbours (KNN), Decision Trees, Support Vector Machines(SVM), and Artificial Neural Networks (ANN) based on their accuracy and suitability for this task

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 basicpatterns in the dataset.
- KNN also achieves an accuracy of approximately 85.06%, indicating its effectiveness in capturing localpatterns 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. Support Vector Machines (SVM) and Artificial Neural Networks (ANN) are more complex algorithms capable of capturing nonlinear relationships in the data.

- 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.

While SVM outperforms ANN in terms of accuracy, ANN offers a balance between performance and interpretability, making it a viable option for heart disease prediction systems where transparency is important. Figure 2 shows the accuracy achieved by various techniques. Here the highest accuracy achieved is using SVM classification.

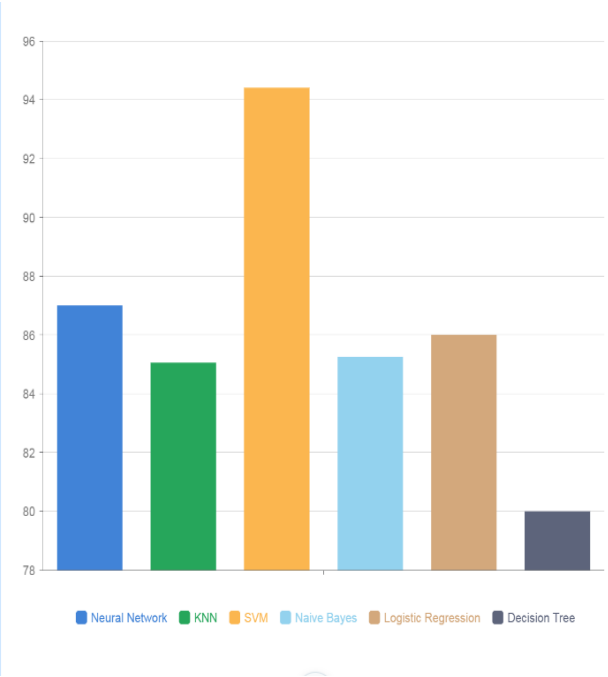

Fig 2: Accuracy comparison

6. CONCLUSION

In this study, the authors developed a heart disease prediction system utilizing machine learning techniques and algorithms, including Naive Bayes, K-NN, Decision Trees, SVM, and Artificial Neural Networks (ANN). Our aim was to accurately classify patients based on their risk of heart disease using relevant medical attributes. 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 findings underscore the potential of machine learning algorithms in aiding medical professionals in the early detection and prediction of heart disease. These predictive models can assist in preventive healthcare measures and personalized treatment strategies, thereby reducing the burden of heart disease-related fatalities.

However, it's very necessary to acknowledge the demerits of our study. The working of machine learning algorithms greatly relies on the quality and quantity of data available, as well as the feature selection process. Additionally, further validation of our models on diverse datasets and in clinical settings is warranted to assess their generalizability and real-world applicability.

In conclusion, our research and efforts adds to the ongoing efforts in leveraging and using machine learning for improving chronical heart disease prediction systems. By harnessing the potential of data-driven approaches, we can potentially revolutionise the field of cardiology, leading to better patient outcomes and reduced healthcare costs.

7. 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.
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.

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.

14. Colleen M. Norrisa,b,*, William A. Ghalid,e,f , L. Duncan Saundersc , Rollin Brante , Diane Galbraithd,f , Peter Farise , Merril L. Knudtsond , for the APPROACH Investigators a Faculty of Nursing, University of Alberta, Edmonton, Alberta, Canada bDivision of Cardiology, Faculty of Medicine, University of Alberta, 4-130F Clinical Sciences Building, Edmonton, Alberta T6G 2G3, Canada cDepartment of Public Health Sciences, University of Alberta, Edmonton, Alberta, Canada dDepartment of Medicine, University of Calgary, Calgary, Alberta, Canada eDepartment of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada fCentre for Health and Policy Studies, University of Calgary, Calgary, Alberta, Canada Accepted 12 September 2005