International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

Heart Disease Prediction - A Review through Machine Learning Technologies

Varsha Sharda, Raghvi Vyas, Rishika Meghani, Ratnam Luhadiya

Abstract: Heart diseases are one of the most fatal threats to humankind. The biggest reason is that these are the most difficult diseases to diagnose due to the complex nature of the organ, the vitality of which is not hidden from the human race. The term "Heart Disease" is also an umbrella term for a range of conditions which affect the heart. According to an article, Heart diseases were singled out as one of the primary factors of death. The article read: "About 697,000 people die of heart Disease in the United States every year - that's 1 in every 5 deaths. Heart disease is the leading cause of death for both men and women. Coronary Heart Disease (CHD) is the most common type of heart disease, killing millions annually." Machine Learning technology has been proven as an effective tool in assisting and making quick and early decisions and diagnosis.

Introduction:

Posing a major threat in the medical world, heart disease is a serious concern among researchers and healthcare professionals. Techniques developed earlier in this realm have not proven to be as efficient, even by professional researchers [3].

There is a large group of people who, even after having the necessary resources succumb to heart diseases, just because of late detection. An early diagnosis can help people seek timely assistance. The general life expectancy of a region can be increased by early detection and treatment of heart diseases.

By enabling healthcare professionals to identify individuals at high risk and initiate preventive measures, AI-ML technologies have emerged as a promising tool for predicting heart disease. Although ML technologies still haven't proven to be a sure solution to the problem, the accuracy is increasing with each attempt. AI takes large data sets, from patients in different settings, studies the diagnosis and forms trends based on different inputs given.

This review paper presents a heart disease prediction system created using Python and Django, which utilises gradient boosting algorithm, logistic regression with decision tree algorithms, and naive Bayes algorithm for predicting the likelihood of heart disease in individuals. One of the major limitations of the system however, is the accuracy with which it draws results. The system is used to warn users of a potential underlying heart disease. If received a warning, the user can further consult a doctor and get professional assistance.

Related Works:

Paper [5] has illustrated how the datasets are generally raw in nature, they are redundant and inconsistentThey need to be pre-processed; high dimensional data set is reduced to low data set. crucial features from the data set are extracted because they contain a vast array of information. Filtration of relevant features reduces work of training the algorithms and hence results in reduction in time complexity. Apart from time, other parameters like accuracy also play vital roles in proving effectiveness of these algorithms. Data from UCI repository is used in Paper [6] to evaluate performance of different machine learning algorithms using Naive Bayes, KNN, Decision Tree, ANN. The results were as follows: ANN, with the highest accuracy of 85.3%, Naïve Bayes and KNN gave almost 78% and Decision Tree gave 80%.

Paper [7] uses the WEKA tool for measuring performance of different machine learning algorithm.ANN with PCA is implemented to increase the speed. With an accuracy of 94.5% and 97.7%before and after applying PCA, respectively, a significant difference was noticed [7]. Apart from the decision tree, various other approaches were also adopted for achieving accurate detection of heart disease in humans [8] collected raw data from an EEG device which was used to train neural networks for pattern classification . Here input outputs are depressive and non depressive categories in the hidden layer. For training to achieve efficient results, a scaled conjugate gradient algorithm was used. Efficiency of up to 95% was achieved with help of trained neural network.This technique to classify and achieve better results in cases where the feature vectors are multi dimensional and non linear. Because of its capability to work under datasets of high dimensionality, these methods proved superior to all other existing quantum contemporary techniques.

Methodology:

The heart disease prediction system developed in this project comprises several stages, including data preprocessing, feature selection, model selection, and performance evaluation. To be considered suitable for analysis, the data must be preprocessed. The preprocessing stages involve cleaning, transforming, and normalising the input data.

Feature selection is performed to identify the most relevant features for predicting heart disease. The system takes several inputs such as age, gender, cholesterol level, blood pressure, and smoking status. The selected features are then used to train the machine learning models. The performance of the models is evaluated using various metrics, including accuracy, precision, recall, and F1-score.

Decision Tree:

An easy to understand, supervised learning algorithm classifier, is the decision tree algorithm.The decision tree resembles the tree structure. There are internal nodes, branches, and leaf nodesEach branch represents the values of a given data set, internal nodes Tests on a given attribute and the Leaf nodes show the class to predict or indicate the results. Depending on the predictive attribute

and the given rules, the classification rule starts from the root node to the leaf nodes,dealing with numerical and categorical data.
Commonly used decision tree algorithms are CART, ID3, C4.5, J48 and CHAID are very important in the prediction of diseases [6].

Artificial Neural Networks:
Artificial Neural Networks are the human neurons type network structure which consists of a number of nodes that are connected through directional links where each node represents a processing unit and the links between them designate the causal relation between them [7]. ANN enables clinical decision making, assisting doctors in analysis and accurate detections

Gradient Boosting:
The gradient boosting algorithm is a machine learning technique that is used for both regression and classification problems. It is an ensemble method that combines multiple weak models into a strong model, by iteratively training them on the residuals of the previous model. In other words, it focuses on reducing the errors of the previous models by adding new models that can handle the remaining errors. The algorithm works by building a decision tree based on the features of the data, and then using a gradient descent optimization method to minimise the loss function. Gradient boosting has been shown to be highly effective in many real-world applications, such as image classification, speech recognition, and fraud detection. It is known for its ability to handle complex relationships between features and labels, and for its high accuracy and robustness. However, it is computationally expensive and requires careful hyperparameter tuning to achieve optimal results

Logistic Regression with Decision Tree Algorithms:
Logistic regression with a decision tree algorithm is a hybrid approach that combines the strengths of two powerful machine learning techniques. Logistic regression is a classification algorithm that is used to predict the probability of a binary outcome based on a set of input variables. On the other hand, decision trees are a popular machine learning method for solving classification and regression problems. The hybrid approach of logistic regression with decision tree algorithm combines these two techniques to improve classification accuracy. The decision tree component of the algorithm helps to identify the most important features for classification, while logistic regression is used to build a model that can predict the probability of a binary outcome based on these features. This approach has proven to be highly effective in many real-world applications, such as credit scoring, fraud detection, and medical diagnosis. With its ability to handle complex datasets and produce highly accurate predictions, logistic regression with decision tree algorithm is a valuable tool for data scientists and machine learning practitioners

Naive Bayes Algorithm:

The Naive Bayes algorithm is a probabilistic machine learning algorithm based on Bayes' theorem. It is a probabilistic classifier that predicts the class of a data point based on the probabilities of it belonging to different classes. The algorithm makes the "naive" assumption that all features in the data are independent of each other, hence the name "naive Bayes". Despite this assumption, Naive Bayes has proven to be a highly effective algorithm in many classification tasks, such as text classification, spam filtering, sentiment analysis, and image recognition. The algorithm is simple to implement, computationally efficient, and requires only a small amount of training data. Due to its simplicity and effectiveness, Naive Bayes is a popular choice for many real-world applications. Naive Bayes is used to compute posterior probability of each class based on conditional probability of classifying data sets [7]. It has been used in predicting the risk of heart disease. A study [2]used the Naive Bayes algorithm to predict the risk of coronary heart disease. The study found that the algorithm had an accuracy of 0.90, outperforming other machine learning algorithms.

The heart disease prediction system developed in this project comprises several stages, including data preprocessing, feature selection, model selection, and performance evaluation. The data preprocessing stage involves cleaning, transforming, and normalising the input data to ensure that it is suitable for analysis, using KDP and data mining. Feature selection: performed to identify the relevant features for the prediction process. The selected features are then used to train the machine learning models, including gradient boosting algorithm, logistic regression with decision tree algorithms, and naive Bayes algorithm.

The system has been trained on a large dataset, comprising over 300,000 records, which increases the reliability, efficiency and robustness of the system.

High accuracy in predicting the risk of heart disease. Identifies the most important risk factors for heart disease, allowing for targeted interventions and prevention strategies. Even if there is a slight risk detected, the patient can seek professional help. Uses AI-ML algorithms which can continually learn and adapt to new data, potentially improving the accuracy of the predictions over time. Moreover, the system is user-friendly and accessible, with a web interface that can be easily used by healthcare professionals and patients alike. The user only needs to input the required data in the specified fields, the system will take care of the rest, giving the user, clear and concise summary.

Discussion:

The heart disease prediction system developed in this project has several advantages, including its ability to predict heart disease with high accuracy, its user-friendliness,

and it's easy integration with existing healthcare systems. The use of AI-ML algorithms in predicting the risk of heart disease has shown promising results.

Gradient boosting, logistic regression with decision tree algorithms, and Naive Bayes algorithm have all been used to accurately predict the risk of heart disease.

These algorithms have demonstrated high accuracy in predicting the risk of heart disease and identifying important risk factors, such as age, gender, cholesterol level, blood pressure, and ECG. The Heart Disease Prediction system enables anyone with access to the internet to get their result, regardless of their location, background or financial status. This is an important factor affecting the people in communities that have limited resources. However, the system's limitations include the need for large amounts of data for accurate predictions and the possibility of biassed results due to the quality and representativeness of the data. These algorithms rely heavily on accurate and complete data inputs. If the data is incomplete or inaccurate, the predictions may be less accurate.
Additionally, the algorithms may not be able to account for all risk factors for heart disease, which could potentially lead to missed diagnoses or inaccurate predictions.
The algorithms may also require ongoing maintenance and updates to ensure that they are up-to-date and accurate. The system's performance may vary depending on the specific population and demographics being analysed. Future research can focus on addressing the limitations of the heart disease prediction system developed in this project, including the need for large amounts of data and the potential for biassed results. Much more optimal than MLP, J48 and KStar, an approach proposed in[5] has worked to improve the accuracy and found that performance of Bayes Net and SMO classifiers increased significantly. Additionally, the system's performance can be evaluated on different populations and demographics to determine its generalizability. Finally, the development of more advanced AI-ML algorithms may further improve the accuracy and effectiveness of heart disease prediction systems.

Conclusion:

The heart disease prediction system developed using Python and Django and based on AI-ML technologies has significant potential in aiding in the early detection and prevention of heart disease. It is an effective tool for predicting the risk of heart disease. The system uses several AI-ML algorithms to accurately predict the risk of heart disease based on several input factors. The system demonstrated high accuracy in predicting the risk of heart disease and identified the most important risk factors for heart disease. The system can be used as a screening tool to identify patients at high risk. The system's accuracy and reliability make it a valuable tool for healthcare professionals in identifying individuals at high risk and initiating preventive measures. However, further research is required to address the limitations of the system and improve its accuracy and effectiveness. Overall, the heart disease prediction system is a promising tool in the fight against heart disease.

References:

[1] Center for Disease Control and Prevention, Heart Disease Facts, "https://www.cdc.gov/heartdisease/facts.htm".

[2] International Journal on Recent and Innovation Trends in Computing and Communication, ISSN: 2321-8169, Prediction of Heart Disease using Machine Learning Algorithms: A Survey, Volume: 5 Issue: 8

[3] M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, "Prediction of risk score for heart disease using associative classification and hybrid feature subset selection," Int. Conf. Intell. Syst. Des. Appl. ISDA, pp. 628–634, 2012.

[4] Harshit Jindal,Rishabh Khera and PreetiNagrath,"Heart disease prediction using machine learning algorithms", doi:10.1088/1757-899X/1022/1/012072

[5] M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," 2016 3rd Int. Conf. Electr. Eng. Inf. sCommun. Technol. iCEEiCT 2016, 2017

[6] Muhammad Usama Riaz, SHAHID MEHMOOD AWAN, ABDUL GHAFFAR KHAN, "PREDICTION OF HEART DISEASE USING ARTIFICIAL NEURAL NETWORK", https://www.researchgate.net/publication/328630348_PREDICTION_OF_HEART_DISEASE_USING_ARTIFICIAL_NEURAL_NETWORK. October 2018

[7] 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.

[8] S. Kumra, R. Saxena, and S. Mehta, "An Extensive Review on Swarm Robotics," pp. 140–145, 2009.