Utilizing Machine Learning Classification for Heart Disease Identification in E-Healthcare: A Comprehensive Approach

Palegari Saraswathamma

Department of Computer Science Sri Venkateswara University, Tirupati

Abstract— *This research focuses on early detection of heart disease symptoms using patient data and real-time user input. Modern healthcare data includes detailed demographic and symptom information, allowing for comprehensive analysis. Our proposed method utilizes this data for classification, comparing recent healthcare data with baseline distributions. Machine learning techniques such as Logistic Regression, K-Nearest Neighbors, Random Forest, and XGBoost are employed for training and testing. Classifier performance is evaluated to make predictions.*

I. INTRODUCTION

Cardiac arrest is a fatal event caused by sudden coronary thrombosis, resulting in the death of heart muscle. It occurs when oxygen-rich blood flow to the heart is blocked due to plaque buildup in the arteries, a condition known as atherosclerosis. This plaque can rupture, causing a blood clot that obstructs blood flow and leads to heart muscle death. The resulting damage may not be immediately apparent, leading to long-term complications. Most heart attacks are linked to coronary heart disease, where plaque accumulates in the coronary arteries. Early prediction is crucial for diagnosing cardiac issues and preventing fatalities. Severe spasms in coronary arteries can also trigger heart attacks, while complications like heart failure and life-threatening arrhythmias may arise. Factors such as education and lifestyle contribute to an individual's predisposition to cardiac disease, with socioeconomic status playing a significant role

II. LITERATURE REVIEW

2.1 Prediction of heart disease by using machine learning

Rohit Murty, Satish Patle, Saurabh Bute, Sneha Bhilkar, Durga Wanjari - 2020

With the rampant increase in the heart stroke rates at juvenile ages, we need to put a system in place to be able to detect the symptoms of a heart stroke at an early stage and thus prevent it. It is impractical for a common man to frequently undergo costly tests like the ECG and thus there needs to be a system in place which is handy and at the same time reliable, in predicting the chances of a heart disease. Thus, we propose to develop an application which can predict the vulnerability of a heart disease given basic symptoms like age, sex, pulse rate etc. The machine learning algorithm neural networks has proven to be the most accurate and reliable algorithm and hence used in the proposed system.

2.2 Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset

Tapas Ranjan Baitharua, Subhendu Kumar Panib - 2016

The accuracy of data classification relies heavily on the dataset used for training. Liver problems have become a leading cause of death worldwide, yet the vast amount of healthcare data remains underutilized. Our research aims to leverage this data to develop intelligent medical decision support systems for liver disorder diagnosis. We propose using decision trees such as J48, Naive Bayes, ANN, ZeroR, 1BK, and VFI algorithms for disease classification and compare their effectiveness and accuracy rates. Early detection of liver disease is crucial for effective treatment, leading to improved performance of classification models and faster learning. We present a comparative analysis of classification accuracy using liver disorder data under various scenarios, quantitatively comparing the predictive performances of popular classifiers

2.3 Improving Disease Prediction by Machine Learning

Smriti Mukesh Singh1, Dr. Dinesh B. Hanchate2 – 2018

Nowadays, Big Data is increasingly used in healthcare for precise medical data analysis, benefiting early disease detection, patient care, and community services. Fragmented medical data reduces analysis accuracy, prompting the use of machine learning algorithms for chronic disease prediction. To address incomplete data, Genetic algorithms reconstruct missing data. The dataset comprises structured and unstructured data; RNN algorithm extracts features from unstructured data. The system proposes SVM and Naive Bayesian algorithms for disease prediction using structured and unstructured hospital data,

respectively. Additionally, a Community Question Answering (CQA) system is proposed to predict questions and answers, with KNN and SVM algorithms facilitating classification for optimal user responses related to diseases.

2.4 Heart Disease Prediction System Using Data Mining Techniques

Abhishiek Taneja – 2015

In today's modern world cardiovascular disease is the most lethal one. This disease attacks a person so instantly that it hardly gets any time to get treated with. So diagnosing patients correctly on timely basis is the most challenging task for the medical fraternity. A wrong diagnosis by the hospital leads to earn a bad name and loosing reputation. At the same time treatment of the said disease is quite high and not affordable by most of the patients particularly in India. The purpose of this paper is to develop a cost-effective treatment using data mining technologies for facilitating data base decision support system. Almost all the hospitals use some hospital management system to manage healthcare in patients. Unfortunately, most of the systems rarely use the huge clinical data where vital information is hidden. As these systems create huge amount of data in varied forms but this data is seldom visited and remain untapped. So, in this direction lots of efforts are required to make intelligent decisions. The diagnosis of this disease using different features or symptoms is a complex activity. In this paper using varied data mining technologies an attempt is made to assist in the diagnosis of the disease in question.

2.5 Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques

Chaitrali S. Dangare, Sulabha S. Apte - 2016

The healthcare sector possesses abundant data, yet not all are effectively utilized for insightful decision-making. Advanced data mining techniques are crucial, especially for predicting heart diseases. This study enhances heart disease prediction systems by incorporating additional attributes like obesity and smoking alongside existing ones such as sex, blood pressure, and cholesterol. Three classification techniques - Decision Trees, Naive Bayes, and Neural Networks - are evaluated based on accuracy using these attributes. Results indicate that Neural Networks achieve the highest accuracy (100%), followed by Decision Trees (99.62%), and Naive Bayes (90.74%), showcasing their effectiveness in heart disease prediction.

Problem Statement

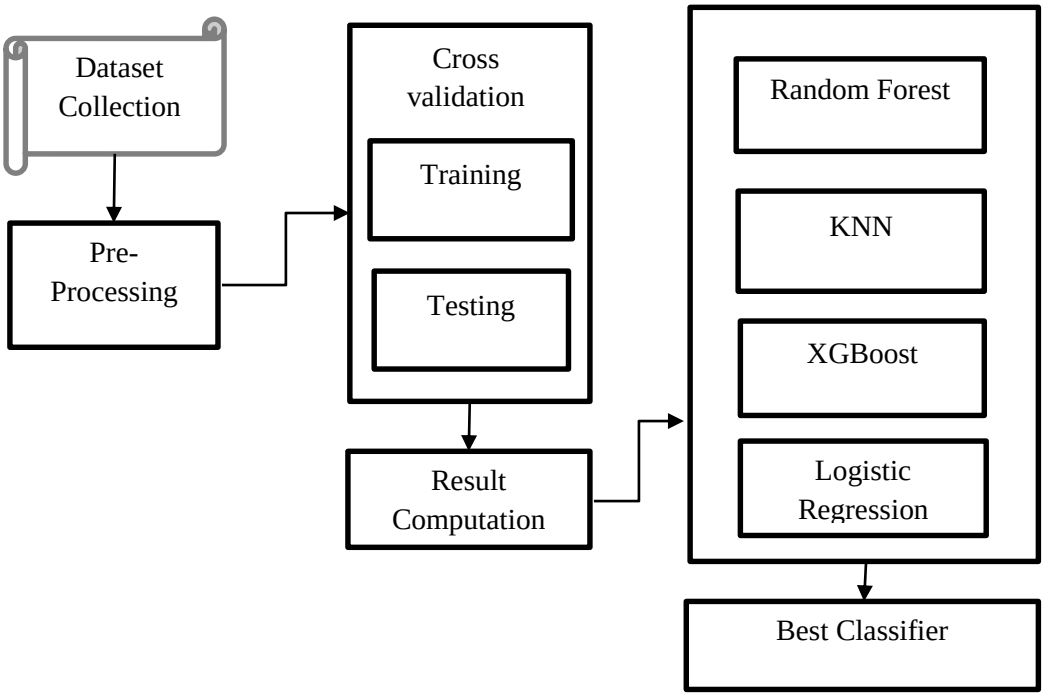
- A heart attack happens if the flow of oxygen-rich blood to a section of heart muscle suddenly becomes blocked and the heart can't get oxygen.
- The heart damages caused long-lasting and severe problems. A majority of the heart attacks occur as a result proportion to coronary heart disease.
- Coronary heart disease is a condition in which a wax like substance that can be termed as plaque builds up inside of the coronary arteries.
- Only early prediction could help to better diagnose the cardiac problems at the benign stage to save a person's life.

III. PROPOSED SYSTEM

- Our proposed system involves Logistic regression, XG Boost, Random Forest classifier, KNN Algorithm in Machine Learning concept used to train the dataset.
- Thus, preventing heart diseases has become more than necessary. Good data-driven systems for predicting heart diseases can improve the entire research and prevention process, making sure that more people can live healthy lives.
- This is where Machine Learning comes into play. Machine Learning helps in predicting the heart diseases, and the predictions made are quite accurate.
- The project involved analysis of the heart disease patient dataset with proper data processing. Then, different models were trained and predictions are made with different algorithms KNN, Decision Tree, Random Forest, Logistic Regression etc.
- This is the jupyter notebook code and dataset I've used for my Kaggle kernel 'Binary Classification with Sklearn and Keras'.
-

Advantages

- Easy detection of the Cardiac disease with the concluded technique.
- Time consuming.
- Best accuracy Model helps in better treatment as early.
- Detection of best Model will quick the treatment which is life saving



IV. IMPLEMENTATION

- Dataset Collection and pre-processing
- Train the model
- Evaluation
- Comparison of existing model
- Performance analysis

4.1 Dataset Collection and Pre-processing

A dataset is essentially a structured collection of data, often organized in tabular form with columns representing variables and rows corresponding to individual entries. Each value within this structure is referred to as a datum. We've opted to work with a publicly-available healthcare dataset, designed to facilitate easy comprehension for medical professionals, allowing them to apply both traditional statistical methods and modern machine learning techniques. Moreover, this dataset's compact size ensures efficient computation on most modern computer systems. The sklearn.preprocessing package offers various tools and transformer classes to preprocess raw feature vectors, making them more compatible with subsequent machine learning models."

4.2 Train the Model

This stage involves evaluating models based on input data. In our study, we'll train the model using four machine learning algorithms to predict heart disease. The K-nearest neighbors algorithm, a supervised classification method, determines object classification based on the nearest neighbors. It calculates distance using Euclidean distance, clustering data based on similarity and filling missing values using K-NN. Various prediction techniques can be applied once missing values are filled, potentially improving accuracy. The random forest algorithm, another supervised classification technique, builds a forest of trees where

each tree provides a class prediction. The class with the most votes becomes the model's prediction. Higher accuracy is achieved with more trees in the random forest classifier. The three common methodologies are:

- Forest RI (random input choice);
- Forest RC (random blend);
- Combination of forest RI and forest RC.

It is used for classification as well as regression task, but can do well with classification task, and can overcome missing values. Besides, being slow to obtain predictions as it requires large data sets and more trees, results are unaccountable.

Random forest algorithm has obtained an accuracy of 91.6% with Cleveland dataset in Using People's dataset, an accuracy of 97% was achieved.

4.3 Evaluation

This stage is to form evaluation the models based on the input data. For our purpose of study, we are going to implement the model using XGboost classifier algorithm in Machine Learning.

We will split the data set into test and train set. After splitting the data first have to train the data and test the data using XGboost classifier, Logistic regression, KNN, Random Forest in Machine learning techniques.

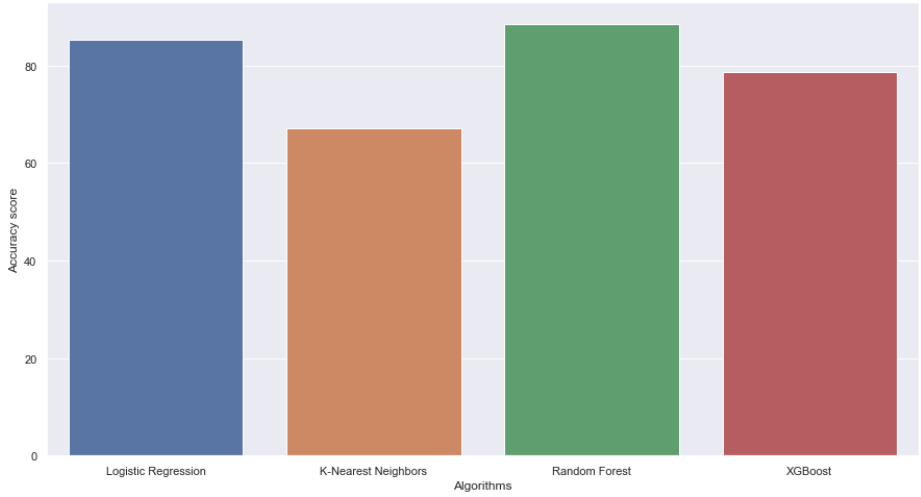
4.4 Comparison of Existing model

This module includes comparison of existing system algorithm accuracy and our proposed model accuracy. Our aim is to improve the accuracy score.

4.5 Performance Analysis

The next stage is to predict the results using Classifier. The best method for the training and test data set is definitely given has the best results for Classification Accuracy and Recall for both validation cases. Now we forward this Random Forest classifier to next stage to predict the disease that may further lead to a cardiac arrest.

The results are compared using a confusion matrix. The consistency of a classification model can be well visualized with a tabular form also called Confusion Matrix (or "classifier") which shows its results over a set of known test data.



The accuracy score Logistic Regression is: 85.25 %

The accuracy score K-Nearest Neighbors is: 67.21 %

The accuracy score Random Forest is: 88.52 %

The accuracy score XGBoost is: 78.69 %

Random Forest Gives Good Accuracy

V. CONCLUSION AND FUTURE WORK

In this paper, a reliable multi-process Machine Learning method for building a Heart disease risk prediction system is proposed, showing higher accuracy compared to existing systems. Detecting heart disease symptoms early is crucial to reduce mortality rates. The study aims to define effective data mining techniques for heart disease prediction using only 14 essential attributes. Four data mining classification techniques were applied, with K-nearest neighbor and random forest showing the best results. Further research can explore additional data mining techniques like time series and clustering to improve accuracy for early prediction of heart disease.

REFERENCES

[1] Baitharu, Tapas Ranjan, and Subhendu Kumar Pani. "Analysis of data mining techniques for healthcare decision support system using liver disorder dataset." *Procedia Computer Science* 85 (2016): 862-870.

[2] Gavhane, Aditi, Gouthami Kokkula, Isha Pandya, and Kailas Devadkar. "Prediction of heart disease using machine learning." In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1275-1278. IEEE, 2018.

[3] Singh, Smriti Mukesh, and Dinesh B. Hanchate. "Improving disease prediction by machine learning." *Int J Res Eng Technol* 5, no. 6 (2018): 1542-1548.

[4] Taneja, Abhishek. "Heart disease prediction system using data mining techniques." *Oriental Journal of Computer science and technology* 6, no. 4 (2013): 457-466.

[5] Dangare, Chaitrali S., and Sulabha S. Apte. "Improved study of heart disease prediction system using data mining classification techniques." *International Journal of Computer Applications* 47, no. 10 (2012): 44-48.

[6] Thomas, J., and R. Theresa Princy. "Human heart disease prediction system using data mining techniques." In *2016 international conference on circuit, power and computing technologies (ICCPCT)*, pp. 1-5. IEEE, 2016.

[7] Kaur, Beant, and Williamjeet Singh. "Review on heart disease prediction system using data mining techniques." *International journal on recent and innovation trends in computing and communication* 2, no. 10 (2014): 3003-3008.

[8] Meyer, Alexander, Dina Zverinski, Boris Pfahringer, Jörg Kempfert, Titus Kuehne, Simon H. Sündermann, Christof Stamm, Thomas Hofmann, Volkmar Falk, and Carsten Eickhoff. "Machine learning for real-time prediction of complications in critical care: a retrospective study." *The Lancet Respiratory Medicine* 6, no. 12 (2018): 905-914.

[9] Rajkomar, Alvin, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. "Ensuring fairness in machine learning to advance health equity." *Annals of internal medicine* 169, no. 12 (2018): 866-872.

[10] Rajamhoana, S. P., C. Akalya Devi, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika. "Analysis of neural networks based heart disease prediction system." In *2018 11th International Conference on Human System Interaction (HSI)*, pp. 233-239. IEEE, 2018.

[11] Ramalingam, V. V., Ayantan Dandapath, and M. Karthik Raja. "Heart disease prediction using machine learning techniques: a survey." *International Journal of Engineering & Technology* 7, no. 2.8 (2018): 684-687.

[12] Kohli, Pahulpreet Singh, and Shriya Arora. "Application of machine learning in disease prediction." In *2018 4th International conference on computing communication and automation (ICCCA)*, pp. 1-4. IEEE, 2018.

[13] Marimuthu, M., M. Abinaya, K. S. Hariesh, K. Madhankumar, and V. Pavithra. "A review on heart disease prediction using machine learning and data analytics approach." *International Journal of Computer Applications* 181, no. 18 (2018): 20-25.

[14] Beyene, Chala, and Pooja Kamat. "Survey on prediction and analysis the occurrence of heart disease using data mining techniques." *International Journal of Pure and Applied Mathematics* 118, no. 8 (2018): 165-174.

[15] Khourdifi, Youness, and Mohamed Bahaj. "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization." *International Journal of Intelligent Engineering & Systems* 12, no. 1 (2019): 242-252.