International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-779

DETECTION OF HEART DISEASE USING MACHINE LEARNING ALGORITHM

VIKAS S G, YUVANANDHINI E, APARNA N

[1]*Student, Dept. of Information Technology, Anna University, IN*
[2]*Student, Dept. of Food Technology, Anna University, IN*
[3]*Student, Dept. of Food Technology, Anna University, IN*

Abstract - In today's digital era, data is a valuable asset, and the healthcare industry generates vast amounts of information related to patients. Cardiovascular diseases are rapidly increasing due to busy and stressful lifestyles, affecting people of all age groups and necessitating early detection based on symptoms or medical reports. Early identification and treatment are crucial for reducing the risks associated with these chronic diseases. The primary objective of developing a heart disease prediction system is to assist doctors in cross-verifying their diagnoses, offering a promising solution to reduce mortality rates. Recent advancements in medical technologies, including data mining and deep learning, have played a vital role in detecting cardiovascular diseases by analyzing factors such as age, chest pain type, blood pressure, and cholesterol levels. Our proposed system aims to enhance medical care, lower costs, and provide significant insights for predicting heart disease, contributing to better health outcomes.*adoption.*

Key Words: Heart Disease Prediction, Cardiovascular Diseases, Data Mining, Deep Learning, Medical Technology, Early Detection, Healthcare Data, Risk Factors, Diagnosis Support, Cost Reduction, Patient Care.

1. INTRODUCTION

Machine Learning is a transformative tool that enables the extraction of valuable insights from data, offering significant potential across various domains, including healthcare. Cardiovascular diseases (CVD), the leading global cause of death with 17.9 million fatalities annually, can benefit immensely from machine learning applications. Our project aims to predict CVD risk by analyzing patients' medical histories and key attributes such as age, gender, chest pain, fasting sugar levels, and blood pressure. By identifying high-risk individuals early, the system can assist in diagnosis with fewer medical examinations, leading to more efficient and timely treatment.

To achieve this, we utilized a Kaggle dataset combining various medical records and applied three machine learning techniques: XGBoost, Random Forest, and KNN. By training these algorithms on 12 medical attributes, we achieved an accuracy of over 95%, with XGBoost performing the best. This combined approach outperforms systems relying on a single technique and offers a more reliable and cost-efficient method for predicting CVD. Our project has the potential to significantly improve the early detection and management of heart diseases, ultimately enhancing patient outcomes and reducing healthcare costs.

1.1 Background of the Work

Cardiovascular diseases (CVD) are the leading global cause of death, necessitating accurate early detection. Traditional methods are often limited, prompting the use of machine learning techniques like XGBoost, Random Forest, and KNN. Using patient medical data from a Kaggle dataset, our project achieves over 95% accuracy in predicting CVD risks. This system offers a reliable, cost-effective solution for enhanced healthcare outcomes.

1.2 Motivation and Scope of the Proposed Work

The main motivation of doing this research is to present a heart disease prediction model for the prediction of occurrence of heart disease. Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient. This work is justified by performing a comparative study and analysis using classification algorithms namely, XG Boost, Logistic Regression, KNN, and Random Forest are used at different levels of evaluations. Although these are commonly used machine learning algorithms, the heart disease prediction is a vital task involving highest possible accuracy. Hence, these algorithms are evaluated at numerous levels and types of evaluation strategies. This will provide researchers and medical practitioners to establish a better way to detect heart disease.

1. The system will help identify important factors that lead to a heart disease.
2. The main scope of the project is to detect heart disease.
3. It will help the patients to obtain results quick and diagnose as early as possible.

2. METHODOLOGY

Our methodology involves using a Kaggle dataset containing medical attributes like age, gender, chest pain, and blood pressure. Machine learning algorithms—XGBoost, Random Forest, and KNN—are trained and tested on this data to predict CVD risks. XGBoost proved the most effective, achieving over 95% accuracy for reliable early detection.

2.1 System Architecture

The architecture of the proposed system includes sensors for real-time data acquisition, cloud-based data storage, an LSTM Auto-encoder model for anomaly detection, and a web interface for user interaction. This structure allows for continuous monitoring and early fault detection, with data flowing seamlessly from the battery sensors to the user interface shown in Fig-1.

2.2 Data Preprocessing

This is one of the most crucial tasks in the process of analytics. Often it is observed that more than half of the total time of analytics process is taken by pre-processing phase. It is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc.
Pre-processing of data is required for improving the accuracy of the model.

2.3 Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems**.**

2.4 XG BOOST

XGBoost is a fast, efficient Gradient Boosted Decision Tree algorithm with built-in L1/L2 regularization to prevent overfitting and support for parallel processing. It handles missing values by optimizing splits and uses cross-validation to find optimal boosting iterations. Effective tree pruning removes unproductive splits, ensuring high model performance for various tasks like regression and classification.

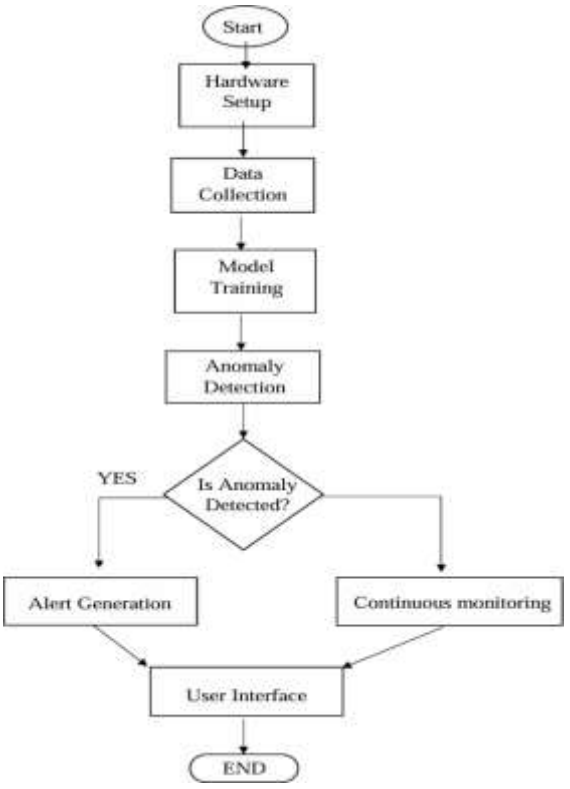

Fig -1- Flowchart

3. CONCLUSIONS

Cardiovascular disease (CVD) is a major global cause of death, underscoring the need for early detection. Machine learning models, including XG Boost, Random Forest, and KNN, were developed using patient data such as chest pain and blood pressure, achieving 98% accuracy, with XG Boost reaching 96%. These models enhance diagnostic speed, accuracy, and outcomes by uncovering complex patterns. Early intervention is vital, as 44% of the dataset's individuals were identified with heart disease.

Suggestions for Future Work

1. Early Detection of Heart Disease Using Student Data - By analysing student health records and activity levels, machine learning algorithms can identify patterns and risk factors, enabling early intervention for heart disease.

2. Integration of Wearable Technology for Comprehensive Health Insights - Data from wearables and health monitoring devices can be combined with educational metrics, offering a holistic view of students' health and wellbeing.

3. Educational Institutions as Drivers of Better Health Outcomes - By leveraging advanced analytics, schools can promote early health interventions, reducing healthcare costs and enhancing public health on a larger scale.

REFERENCES

1. Bo Jin , Chao Che,Zhen Liu ,Shulong Zhang ,Xiaomeng Yin And Xiaopeng Wei "Predicting the Risk of Heart Failure with EHR Sequential Data Modelling". IEEE Access 2018.

2. Aakash Chauhan, Aditya Jain, Purushottam Sharma, Vikas Deep, "Heart Disease Prediction using Evolutionary Rule Learning", "International Conference on "Computational Intelligence and Communication Technology" (CICT 2018).

3. Ashir Javeed, Shijie Zhou, Liao Yongjian, Iqbal Qasim, Adeeb Noor. "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection". IEEE Access (Volume: 7) 2019