iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.61589

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

A Review on Heart Disease Prediction System

Bodla Swathi[1], Vemulapalli Krishna Teja[2], Putta Akhil Kumar[3], Koppol Nikhethan Goud[4], Jatavallabula G K Somayajulu[5]

[1]*Assistant Professor, Department of CSE-DS, Malla Reddy College of Engineering,*
[2, 3, 4, 5] *UG students, Department of CSE-DS, Malla Reddy College of Engineering Hyderabad, TS, India*

Abstract: ***Nowadays, heart failure symptoms can manifest at any stage of life, with older individuals more commonly affected than younger ones. Cardiovascular disease remains a major global health problem requiring accurate predictive tools for early intervention and prevention. This study presents an integrated approach to heart disease prediction using genetics, artificial neural network (ANN) and TPOT classifiers, leading to the development of user relationships to predict new conditions. Using ANN, a powerful learning technique inspired by the human brain, the system can learn complex patterns in data to improve the accuracy of predictions. Additionally, the TPOT classifier can modify model selection and hyperparameter tuning methods to improve prediction. Integrating these classifications into the user interface enables better interactions, allowing physicians and individuals to instantly access relevant information and receive predictions. We've designed an easy-to-use interface that helps catch heart disease early and manage it proactively. This tool is all about making healthcare better and lives healthier.***
Keywords: Cardiovascular disease, Genetics, accuracy, model selection, hyperparameter tuning, prediction, healthcare.

I. INTRODUCTION

Genetic algorithms play a significant role in enhancing the efficiency and accuracy of early medical diagnosis of heart disease. By leveraging genetic algorithms, medical professionals can analyze vast amounts of patient data, identifying patterns and trends that may indicate a predisposition to heart disease. These algorithms excel at identifying complex relationships within data sets, including genetic markers that may contribute to cardiovascular risk. Moreover, genetic algorithms can aid in optimizing diagnostic models by selecting the most relevant features from extensive datasets. This feature selection process helps streamline the diagnostic process, ensuring that healthcare providers focus on the most critical indicators of heart disease. Factors such as obesity, hypertension, high blood cholesterol, and pre-existing heart conditions are among the habitual and physiological risk factors associated with heart disease. Genetic algorithms can assist in analyzing how these factors interact and contribute to an individual's overall cardiovascular health, leading to more personalized and effective preventive measures and treatment plans. Artificial Neural Networks (ANN) are like digital brains that learn from data to spot patterns and make predictions, just like humans do. They're incredibly versatile and have found their way into many areas, including healthcare, where they're invaluable for sorting through complex information and finding hidden insights. Another influential asset in medical diagnosis is the Tree-based Pipeline Optimization Tool (TPOT) classifier. TPOT streamlines the creation and enhancement of machine learning pipelines, encompassing tasks such as data preprocessing, feature selection, and model selection. By exploring an extensive array of algorithms and parameters, TPOT identifies the most suitable combination for a specific dataset, thereby conserving time and effort for researchers and healthcare practitioners.

II. OBJECTIVE

Heart disease affects millions of people around the world and remains a leading cause of death worldwide. To reduce costs and increase the accuracy of diagnostic tests, efficient and reliable medical diagnosis using computer technology is needed. Data mining is a powerful software technique that helps computers generate and classify various properties, making it a valuable tool in medical research. This project uses a genetic algorithm-enhanced classification method to predict heart disease. Optimizing classification models using genetic algorithms inspired by natural selection processes allows for more accurate and efficient predictions. This project covers a variety of related topics including machine learning and its methods, with brief explanations. Data preprocessing techniques are discussed to ensure that the data sets used for analysis are clean and suitable for modeling.

III. LITERATURE SURVEY

Jamin Patel explores the urgent need for improved methods for heart disease detection in his 2015 study, “Improving Heart Disease Detection Using Machine Learning and Data Mining Techniques.”

Using data mining techniques, Patel aims to help medical professionals diagnose heart disease more effectively. In this study, we compare the performance of various algorithms, including J48, Logistic Model Tree, and Random Forest, in the context of heart disease diagnosis. Patel evaluates these algorithms using the Cleveland database from the UCI repository, which contains 303 instances and 76 attributes, to determine the most efficient approach. Ultimately, the goal of this research is to uncover hidden patterns in heart disease data and develop predictive models to identify at-risk patients, potentially reducing heart disease-related mortality.[1]

In a 2017 paper, "Can machine learning improve cardiovascular risk prediction using routine clinical data" Stephen F. Weng explores the potential of machine learning to improve cardiovascular risk prediction. This highlights the limitations of current approaches, which often fail to identify people who could benefit from preventive treatment or lead to unnecessary interventions. Weng aims to use machine learning techniques to increase the accuracy of predictions by considering complex interactions between various risk factors. In the evaluation, Weng will use routine clinical data to address these challenges and explore the effectiveness of machine learning in improving cardiovascular risk prediction.[2]

V. V. Ramalingam's 2018 paper "Predicting Cardiovascular Disease Using Machine Learning Techniques" addresses the urgent need for reliable and accurate diagnostic systems for cardiovascular disease (CVD), which has become a leading cause of death worldwide. Ramalingam emphasizes the importance of applying machine learning algorithms to medical datasets to automate complex data analysis and help medical professionals diagnose cardiovascular disease. He analyzes the performance by examining various models based on machine learning algorithms and techniques. In particular, supervised learning algorithms such as support vector machines (SVM), K-Nearest Neighbor (KNN), Naive Bayes, decision trees (DT), random forests (RF), and ensemble models have become popular choices among cardiac diagnostic researchers.[3]

In 2020 a research paper titled, "Exploring Genetic Algorithms in Complex Optimization Problems" by Jonathan A. Smith and Emma K. Johnson, the authors discuss the applications and advancements of genetic algorithms (GA) in various fields. Provides in-depth research on carefully trace the evolution of GA, explaining its basic principles and detailing its effectiveness in solving complex optimization problems. In this study, we carefully investigate the main components of GA, including population initialization, fitness function development, crossover and mutation operators, and selection strategy. The paper also highlights recent advances in genetic algorithm techniques, especially hybrid approaches that integrate GA with other optimization techniques such as simulated annealing and particle swarm optimization. Through insightful case studies covering engineering design, financial modeling, and data analysis, the authors demonstrate successful applications of GA in real-world scenarios and highlight the versatility and effectiveness of GA in finding optimal solutions. Overall, this study provides a comprehensive understanding of the key role played by genetic algorithms in solving complex optimization problems in various domains.[4]

The study concluded with best algorithm and optimization techniques have trained and tested to give good results in heart disease prediction. Experimental results confirmed the achieving disease prediction with an impressive accuracy of 99.02%.

IV. EXISTING SYSTEM

1) *Support Vector Machine (SVM):* It is a supervised learning algorithm used for classification tasks. In heart disease prediction, SVM aims to find the optimal hyperplane that separates the data points into different classes, such as "heart disease" and "no heart disease." It works by maximizing the margin between classes, thus enhancing generalization to unseen data. SVMs in heart disease prediction utilize features like age, cholesterol levels, blood pressure, etc., to classify patients into risk categories, aiding in diagnosis and treatment planning. Model performance is optimized through appropriate kernel selection and parameter tuning.
2) *K-Nearest Neighbours:* In this method, K- Nearest Neighbours showed poor performance because KNN classifies test data directly from the dataset, no training was performed before testing.
3) Decision Tree (ID3): At training stage, it converted the continuous value's data into categorical values and given arrange. When test data pattern contained values out of this given range, the classifier performance was affected and thus predicted wrong class label.
4) *Gaussian Naive Bayes:* At the training stage, it calculated the mean and standard deviation of each attribute. This mean and standard deviation were used to calculate the probabilities for the test data. For this reason, some attributes values are too big or too small from the mean. When testing data pattern contains those attributes values, it affects the classifier performance and sometimes gives wrong output label.

5) *Logistic Regression:* At the training stage, Logistic Regression algorithm estimated coefficient values by using stochastic gradient descent. The model can be trained for a fixed or as much as No of epochs by using stochastic gradient descent. Coefficients values are updated until the model predicts the correct class label for each training data.
6) *Random Forest:* Random Forest is an ensemble classification method which is based on Decision Tree algorithm. This algorithm takes a portion of the dataset and then builds a tree, repeating this step for creating a forest by combining the generated trees. At the test stage, each tree predicts a class label for each test data and the majority values of the class label are assigned to the test data. Therefore, it showed reasonable performance than conventional decision tree algorithm for this data.

V. LIMITATIONS OF EXISTING SYSTEM

1) *Support Vector Machines (SVMs):* SVMs can be sensitive to kernel parameter selection and may not work efficiently on large or high-dimensional data sets. It also requires extensive data preprocessing.
2) *Decision tree:* Decision trees are prone to overfitting, especially if the tree depth is not properly constructed. Additionally, even small changes to the data can make it unstable, causing different trees and inconsistent predictions.
3) *Random forests:* Random forests generally have high prediction accuracy and handle nonlinear relationships well, but are computationally expensive and not easily interpretable, especially when the ensemble consists of many trees.
4) *K-Nearest Neighbours (KNN):* KNN suffers from the "curse of dimensionality" and becomes less efficient as the number of features increases. Also, the entire training dataset must be stored and retrieved for each prediction, making it computationally expensive during inference.
5) *Logistic Regression:* Logistic regression assumes a linear relationship between the characteristic and the log probability of the target variable. This may not always be true in complex data sets.
6) *Navi Bayes:* This may not be true for complex relationships between cardiovascular risk factors and can potentially lead to simplistic or inaccurate predictions.

VI. PROPOSED SYSTEM

Advanced classification methods combined with automated tools such as genetic and plays an important role in uncovering hidden relationships between correlated features. This approach significantly improves the accuracy of class label prediction, including identifying patients with cardiovascular disease in large datasets. Using genetic algorithm, Artificial neural network these methods can effectively determine optimal feature combinations and model parameters, resulting in more accurate and reliable predictions known for its automated machine learning capabilities, Tpot Classifier explores a variety of classification models and hyperparameter configurations to further simplify the process, reducing diagnostic time and costs. The integration of genetic algorithms and Tpot classifier transforms the diagnostic process into an expert system. The system can accurately distinguish between patients with and without cardiovascular disease, mimicking the experience of a healthcare worker while increasing accuracy, efficiency and cost-effectiveness.

VII. ARCHITECTURE

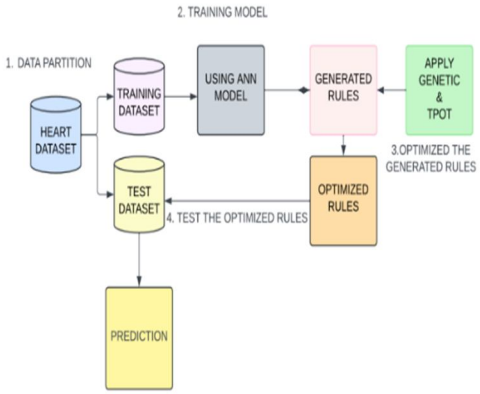


fig. block diagram

This architecture is designed to create, optimize, and evaluate outcome prediction models using cardiac datasets. The architecture follows a systematic approach to thoroughly evaluate and optimize model performance.

1) *Data partition:* The cardiac data set is the starting point for the process. This dataset is divided into two parts as training dataset and testing dataset. Data partitioning is important for training and evaluating model performance.
2) *Training model:* The training dataset is used to train an artificial neural network (ANN) model. The rules generated in this step are applied to the ANN model. These rules can be initial weights, biases or other hyperparameters that affect the model's training process.
3) *Optimization of generated rules:* After training the ANN model, the generated rules are optimized using genetic algorithms and the Tree-Based Pipeline Optimizer (TPOT). The goal of this step is to fine-tune the rules to improve the performance of the model. We then use the test dataset to test the optimized rules.
4) *Test optimized rules:* Finally, the optimized rules are applied to the ANN model to make predictions on the test dataset. This step evaluates the performance of the model with optimized rules compared to the original model with generated rules.

VIII. RESULTS

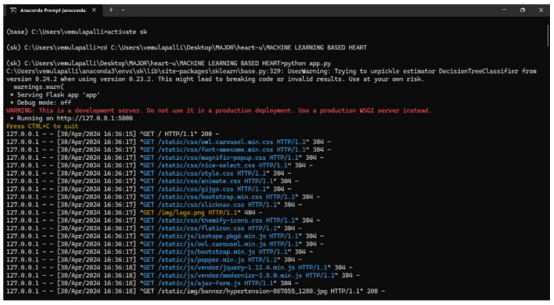


fig.1 generation of user interface using flask framework

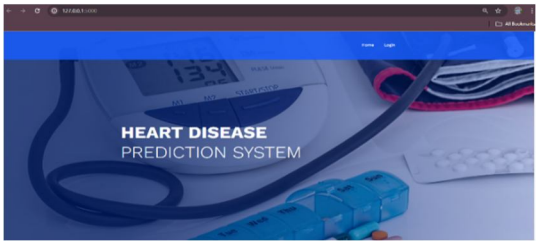


fig.2 homepage of user interface

Fig.3 training and testing

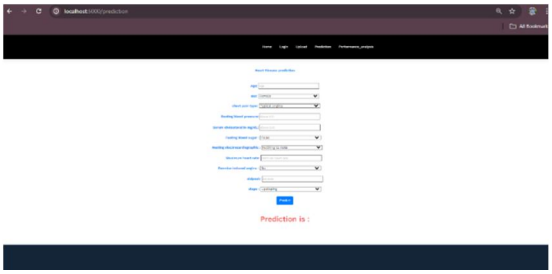


fig.4 prediction page

IX. CONCLUSION

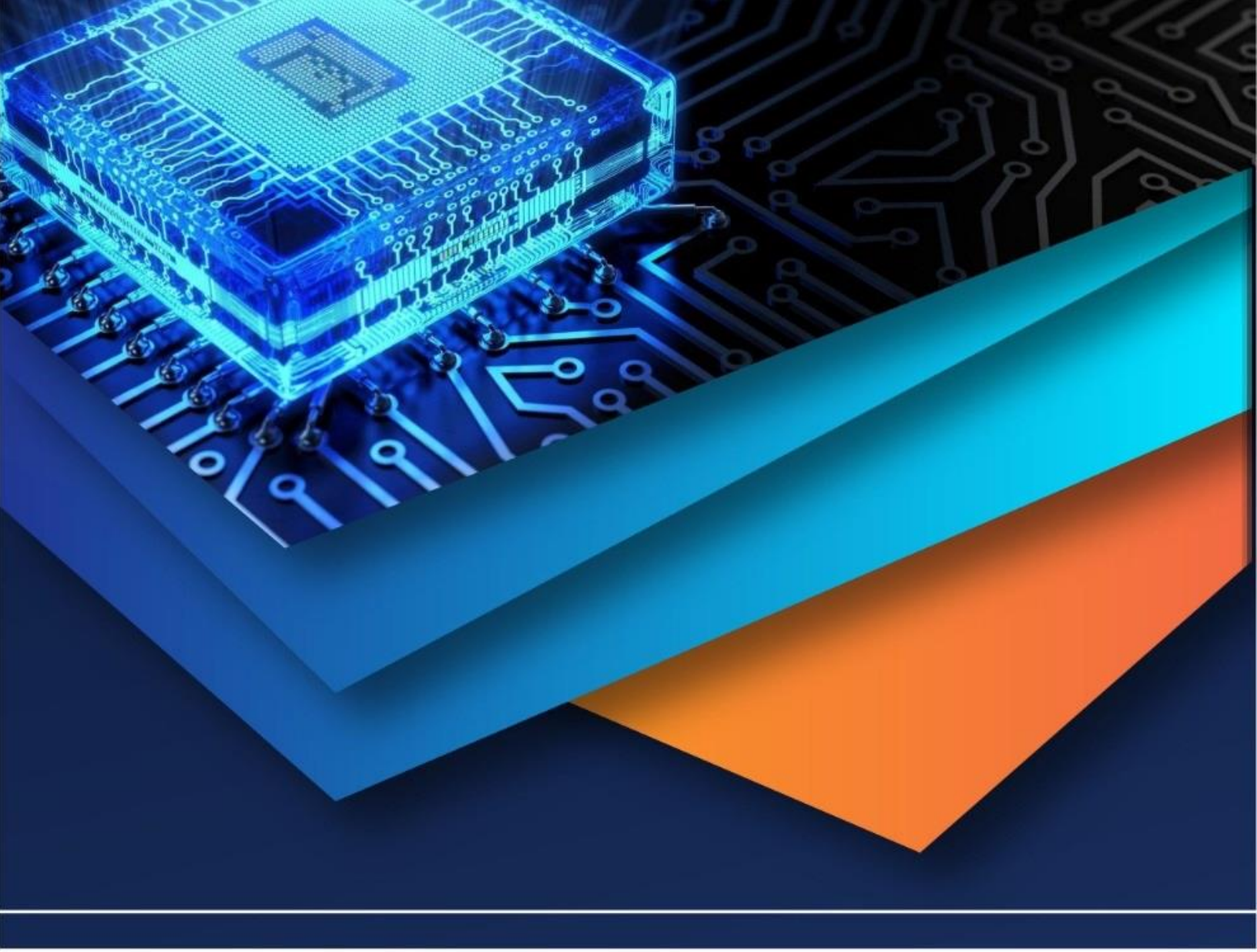
In this study, cardiovascular disease prediction project uses machine learning algorithms, classification methods, and real-world datasets to create robust and reliable prediction models. By analysing key factors such as blood pressure, cholesterol levels, and other medical indicators, the model can accurately predict an individual's likelihood of developing cardiovascular disease. Through data preprocessing, feature selection, and model optimization, the project ensures the accuracy and efficiency of predictions. Additionally, integrating advanced technologies such as genetic algorithms and automatic classifiers such as TPOT improves the performance and scalability of the model.

X. FUTURE SCOPE

One viable future improvement for predictive cardiovascular disease research is the integration of real-time data streaming and continuous monitoring capabilities. These enhancements allow models to receive and process data in real time, enabling immediate prediction and intervention. This integration may involve the use of medical monitoring devices, such as IoT devices or wearable sensors, that continuously collect relevant health data, including vital signs such as heart rate, blood pressure, and activity levels. Machine learning models analyse this streaming data to provide continuous predictions and warnings about potential cardiovascular disease risks. To further enhance the model's capabilities, we incorporate a continuous learning mechanism, allowing it to adapt and evolve in real time as new data streams become available. These improvements not only improve the predictive accuracy of the model, but also enable proactive medical intervention, improving patient outcomes and quality of care.

REFERENCES

[1] Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. pp.vol. 7(1):129–37, 2015.
[2] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine learning improve cardiovascular risk prediction using routine clinical data, PLOS, pp. vol. 12(4), 2017.
[3] Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol, pp.vol. 7(2.8):684–687, 2018
[4] Jonathan A. Smith and Emma K. Johnson,"Exploring Genetic Algorithms in Complex Optimization Problems" a survey in 2020.
[5] Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl. 2017; 9:1–16.
[6] Pahwa K, Kumar R. Prediction of heart disease using hybrid technique for selecting features. In: 2017 4th IEEE Uttar Pradesh section international conference on electrical, computer and electronics (UPCON). IEEE. p. 500–504.
[7] Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In: 2017 IEEE symposium on computers and communications (ISCC). IEEE. p. 204–207.
[8] Chauhan R, Bajaj P, Choudhary K, Gigras Y. Framework to predict health diseases using attribute selection mechanism. In: 2015 2nd international conference on computing for sustainable global development (INDIACom). IEEE. p. 1880–84.
[9] Bouali H, Akaichi J. Comparative study of different classification techniques: heart disease use case. In: 2014 13th international conference on machine learning and applications. IEEE. p. 482–86.
[10] Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol. 2011; 3:67.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY