iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** I **Month of publication:** January 2024

DOI: https://doi.org/10.22214/ijraset.2024.58148

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 I Jan 2024- Available at www.ijraset.com

Machine Learning Oriented Heart Disease Prediction using Clinical Data with Genetic Factors: A Overview

Sharayu Ukirde[1], Prof. Anup Date[2], Mayur Jambe[3], Prof. Hemant Shiyal[4]
[1, 2, 3, 4]*DY Patil University Ambi, Pune*

Abstract: ***The critical global health concern of heart disease is to focus on improving its prediction accuracy through the integration of clinical data and genetic factors using machine learning. The primary aim is to develop a highly accurate predictive model that can revolutionize heart disease diagnosis and enhance healthcare outcomes. Machine learning, a branch of artificial intelligence, offers the potential to create systems that can learn from historical data and make informed predictions, particularly in the medical field. The research measures the likelihood of patients being diagnosed with cardiovascular heart disease based on a range of medical attributes, including age, gender, chest pain, and fasting sugar levels. A dataset from Kaggle, comprising comprehensive patient medical histories and attributes, is leveraged, utilizing 14 medical attributes to predict heart disease risk. The research incorporates four unique machine learning algorithms, namely Logistic regression, Support Vector Machines (SVM), Decision Trees, and Random Forest Classifier, to classify individuals into two groups: those who are susceptible to heart disease and those who are not. The research endeavours to significantly enhance heart disease prediction accuracy, providing a promising avenue for personalized healthcare and early detection. Through the synergy of clinical and genetic data and the capabilities of machine learning, it aims to contribute to more precise risk assessment, timely treatment, and ultimately better outcomes for individuals at risk of heart disease.***

Keywords: Heart disease prediction, Clinical data, Genetic factors, Machine learning, Kaggle dataset, Logistic regression, Support Vector Machines (SVM), Decision Trees, Random Forest Classifier.

I. INTRODUCTION

Heart disease is a critical global health concern, and this research focuses on significantly enhancing its prediction accuracy by integrating clinical data with genetic factors through the power of machine learning. Our aim is to develop a highly accurate predictive model that can transform heart disease diagnosis and improve healthcare outcomes. Machine learning, a subdivision of artificial intelligence, provides the capability to develop systems that can acquire knowledge from data and make timely decisions [2].It completes this by training algorithms on historical data to build models capable of predicting heart disease based on new input data [3]. These models excel at detecting intricate patterns within datasets, making them valuable tools for accurate predictions [5].

Medical databases predominantly consist of discrete information, which can complicate decision-making processes [4]. Machine learning, a specialized branch of data mining, adeptly handles extensive and organized datasets [6].In the medical field, it finds applications in diagnosis, detection, and disease prediction[8]. Our primary objective is to provide doctors with a tool for the early-stage detection of heart disease, facilitating effective treatments and preventing severe consequences[1]. Machine learning plays a pivotal role in uncovering hidden discrete patterns within the data, ultimately contributing to heart disease prediction and early diagnosis [7].

In this study, we assess the probability of patients being diagnosed with cardiovascular heart disease by analyzing their medical characteristics, including gender, age, chest pain, fasting sugar levels, and other factors [9].Leveraging a dataset sourced from Kaggle, about complete patient medical histories and attributes [10], we employ 14 medical attributes to predict the likelihood of heart disease. Four different machine learning algorithms, namely Logistic regression, Support Vector Machines (SVM), Decision Trees, and Random Forest Classifier [11], are utilized to train these attributes. The result of this training process is the classification of patients into two categories: those who are at risk of heart disease and those who are not [12]. This research endeavours towards significantly improve heart disease prediction accuracy, offering a promising avenue for personalized healthcare and early detection [13]. Through the synergy of clinical and genetic data and the capabilities of machine learning, we aim to contribute to more precise risk assessment, timely treatment, and ultimately, better outcomes for individuals at risk of heart disease [14].

II. LITERATURE REVIEW

This paper highlights the importance of early detection of heart disease through the utilization of machine learning. It emphasizes the significance of predictive models in healthcare and their potential to enhance patient outcomes [1]. The research specifically focuses on the application of machine learning and data mining to predict diseases within the healthcare field. It underscores the role of machine learning in healthcare and emphasizes the use of predictive models for improved disease prediction. A survey was conducted to predict heart disease using various classification algorithms, including Naive Bayes, KNN (K-Nearest Neighbour), Decision tree, and Neural network. The accuracy of these classifiers was projected for different numbers of attributes [21]. Furthermore, heart disease prediction was carried out using Naive Bayes classification and SVM (Support Vector Machine). The analysis utilized performance measures such as Mean Absolute Error, Sum of Squared Error, and Root Mean Squared Error. The results demonstrated that SVM outperformed Naive Bayes in terms of accuracy [22].

A. Support Vector Machine

Support vector machines exist in different forms, linear and non-classifier. What is usual in this context, two different datasets are involved with SVM training, and a test set. Propose a system containing two models based on linear Support Vector Machine (SVM). The first one is called L1 regularized and the second one is called L2 regularized. The first model is used for removing unnecessary features by making the coefficient of those features zero. The second model is used for prediction. Prediction of disease is done in this part. A hybrid grid search algorithm was suggested to enhance the performance of both models. This algorithm takes into account various metrics such as accuracy, sensitivity, specificity, the Matthews correlation coefficient, ROC chart, and area under the curve to improve the models. The Cleveland dataset was utilized for this purpose, with a data split of 70% for training and 30% for testing using holdout validation. There are two experiments carried out and each experiment is carried out for various values of C1, C2, and k where C1 is the hyperparameter of the L1 regularized model, C2 is the hyperparameter of the L2 regularized model k is the size of a selected subset of features. The initial trial involves combining an L1-linear SVM model with an L2-linear SVM model, resulting in a maximum testing accuracy of 91.11% and a training accuracy of 84.05%. In the second experiment, we cascade an L1-L1-linear SVM model with an L2-linear SVM model using an RBF kernel, achieving a maximum testing accuracy of 92.22% and a training accuracy of 85.02%.They have obtained an improvement in accuracy over predictable SVM models by 3.3% [23].

B. Random Forest

Random Forest is an influential ensemble learning method that merges numerous decision trees. By randomly sampling the training dataset with replacement (bootstrapping) and selecting random subsets of features, it creates a collection of decision trees. In the field of heart disease prediction, Tikotikar and Kodabagi [4] have successfully utilized the Random Forest method, achieving an impressive accuracy of 86.90% [5]. Their study directly aligns with the primary objective of your research, making the utilization of the Random Forest algorithm highly relevant to your work. Furthermore, Pal and Parija [5] have also made significant contributions to this field through their research on heart disease prediction using Random Forests. Additionally, the study conducted by Riyaz and colleagues [7] highlights the importance of enhancing coronary heart disease prediction, which is a crucial aspect of your own study. Their work also explores methods for outlier elimination, which can greatly improve the accuracy of predictive models.

C. Naive Bayes

The Naive Bayesian (NB) classifier, also known as the "independent feature model," is based on the Bayesian theorem and serves as a straightforward probabilistic classifier with a strong assumption of independence. In general, the NB classifier assumes that the presence or absence of a specific class feature is independent of the presence of other class features. NB classifiers are commonly used in supervised learning. Furthermore, the proposed system incorporates Naive Bayesian techniques to classify datasets and employs the Advanced Encryption Standard (AES) algorithm to ensure secure data transfer in disease prediction [20]. This research study is directly relevant to your area of interest as it specifically focuses on utilizing machine learning to predict heart disease. It provides valuable insights into the methodologies employed for this objective. The study conducted by Rajdhan et al. [8], titled "Heart Disease Prediction Using Machine Learning," emphasizes the significance of this research. Additionally, Hasan and his team have conducted a comparative analysis of various classification approaches for heart disease prediction. This research is relevant to your work as it provides insights into model selection and performance evaluation. The publication entitled "Comparative Analysis of Classification Approaches for Heart Disease Prediction" authored by Hasan, Mamun, Uddin, and Hossain [13] provides significant insights within this particular domain.

D. Logistic Regression

LR models are used to estimate the possibilities of the target belonging to a specific category. In the logistic function equation, the variable x represents the input.

We have the flexibility to input values within the range of -20 to 20 into the logistic function. Subsequently, these inputs undergo a transformation, resulting in a range between 0 and 1.

The fundamental model of this function finds its primary application in binary classification problems, where it categorizes each sample into one of two groups: Yes or No.

This research conducted by Rahim and colleagues introduces an integrated machine-learning framework that aims to predict cardiovascular diseases, aligning with the objectives of your own research. This framework offers significant insights into the creation of efficient predictive models for heart disease. The research paper titled "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases" by Rahim, Rasheed, Azam, Anwar, Rahim, and Muzaffar [15] highlights the significance of this framework.

Furthermore, Fahd Saleh Alotaibi has developed a machine-learning model that evaluates the performance of five distinct algorithms: Decision Tree, Logistic Regression, Random Forest, Naive Bayes, and SVM classification algorithms. The research compares the accuracy of these algorithms.

E. Decision Tree

Decision trees offer a straightforward approach in situations where the connection between features and outcomes is non-linear or when features exhibit interaction with one another. By repeatedly splitting the data based on a measure like information gain, decision trees can effectively capture the information contained in the data (Kumar et al., 2013). In terms of accuracy, the Decision Tree algorithm has been shown to perform exceptionally well, as demonstrated by a study that reported the highest accuracy [19].

Version 1:

The research paper titled "Evaluation Measures for Models' Assessment over Imbalanced Data Sets" authored by Bekkar, Djemaa, and Alitouche [14], provides a specific focus on evaluation techniques for assessing models, particularly in the context of imbalanced data sets. These evaluation measures are of utmost importance in guaranteeing the accuracy and effectiveness of predictive models.

F. K-Nearest Neighbor

This is a method of machine learning that is supervised in a sluggish manner, allowing for prediction and classification. It is easy to concept and understand, requires minimal training time, and utilizes the entire training set.

This nonparametric approach involves measuring the distance between two sets of data in order to predict and label unknown data based on known data. The distance metric is employed to compute the distance between every point in the testing data and each point in the training data (Reddy et al., 2019).K. Pramanik et al. propose a Hybrid Algorithm that combines the ID3 and KNN algorithms for predicting heart disease.

The data is pre-processed using the KNN algorithm, making it a pre-processed algorithm. The pre-processed data forms the training set, which is then classified using a tree structure.

The ID3 algorithm is implemented as the classifier for predicting heart disease. The KNN Algorithm is used to classify incorrect values [25]. To assess the accuracy of the computer-based prediction algorithm, SVM, KNN, and artificial neural network (ANN) are employed. KNN (82.963%) and ANN (73.3333%) are among the algorithms employed.

The authors suggest SVM as the optimal classification algorithm, exhibiting the highest accuracy in predicting heart disease. In comparison to other prediction models, the KNN model attains the highest accuracy (88.52%). Additionally, Rajdhan et al. [8] utilized machine learning algorithms to predict heart disease.

III. PROPOSED SYSTEM

The methodology for developing a heart disease prediction model that integrates clinical data with genetic factors involves a systematic approach to data collection, preprocessing, model development, evaluation, and deployment.

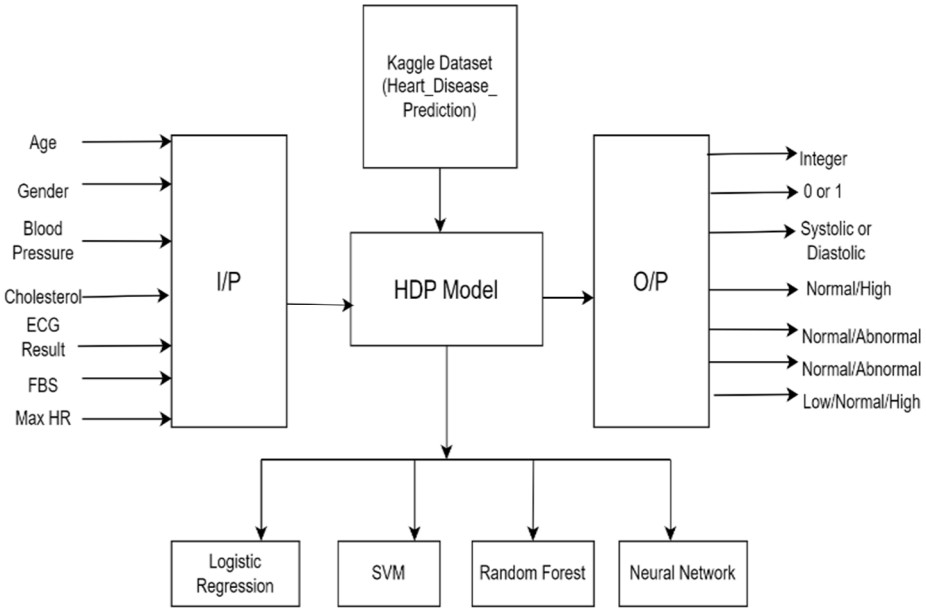


Fig. 1 - Block Diagram of Heart Disease Prediction

1) *Data Collection:* The first step involves gathering a diverse and comprehensive dataset that combines various clinical data attributes (such as age, gender, blood pressure, cholesterol, FBS, ECG results, and heart rate) with genetic factors. The dataset should be sufficiently large and representative to capture a wide range of patient profiles.
2) *Data Analysis and Preprocessing:* Once the data is collected, a thorough analysis is conducted to understand its structure and quality. This includes examining data distributions, identifying potential outliers, and dealing with missing values. Data preprocessing methods are applied to confirm that the dataset is fresh and suitable for analysis.
3) *Feature Selection:* Enhancing the accuracy of heart disease prediction models involves a precarious process known as feature selection. It involves identifying the most appropriate attributes that have a significant impact on predicting heart disease. Feature selection can be based on statistical analysis, domain knowledge, or machine learning techniques.
4) *Data Integration:* Integrating clinical data with genetic factors is a core part of this research. The selected features from both clinical and genetic data sources are combined to create a unified dataset for heart disease prediction. This integration allows the model to consider a broader range of factors that contribute to heart disease risk.
5) *Machine Learning Model Development:* Machine learning models are employed to develop accurate predictive models for heart disease. This research specifically selects the following algorithms: logistic regression, support vector machines (SVM), decision trees, and random forest classifier.
6) *Model Training and Evaluation:* Once the models are developed, they are accomplished on the combined dataset. Training involves adjusting model parameters to fit the data and abate forecast errors. After training, the models are rigorously evaluated using various performance metrics, such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve. This assessment guarantees the efficacy of the models in accurately predicting heart disease.
7) *Challenges Addressed:* The research acknowledges and addresses several challenges related to data preprocessing, feature selection, and model development. Data preprocessing tasks may include handling missing values and outliers, while feature selection objectives to identify the most useful attributes. Model development challenges involve choosing the right algorithms and optimizing their parameters to achieve the best predictive performance.

This comprehensive methodology follows a systematic approach to integrate clinical and genetic data, develop accurate predictive models, and address the complexities of data analysis and model development. The research aims to enhance early heart disease detection, facilitate personalized healthcare, and improve healthcare outcomes for individuals at risk of heart disease.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue I Jan 2024- Available at www.ijraset.com

IV. METHODS

A. Random Forest Algorithm

The Random Forest algorithm, as stated in the research paper "Prediction of Heart Diseases Using Random Forest" by Pal and Parija[5], is a collaborative learning technique. It operates by creating numerous decision trees and merging their predictions to enhance the accuracy and reliability of the overall predictions. Random Forest is particularly useful for classification tasks, making it suitable for predicting heart diseases. It can handle a large number of features and is less prone to overfitting.

This ensemble classifier builds several decision trees and incorporates them to get the best result. For tree learning, it mainly applies bootstrap collecting or bagging.

Given a dataset, X = {x1, x2, x3, ..., xn}, with corresponding responses Y = {x1, x2, x3, ..., xn}, the bagging process is repeated from b = 1 to B.

B. Support Vector Machines (SVM) Algorithm

In the scholarly paper titled "Utilizing Machine Learning for Heart Disease Prediction" authored by Rajdhan et al. [7], the authors emphasize the importance of Support Vector Machine (SVM) as a resilient algorithm for supervised machine learning. SVM is widely employed in various classification tasks, particularly in the prediction of heart disease. By identifying a hyperplane that optimally distinguishes different classes, SVM maximizes the margin between them. This characteristic makes SVM highly effective in both linear and non-linear classification tasks, showcasing its versatility in the field of heart disease prediction.

C. Logistic Regression Algorithm

Logistic regression is a statistical model utilized in binary classification, as highlighted in the research paper "Disease Prediction by Machine Learning Over Big Data from Healthcare Communities" authored by Chen et al.[2]. This model estimates the probability of a binary outcome, such as the existence or nonexistence of heart disease. Logistic regression is a straightforward yet powerful algorithm utilized for classification purposes, especially when interpretability holds significance. The study "Improving Coronary Heart Disease Prediction" conducted by Riyaz, Butt, and Zaman [6] recognizes the exceptional predictive performance of this model.

$$f(x) = \frac{1}{1+e^{-x.}}$$

D. Naive Bayes Classification Algorithm

Naive Bayes classification, as discussed in "A Survey on Techniques for Prediction of Disease in Medical Data" by Tikotikar and Kodabagi.[4], is based on Bayes' theorem. It's a probabilistic classification algorithm that assumes that features are conditionally independent, hence the term "naive." Naive Bayesian classification is widely used in healthcare for disease prediction due to its simplicity and effectiveness. It's also declared in "Evaluation Measures for Models' Calculation over Excessive Data Sets" by Bekkar, Djemaa, and Alitouche.[15] in the context of model valuation. Naive Bayes is a simple yet capable categorization algorithm built on Bayes Theorem. It presupposes predictor independence, which means that the traits or characteristics are unrelated or connected in any way.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

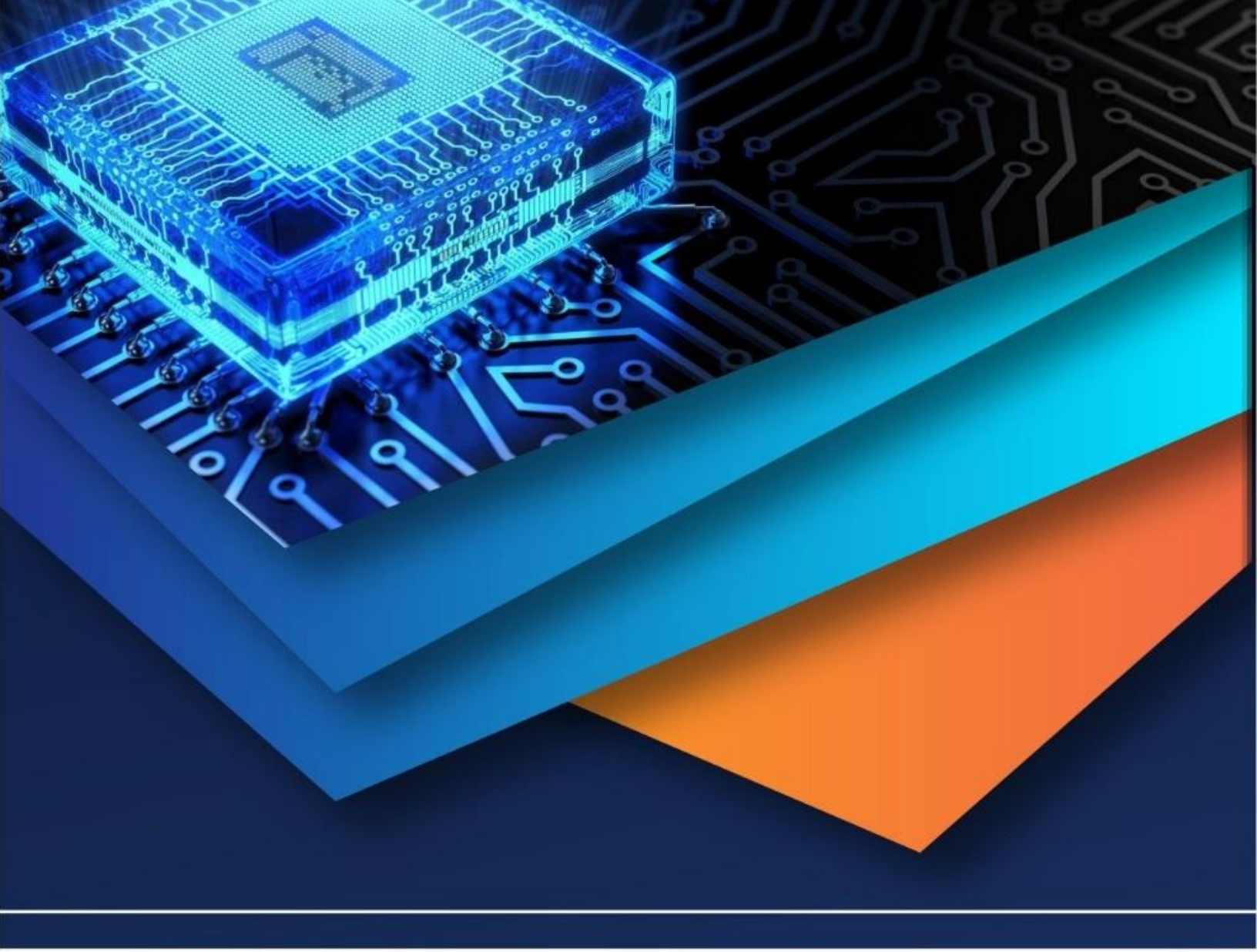
V. CONCLUSION

By integrating crucial health data such as age, gender, blood pressure, cholesterol, etc., along with genetic factors, we employ this system to develop a predictive model for Heart Disease.

Our approach involves a series of stages, including data collection, analysis, and the utilization of various machine-learning techniques. We test challenges like cleaning up the data, choosing the right information, and building a smart model. The big idea is to use both regular health info and genetic details to make our heart disease predictions more accurate. This will help to improve how accurately we can predict diseases and create useful tools for predicting them. Teamwork between technology and medical fields is key to making healthcare better.

REFERENCES

[1] Yilmaz R, Yagin FH. "Early detection of coronary heart disease based on machine learning methods." International Medical Journal. 2022 Jan 1; 4(1): 1–6.
[2] Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L., 2017. Disease prediction by machine learning over big data from healthcare communities. IEEE Access, 5, pp.8869-8879. 18. Huang, F., Wang, S. and Chan, C.C., 2012, August. Predicting disease by using data mining based on healthcare information system. In 2012 IEEE International Conference on granular computing (pp. 191-194).
[3] Boukhatem C, Youssef HY, Nassif AB. Heart Disease Prediction Using Machine Learning. IEEE Advances in Science and Engineering Technology International Conferences (ASET). 2022 Feb 21–24, Dubai, United Arab Emirates.
[4] Tikotikar, A., & Kodabagi, M., 2017. A survey on technique for prediction of disease in medical data. In 2017 International Conference on Smart Technologies for Smart Nation (Smart Tech Con) (pp. 550-555). IEEE.
[5] Pal M, Parija S. Prediction of Heart Diseases Using Random Forest. Journal of Physics: Conference Series. 2021 Mar 15; 1817(1): 1–9. doi: 10.1088/1742-6596/1817/1/012009.
[6] Banu, M.N. and Gomathy, B., 2014, March. Disease forecasting system using data mining methods. In 2014 International Conference on Intelligent Computing Applications (pp. 130-133). IEEE.
[7] Riyaz L, Butt MA, Zaman M. Improving Coronary Heart Disease Prediction by Outlier Elimination. Applied Computer Science. 2022 Mar 28; 18(1): 70–88.
[8] Rajdhan A, Sai M, Agarwal A, Ravi D, Ghuli DP. Heart Disease Prediction Using Machine Learning. International Journal of Research and Technology. 2020; 9(4): 659–662.
[9] Pal M, Parija S. Prediction of Heart Diseases Using Random Forest. Journal of Physics: Conference Series. 2021 Mar 15; 1817(1): 1–9.
[10] Kaggle.com. Kaggle Cardiovascular Disease Dataset. [Internet]. 2019 [updated 2019 Jan 20; cited 2022 Sep 01]; Available from: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.
[11] Rahman MM, Rana MR, Alam MNA, Khan MSI, Uddin KMM. A Web-Based Heart Disease Prediction System Using Machine Learning Algorithms. Network Biology. 2022 Jun 1; 12(2): 64–80.
[12] Sing A, Kumar R. Heart disease prediction using machine learning algorithms. IEEE international conference on electrical and electronics engineering (ICE3). 2020 Feb 14–15, pp. 452–457, Gorakhpur, India.
[13] Hasan SMM, Mamun MA, Uddin MP, Hossain MA. Comparative analysis of classification approaches for heart disease prediction. IEEE International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2). 2018 Feb 8–9, Rajshahi, Bangladesh.
[14] Bekker, M., Djemaa, H. K., and Alitouche, T. A. "Evaluation measures for models' assessment over imbalanced data sets." J Inf EngAppl, 3(10).
[15] Rahim, A., Rasheed, Y., Azam, F., Anwar, M. W., Rahim, M. A., and Muzaffar, A. W. "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases." IEEE Access, 9, 106575-106588.
[16] Raju, C., Philipsy, E., Chacko, S., Suresh, L.P. and Rajan, S.D., 2018, March. A Survey on Predicting Heart Disease Using Data Mining Techniques. In 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), pp. 253-255. IEEE.
[17] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259-64). IEEE.
[18] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava, "Effective Heart Disease Prediction using Hybrid Machine Learning Techniques", IEEE Access 2019.
[19] Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.
[20] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "Design And Implementation Heart Disease Prediction Using Naive Bayesian", International Conference on Trends in Electronics and Information(ICOEI 2019).
[21] Theresa Princy R, J. Thomas,' Human Heart Disease Prediction System using Data Mining Techniques', International Conference on Circuit Power and Computing Technologies, Bangalore,2016.
[22] Nagaraj M Lutimath, Chethan C, Basavaraj S Pol.,' Prediction Of Heart Disease using Machine Learning', International Journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.
[23] Ali, Liaqat, et al, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure" IEEE Access 7 (2019): 54007-54014.
[24] Beant Kaur h, Williamjeet Singh, "Review on Heart Disease Prediction System using Data Mining Techniques", © IJRITCC, Vol.2, Issue: 10, p.p.3003-08, 2014.
[25] Anup Date, Dr. P. V. Ingole, "Low Light Video Enhancement: A Survey", International Journal of Computer Applications (0975 – 8887) National Conference on Recent Trends in Computer Science & Engineering, MEDHA 2015.
[26] Anup Date, " A Video Upgradation of Low Vision AVI Video by Individual Pixel Channel Intensity Measurement and Its Enhancement", IJRITCC, Vol-4, Issue-4, PP 486-490.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY