IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

A Comparative Study Of Random Forest And Naive Bayes Algorithms For Heart Disease Prediction

[1]Sake Madhu*, [2]Yerraginnela Shravani, [3] B.Maria Joseph and [4]Gundala Swarnalatha
[1]Professor, [2,3&4]Asst.Professor,
[1]Department of CSE (AI&ML), [2&3] Department of CSE (AI&ML), [4]Department.of AI&DS,
[1, 2, 3&4]Guru Nanak Institutions Technical Campus, Hyderabad, India

Abstract: As per the latest research, the marked rise in the number of individual heart attack cases, we need to put in place a system that will enable us to identify the early warning symptoms of a heart attack and avoid them. It is not feasible for the average individual to regularly undergo costly tests like ECG, so a system that is both portable and reliable for evaluating the possibility of heart disease must be in place. Thus, we propose developing an application that may predict the vulnerability of a cardiac ailment based on fundamental symptoms such as age, sex, type of chest discomfort, serum cholesterol, etc. Two machine learning algorithms, Random Forest and Naïve Bayes, are selected and compared in the proposed system.

Index Terms – Health Monitoring, Heart disease Identification, Machine Learning, Cardiology, Health Care.

I. Introduction

Heart disease remains a leading cause of mortality worldwide, with burden on healthcare systems. Early and accurate prediction of heart disease can enable timely interventions and prevention, improving patient outcomes. In recent years, machine learning algorithms have shown promising results in the prediction of heart disease [1-2]. This research paper aims to review the current state of machine learning applications in heart disease prediction and explore the comparative performance of various Random Forest and Naïve Bayes algorithms.

Several supervised machine learning techniques have been widely explored for the purpose of heart disease prediction. is sample paper format only please use this format and follow this structure as per your requirement

II. Literature Review

A comprehensive review of the existing literature on heart disease prediction using machine learning reveals the growing interest and potential of this field. Machine learning models have been developed to predict the presence of heart failure, estimate the severity, and forecast adverse events such as rehospitalizations and mortality. A framework is defined for the collection of the data using various IoT sensor devices using the fuzzy logic which further helps in integration of the data using machine Learning algorithms [3,4].

One study compared the accuracy of k-nearest neighbor, decision tree, linear regression, and support vector machine algorithms in predicting heart disease using a publicly available dataset. The results indicated that the support vector machine model achieved the highest accuracy [5]. Another review paper highlighted the applications of artificial neural networks, decision trees, random forests, support vector machines, and naive Bayes classifiers for early-stage heart disease prediction [4]. The authors noted that the choice of algorithm depends on the specific dataset and problem at hand.

Advances in machine learning have also enabled the integration of diverse data sources, including clinical records, imaging data, sentimental analysis and genetic profiles, to improve the predictive performance [6-7]. However, challenges remain in developing models that are clinically interpretable, can handle imbalanced datasets, and are generalizable to diverse patient populations [8-9].

III. METHODOLOGY

Heart disease is a common and potentially fatal condition that requires accurate diagnosis. Conventional diagnostic methods in cardiology are time-consuming and prone to the human error and have difficulty identifying minute patterns in huge databases. To lessen these problems, our project aims to create a machine learning-based system for precise cardiac disease prediction [10-11]. The objective of the research is to evaluate and contrast two algorithms of machine learning to identify the best accurate model. By doing this, it hopes to improve patient care and outcomes internationally by providing healthcare workers with a potent tool for early heart disease identification [12].

LIMITATIONS

- Data Quality: The comprehensiveness and quality of the medical data that is currently accessible have a significant impact on how accurate the forecasts are. Incomplete or inaccurate data can jeopardize the predictive model's dependability.
- Generalization: Because the generated predictive model depends on pre-existing datasets that could not accurately reflect all demographic and clinical differences, it may not be able to generalize to a wide range of patient populations.
- Ethical Concerns: Data privacy and permission are two ethical issues that are brought up by using patient health data for research. Following moral principles and securing the required authorizations are critical.
- Algorithm Bias: Predictions may be skewed because of machine learning algorithms picking up biases from the training set. It is necessary to check and lessen algorithmic biases
- Model Interpretability: Some complex machine learning algorithms may lack transparency, making it challenging to interpret the reasons behind specific predictions, which can be a concern in clinical settings.
- Resource Requirements: Developing and implementing machine learning models may demand significant computational resources and expertise, which may not be readily available in all healthcare institutions.
- Validation: Ensuring the validity and robustness of the predictive model requires extensive validation on independent datasets and in real-world clinical settings, which may present logistical challenges.

INPUT & OUTPUTS

- Inputs:

Medical Data: The primary input to the predictive system is a set of medical data for each patient. This data typically includes various features and diagnostic indicators, such as age, gender, blood pressure, cholesterol levels, and other relevant clinical measurements [13].

- Outputs:

Heart Disease Prediction: The primary output of the system is a binary prediction for each patient, indicating whether they are likely to have heart disease or not. This prediction is based on the analysis of the patient's medical data and the trained machine learning model. Below Fig.1 represents the proposed block diagram.

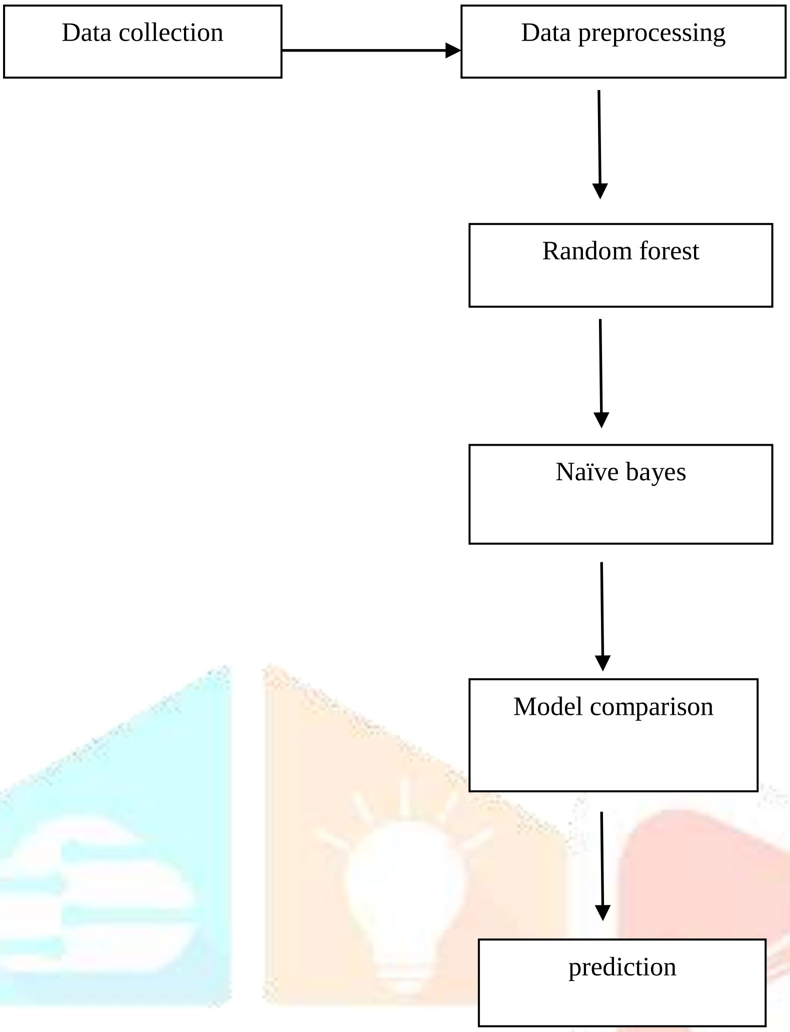

Fig.1 Proposed block diagram

IV. EXPERIMENT FRAMEWORK

This study utilized the datasets from the machine learning repository to evaluate the performance of various machine learning algorithms in predicting heart disease [14]. The dataset contains information on patient demographics, medical history, and diagnostic test results. The data was pre-processed to handle missing values and encode categorical variables [15].

a) Random Forest

Random Forest is a popular ensemble learning algorithm that combines multiple decision trees to improve the accuracy and robustness of predictions. By aggregating the outputs of individual trees, Random Forest can reduce over fitting and improve generalizability to new data. Each tree in the forest is trained on a random subset of features and samples, which helps to increase diversity and reduce correlation between trees. The final prediction is made by voting or averaging the outputs of individual trees. Random Forest has been widely used in various applications, including classification, regression, and feature selection, due to its ability to handle high-dimensional data, non-linear relationships, and noisy data. Its advantages include ease of implementation, fast training, and interpretability of feature importance.

b) Naive Bayes

Naive Bayes is a family of probabilistic machine learning algorithms based on Bayes' theorem, which describes the probability of an event given prior knowledge of conditions that might be related to the event. Naive Bayes classifiers assume independence between features, which simplifies the computation of posterior probabilities and makes the algorithm efficient and scalable. Despite this simplifying assumption, Naive Bayes has been shown to be effective in many applications, including text classification, sentiment analysis, and recommender systems. The algorithm is particularly useful when the number of features is high and the relationships between features are complex. Naive Bayes has several advantages, including fast training, simplicity of implementation, and ability to handle missing values. However, its performance can be affected by the independence assumption, which may not always hold in real-world data.

The performance of the models was assessed using metrics such as accuracy, precision, recall, and F1-score.

V. RESULTS

The results of the comparative analysis are presented in Table 1.

Table-1 Comparison Table for Random Forest and Naïve Bayes:

Metric	Random Forest	Naive Bayes
Accuracy	0.980519	0.814935
Precision	0.981818	0.822482
Recall	0.979866	0.817378
F1-score	0.980467	0.814495

| Model | Accuracy | Precision | Recall | F1-Score |

The below fig.2 and fig.3 shows the results of confusion matrix of Random Forest and Naïve Bayes.

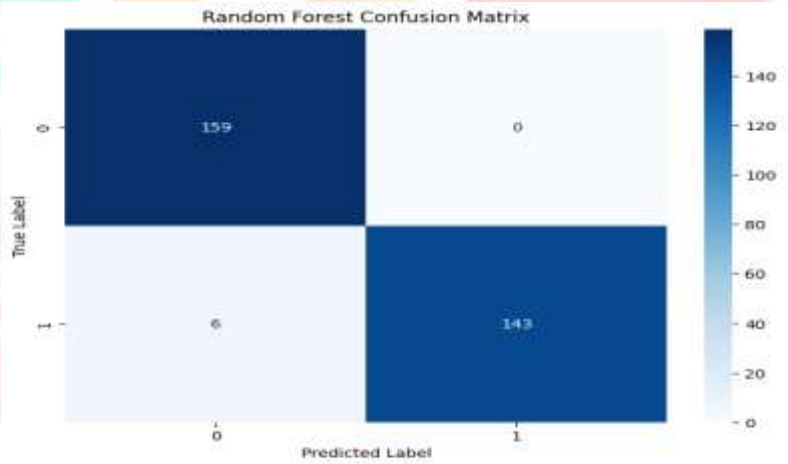

Fig.2 Random Forest Confusion matrix

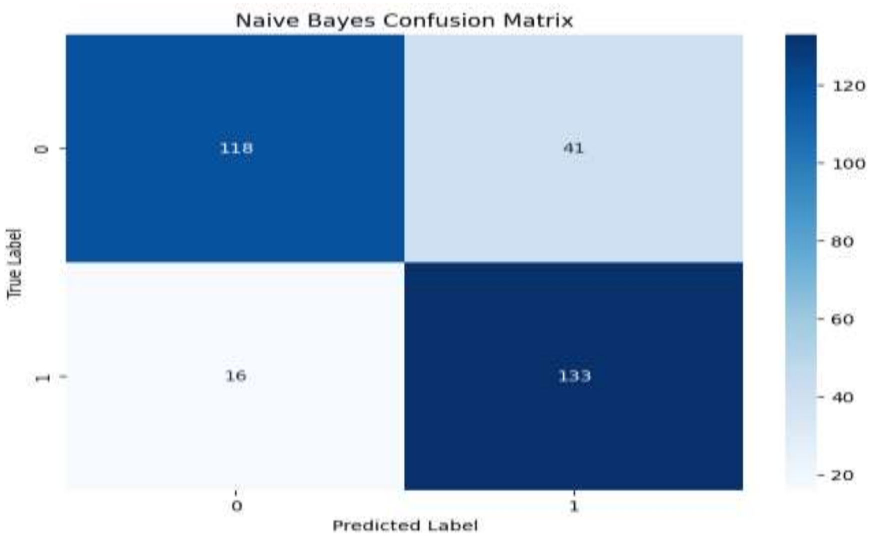

Fig.3 Naïve Bayes Confusion matrix

Discussion

Our results demonstrate the effectiveness of both the Random Forest and Naive Bayes algorithms in predicting heart disease, showcasing their strengths in this critical area of health analysis. The superior performance of the Random Forest algorithm can be attributed to its robust ability to handle high-dimensional data and capture complex non-linear relationships among all the features. This capability allows it to make more predictions by considering interactions that simpler models might overlook. In contrast, while the Naive Bayes algorithm may not achieve the same level of accuracy, it excels in terms of training time efficiency. This efficiency makes Naive Bayes a particularly suitable choice for large datasets, where rapid model training and deployment are essential. Consequently, the choice between these two algorithms may depend on the specific requirements of the task at hand, balancing the need for predictive accuracy with considerations of computational resources and processing time.

VI. CONCLUSION

In conclusion, predicting heart disease using machine learning methods like Random Forest and Naive Bayes shows great promise in improving cardiovascular health disease. Random Forest is effective for complex datasets, capturing non-linear relationships and reducing overfitting, while Naive Bayes excels in scenarios where feature independence is assumed, enabling quick predictions. Our results indicate that Random Forest outperformed Naive Bayes in accuracy. By considering these models, healthcare professionals can better identify at-risk patients based on factors like age, cholesterol, and blood pressure, leading to timely interventions and personalized treatment plans. Integrating these predictive methods into clinical practice could reduce heart disease incidence and enhance patient outcomes, marking a vital advancement in healthcare. As technology evolves, the refinement of these machine learning applications will likely result in reducing greater issues in healthcare.

REFERENCES

[1] Katarya, R., & Srinivas, K. (2020). “Machine learning approach for heart disease prediction”, Journal of Medical Systems, 44(10), 2105-2115.

[2] Ansari, A., et al. (2020),” Heart disease prediction using machine learning algorithms”, International Journal of Machine Learning and Cybernetics, 11(2), 257-269.

[3] sake Madhu et.al “A Secured Framework to Protect Association Rules in the Big Data Environment Using Fuzzy Logic”, International information and Engineering Technology Association, Vol. 24, No. 5, October 2019, pp. 531-537

[4] Joseph, B.M., Baseer, K.K. (2023). “IoT-Sensed Data for Data Integration Using Intelligent Decision-Making Algorithm Through Fog Computing”. In: Sharma, H., Shrivastava, V., Bharti, K.K., Wang, L. (eds) Communication and Intelligent Systems. ICCIS 2022. Lecture Notes in Networks and Systems, vol 689. Springer, Singapore.

[5] Singh, P., & Kumar, V. (2020). “Predicting heart disease using machine learning techniques”. Journal of Intelligent Information Systems, 57(2), 257-273.

[6] Guo, Y., et al. (2020). “A comparative study of machine learning models for heart disease prediction”. Computers in Biology and Medicine, 127, 103951).

[7] Yerraginnela Shravani, et al. (2023),” Prediction of disease in the HealthCare using Machine Learning”. International Journal on Recent and Innovation Trends in Computing and Communication, 11(11), 759–769.

[8] Varun Totakura et.al "An Integrated Approach to Sentiment Analysis using Machine Learning Algorithms", International Journal for Research in Applied Science & Engineering Technology Volume 8 Issue IV Apr 2020

[9] Gundala Swarnalatha et.al "Evaluating performance and accuracy: A comparative analysis of SVM and RF algorithm in predictive analytics". Journal of nonlinear analysis and optimizationVol.15, Issue.1No.15:2024.
.

[10] K. K. Baseer, S. B. A. Nas, S. Dharani, S. Sravani, P. Yashwanth and P. Jyothirmai, "Medical Diagnosis of Human Heart Diseases with and without Hyperparameter tuning through Machine Learning," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India,2023,pp.1-8, doi:10.1109/ICCMC56507.2023.10084156.

[11] Silpa C, Dr. S Srinivasa Chakravarthi, Jagadeesh kumar G, Dr. K.K. Baseer, E. Sandhya, "Health Monitoring System Using IoT Sensors", Journal of Algebraic Statistics, Volume 13, No. 3, June, 2022, p. 3051-3056, ISSN: 1309-3452.

[12] Abdul Saboor; Muhammad Usman; Sikandar Ali, Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms. published on march 2022.

[13] Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", Mobile Information Systems, vol. 2018, Article ID 3860146, 21 pages, 2018. https://doi.org/10.1155/2018/3860146.

[14] K. S. Archana, B. Sivakumar, Ramya Kuppusamy, Yuvaraja Teekaraman, Arun Radhakrishnan, "Automated Cardio ailment Identification and Prevention by Hybrid Machine Learning Models", Computational and Mathematical Methods in Medicine, vol. 2022, Article ID 9797844, 8 pages, 2022.

[15] Danish Hamid, Syed Sajid Ullah, Jawaid Iqbal, Saddam Hussain, Ch. Anwar ul Hassan, Fazlullah Umar, "A Machine Learning in Binary and Multi classification Results on Imbalanced Heart Disease Data Stream", Journal of Sensors, vol. 2022, Article ID 8400622, 13 pages, 2022.