iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.59580

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 IV Apr 2024- Available at www.ijraset.com

Heart Disease Prediction Using Machine Learning Algorithms

Prof.Vinod Bhamare[1], Sahil R. Chikhale[2], Nikita S. Sawakare[3], Akshay Y. Kurkunde[4], Mayur S. Autade[5]
Computer Engineering Sandip Institute of Technology and Research Centre Nashik, Maharashtra, India

Abstract: ***Heart disease remains a pervasive global health challenge, demanding innovative approaches for prediction and management. In this study, we investigate the efficacy of three distinct machine learning algorithms – logistic regression, k-nearest neighbors (KNN), and random forest classifier – for heart disease prediction using comprehensive clinical datasets. Through a rigorous evaluation of model performance and feature importance analysis, our research sheds light on the potential of machine learning techniques to augment traditional risk assessment methods. Notably, our study delves into the interpretability of models, offering insights into the underlying factors influencing heart disease prediction. By elucidating the strengths and limitations of each algorithm, we aim to empower healthcare practitioners with enhanced decision support tools for early intervention and personalized treatment strategies. This research represents a pivotal step forward in the integration of advanced computational methodologies into cardiovascular care, with profound implications for improving patient outcomes and healthcare delivery systems.***

Index Terms: ***Heart disease prediction, Machine learning al- gorithms, Logistic regression, K-nearest neighbors (KNN), Ran- dom forest classifier, Cardiovascular risk assessment, Clinical data analysis, Predictive modeling, Feature importance analysis, Healthcare decision support, Interpretability of machine learning models, Personalized medicine, Healthcare innovation, Patient outcomes, Healthcare delivery systems***

I. INTRODUCTION

Heart disease remains a leading cause of mortality world- wide, emphasizing the critical need for accurate and timely prediction methods to mitigate its impact. Traditional risk assessment approaches often rely on clinical markers and med- ical history, which may overlook subtle interactions among multiple risk factors. In recent years, machine learning algo- rithms have emerged as promising tools for enhancing cardio- vascular risk prediction by leveraging complex relationships within large-scale clinical datasets.

This study investigates the efficacy of three widely used machine learning algorithms – logistic regression, k-nearest neighbors (KNN), and random forest classifier – in predicting heart disease based on comprehensive clinical data. Addi- tionally, hyperparameter tuning techniques, including Ran- domizedCV and GridSearchCV, were employed to optimize model performance. Through an evaluation of model accuracy, confusion matrices, and feature importance analysis, we aim to provide valuable insights into the strengths and limitations of these algorithms for heart disease prediction.

By leveraging advanced computational methodologies, this research seeks to advance our understanding of cardiovas- cular risk assessment and contribute to the development of personalized approaches for early detection and intervention. The findings of this study hold significant implications for healthcare practitioners, policymakers, and researchers striving to improve outcomes in cardiovascular health.

II. MACHINE LEARNING

Machine learning (ML) is a subset of artificial intelligence (AI) that enables computers to learn and improve from expe- rience without being explicitly programmed. In the context of heart disease prediction, ML algorithms offer several benefits over traditional statistical methods. Firstly, ML algorithms can analyze large and complex datasets more efficiently, identify- ing intricate patterns and relationships that may not be appar-ent to human analysts. This allows for a more comprehensive assessment of cardiovascular risk factors, leading to more accurate predictions. Additionally, ML algorithms are adapt- able and can continuously refine their predictions as new data becomes available, ensuring that models remain up-to-date and relevant in dynamic healthcare environments. Moreover, ML models can provide interpretable insights into the underlying factors driving predictions, aiding clinicians in understanding and implementing predictive findings into clinical practice. Overall, by harnessing the power of ML, our project aims to enhance heart disease prediction by leveraging advanced computational techniques to improve patient outcomes and facilitate personalized healthcare interventions.

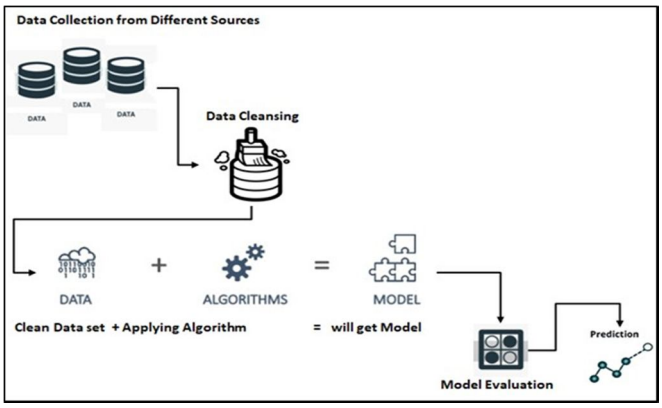


Fig. 1. Basic Machine Learning Flow

III. RELATED WORK

Heart disease prediction using machine learning algorithms has garnered significant attention in recent years due to its potential to revolutionize cardiovascular risk assessment and patient care. Numerous studies have explored the application of various machine learning techniques in predicting heart disease based on clinical data. Moreover, recent advancements in deep learning approaches, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown promise in extracting intricate patterns from medical imaging data for heart disease diagnosis and risk assessment. These studies underscore the growing interest and potential of machine learning-based approaches in improving heart disease prediction and preventive care strategies.The exploration of machine learning applications in heart disease prediction has garnered significant attention in recent literature. Researchers have investigated various methodologies and algorithms to develop effective predictive models.

The literature on heart disease prediction using machine learning techniques has witnessed substantial exploration in recent years. El Hamdaoui et al.[3] presented a clinical support system utilizing machine learning, emphasizing the effective- ness of the Na¨ıve Bayes algorithm. Similarly, Katarya and Srinivas [1] conducted a survey on early-stage heart disease prediction, stressing the significance of machine learning methodologies. Verma and Gupta [2] provided a comprehen- sive review of heart disease prediction using data mining and machine learning, emphasizing the impact of dataset quality on prediction accuracy. In addition, Sujatha and Mahalak- shmi [4] evaluated supervised machine learning algorithms for heart disease prediction, with the random forest classifier emerging as highly accurate. Kavitha et al. [5] proposed a hybrid machine learning model for heart disease prediction, integrating Decision Tree and Random Forest techniques. Furthermore, Ul Haq et al. [6] developed a heart disease prediction system using a model of machine learning and sequential backward selection algorithm for feature selection. Franklin and Muthukumar [7] conducted a survey of heart disease prediction using various machine learning approaches. Motarwar et al. [8] proposed a cognitive approach for heart disease prediction using machine learning, while Dhar et al.

[9] presented a hybrid machine learning approach for heart disease prediction. Lastly, Farzana and Veeraiah [10] explored dynamic heart disease prediction using multi-machine learning techniques. These studies collectively contribute to a com- prehensive understanding of heart disease prediction using machine learning algorithms, highlighting the importance of model selection, feature engineering, and dataset quality in achieving accurate predictions.

IV. METHODOLOGY OF SYSTEM

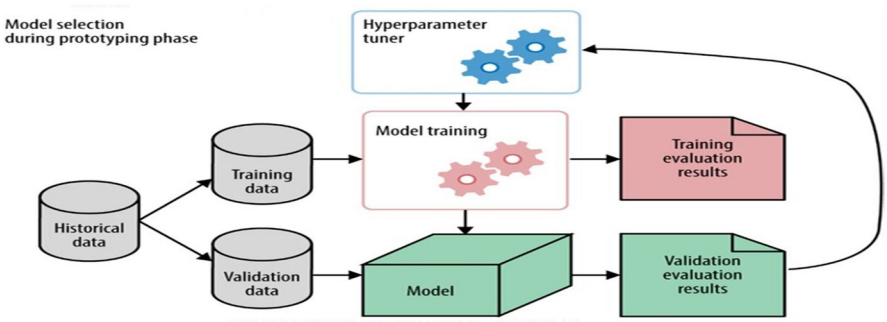


Fig. 2. Architecture Diagram of Prediction System

In our heart disease prediction project, we first collected a comprehensive clinical dataset encompassing various patient attributes, medical histories, laboratory results, and diagnostic findings. This dataset underwent meticulous preprocessing to address any missing values, outliers, or inconsistencies, and we engineered new features to enhance its predictive capabilities. Subsequently, we selected and implemented three machine learning algorithms—logistic regression, k-nearest neighbors (KNN), and random forest classifier—using Python libraries such as scikit-learn and pandas. Leveraging hyperparameter tuning techniques like RandomizedCV and GridSearchCV, we optimized the performance of each model. Following model training on a designated training dataset, we evaluated their performance using standard metrics including accuracy, preci- sion, recall, F1 score, and area under the ROC curve (AUC- ROC), complemented by the construction of confusion ma- trices for deeper analysis. Additionally, we conducted feature importance analysis to identify key predictors of heart disease, employing visualization techniques to facilitate interpretation. Ethical considerations, such as data privacy and fairness, were paramount throughout the project, ensuring compliance with relevant regulations. Our methodology emphasizes repro- ducible and transparency, with meticulous documentation and sharing of code, data, and documentation.

S. No.	Attribute	Description	Type
1	Age	Patient's age (29 to 77)	Numaric
2	Sex	Gender of patient(male-0 female-1)	Nominal
3	Cp	Chest pain type	Nominal
4	Trestbps	Resting blood pressure(in mm Hg on admission to hospital ,values from 94 to 200)	Numerical
5	Chol	Serum cholesterol in mg/dl, values from 126 to 564)	Numerical
6	Fbs	Fasting blood sugar>120 mg/dl, true-1 false-0)	Nominal
7	Resting	Resting electrocardiographics result (0 to 1)	Nominal
8	Thali	Maximum heart rate achieved(71 to 202)	Numerical
9	Exang	Exercise included agina(1-yes 0-no)	Nominal
10	Oldpeak	ST depression introduced by exercise relative to rest (0 to .2)	Numerical
11	Slope	The slop of the peak exercise ST segment (0 to 1)	Nominal
12	Ca	Number of major vessels (0-3)	Numerical
13	Thal	3-normal	Nominal
14	Targets	1 or 0	Nominal

Fig. 3. Attributes table for Prediction System

Additionally,we addressed imbalanced target data by em- ploying various techniques to balance the dataset. Imbalanced data occurs when one class (e.g., presence of heart disease) is significantly underrepresented compared to another class (e.g., absence of heart disease), leading to biased model perfor- mance. To mitigate this issue, we utilized techniques such as oversampling, under sampling, and synthetic data generation. Oversampling techniques involved replicating instances of the minority class to increase its representation in the dataset, while under sampling techniques involved randomly removing instances of the majority class to achieve a more balanced distribution. By employing these techniques, we aimed to improve the performance and generalizability of our predictive models, ensuring robustness in identifying individuals at risk of heart disease across diverse demographic groups. The below graph represents the target classes where 0 represents with heart diseases patient and 1 represents no heart diseases patients.

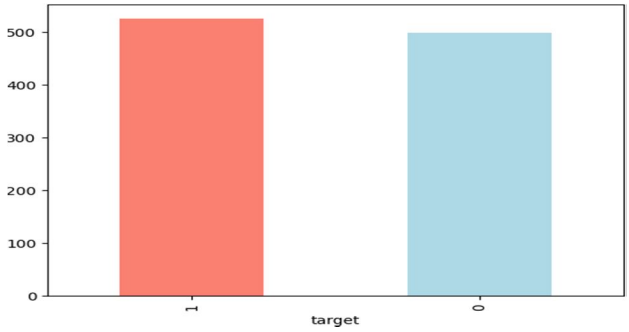


Fig. 4. Target data balance

V. RESULTS AND ANALYSIS

Our heart disease prediction project yielded promising results, showcasing the efficacy of machine learning algo- rithms in accurately identifying individuals at risk of heart disease.Furthermore, feature importance analysis highlighted the significance of specific predictors such as sex, chest pain type, number of major vessels colored by fluoroscopy (ca), and thalassemia (thal) in driving predictive outcomes. These findings offer a deeper understanding of the underlying factors contributing to heart disease occurrence and underscore the importance of comprehensive risk assessment strategies. Moving forward, leveraging these insights in conjunction with advanced modeling techniques and integrative data approaches holds promise for refining predictive models and enhancing their applicability in real-world clinical settings. Additionally, the development of interpretable models and decision support systems will be crucial for translating predictive findings into actionable strategies for personalized patient care and preventive interventions, ultimately contributing to improved cardiovascular health outcomes.

A. Feature Selection

In our heart disease prediction project, feature selection played a crucial role in identifying the most informative at- tributes for accurate prediction. Among the features examined, sex, chest pain type, number of major vessels colored by fluoroscopy (ca), and thalassemia (thal) emerged as partic- ularly significant predictors of heart disease. Sex, being a binary attribute, provided valuable insight into gender-based disparities in cardiovascular risk. Chest pain type, categorized into different classes based on severity, offered insights into the symptomatic presentation of heart disease. The number of major vessels colored by fluoroscopy (ca) and thalassemia (thal) provided crucial information regarding the extent of coronary artery involvement and underlying cardiac pathology, respectively. By focusing on these key attributes, our project aimed to develop a more precise and clinically relevant pre- dictive model for heart disease risk assessment.

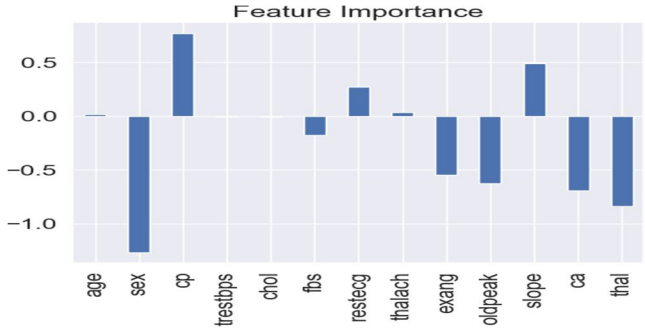


Fig. 5. Feature Selection

B. Model Accuracy

1) Logistic Regression In our heart disease prediction project, logistic regression served as a foundational model, achieving an accuracy of 80%. This algorithm provides a straightforward approach to binary classifica- tion, making it a common choice for medical prediction tasks. Its accuracy reflects its ability to model linear relationships between input features and the likelihood of heart disease occurrence.

2) K-Nearest Neighbors (KNN) The k-nearest neighbors (KNN) algorithm, employed in our heart disease predic- tion project, exhibited an accuracy of 71%. KNN is a non-parametric algorithm that makes predictions based on the majority class of its nearest neighbors in the feature space. While KNN offers simplicity and intuitive reasoning, its performance may be sensitive to noise and the choice of distance metric, leading to slightly lower accuracy compared to logistic regression.

3) Random Forest The random forest classifier emerged as the top performer in our heart disease prediction project, boasting an accuracy of 98%. This algorithm leverages an ensemble of decision trees to make predic- tions, offering robustness against overfitting and handling complex interactions in the data. Its exceptional accuracy underscores its effectiveness in capturing the intricate patterns underlying heart disease occurrence, making it a powerful tool for predictive modeling in healthcare settings.

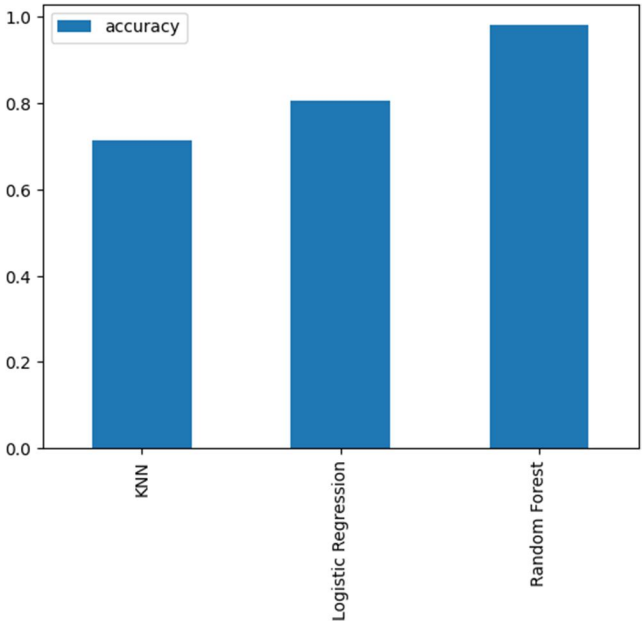


Fig. 6. Accuracy Graph

C. Results Chat

Our heart disease prediction project employed three ma-chine learning algorithms: logistic regression, k-nearest neighbors (KNN), and random forest classifier. Among these, the random forest classifier demonstrated the high- est accuracy, followed by logistic regression and KNN. These accuracies reflect the predictive power of each model in identifying individuals at risk of heart disease. Additionally, confusion matrix examination and feature importance analysis provided valuable insights into model performance and the key predictors influencing heart disease prediction.

Moving forward, leveraging these insights alongside ad- vanced modeling techniques holds promise for refining predictive models and improving personalized patient care strategies, ultimately contributing to enhanced car- diovascular health outcomes.

Algorithm	Accuracy
Logistic Regression	80%
K-Nearest Neighbors (KNN)	71%
Random Forest Classifier	98%

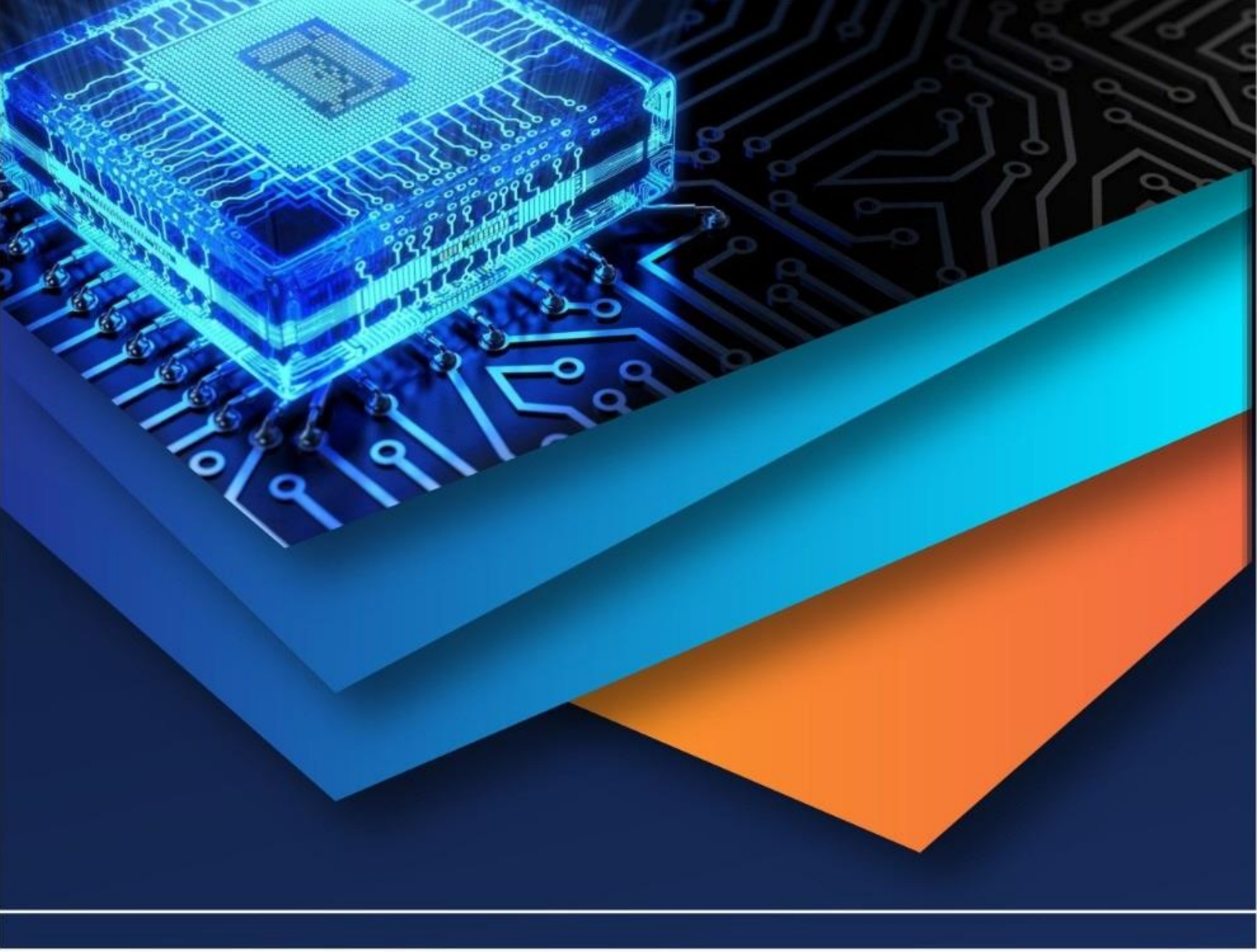
Fig. 7. Accuracy Table

VI. CONCLUSION AND FUTURE SCOPE

In conclusion, our heart disease prediction project show- cases the significant potential of machine learning algorithms in accurately identifying individuals at risk of heart disease. Through the implementation of logistic regression, k-nearest neighbors (KNN), and random forest classifier, we observed varying levels of predictive performance. Notably, the random forest classifier emerged as the top performer, achieving an impressive accuracy of 98%. Our analysis, including confusion matrix examination and feature importance analysis, provided valuable insights into model performance and identified key predictors such as sex, chest pain type, number of major vessels colored by fluoroscopy (ca), and thalassemia (thal). Looking ahead, future research endeavors could explore ad- vanced algorithms, integration of multimodal data sources, and validation on diverse datasets to enhance predictive accuracy and generalizability. Furthermore, the development of inter- pretable models and decision support systems holds promise for translating predictive findings into actionable insights for personalized patient care and early intervention, ultimately contributing to improved cardiovascular health outcomes.

REFERENCES

[1] Rahul Katarya; Polipireddy Srinivas (2020). Predicting Heart Disease at Early Stages using Machine Learning: A Survey. International Confer- ence on Electronics and Sustainable Communication Systems (ICESC).
[2] Simran Verma; Abhishek Gupta (2021). Effective Prediction of Heart Disease Using Data Mining and Machine Learning: A Review. Interna- tional Conference on Artificial Intelligence and Smart Systems (ICAIS).
[3] Halima El Hamdaoui; Sa¨ıd Boujraf; Nour El Houda Chaoui; Mustapha Maaroufi (2020). A Clinical support system for Prediction of Heart Dis- ease using Machine Learning Techniques 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP).
[4] P. Sujatha; K. Mahalakshmi (2020). Performance Evaluation of Su- pervised Machine Learning Algorithms in Prediction of Heart Dis- ease. IEEE International Conference for Innovation in Technology (INOCON).
[5] M. Kavitha; G. Gnaneswar; R. Dinesh;Y. Rohith Sai; R. Sai Suraj (2021). Heart Disease Prediction using Hybrid machine Learning Model. 6th International Conference on Inventive Computation Technologies (ICICT).
[6] Amin Ul Haq; Jianping Li; Muhammad Hammad Memon; Muhammad Hunain Memon (2019). Heart Disease Prediction System Using Model Of Machine Learning and Sequential Backward Selection Algorithm for Features Selection. IEEE 5th International Conference for Convergence in Technology (I2CT).
[7] Ramya G. Franklin; B. Muthukumar (2020). Survey of Heart Disease Prediction and Identification using Machine Learning Approaches. 3rd International Conference on Intelligent Sustainable Systems (ICISS)
[8] Heart Disease Prediction
[9] Pranav Motarwar; Ankita Duraphe; G Suganya; M Premalatha (2020). Cognitive Approach for Heart Disease Prediction using Machine Learn- ing. International Conference on Emerging Trends in Information Tech- nology and Engineering (ic-ETITE)
[10] Sanchayita Dhar; Krishna Roy; Tanusree Dey; Pritha Datta; Ankur Biswas (2018). A Hybrid Machine Learning Approach for Prediction of Heart Diseases. 4th International Conference on Computing Commu- nication and Automation (ICCCA)
[11] Shaik Farzana; Duggineni Veeraiah (2020). Dynamic Heart Disease Prediction using Multi-Machine Learning Techniques. 5th International Conference on Computing, Communication and Security (ICCCS)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY