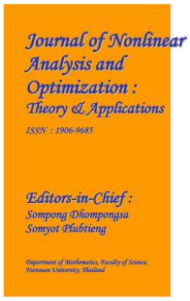
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1 : 2024
ISSN : **1906-9685**



HEART DISEASE DETECTION USING XGB-CLASSIFIER AND FAILUREPREDICTION USING GRADIENT BOOSTING

Dr. S. NagaMallik Raj, Associate Professor of Computer Science Department (CSE),Vignan's Institute of Information Technology, Vishakapatnam, India, mallikblue@gmail.com
R Sree Vani , Computer Science Department (CSE), Vignan's Institute of InformationTechnology, Vishakapatnam, India, sreevaniramisetti@gmail.com
Bandi Raja, Computer Science Department (CSE), Vignan's Institute of InformationTechnology, Vishakapatnam, India, bandiraja2018@gmail.com
Tirukkovalluri Sri Harsha, Computer Science Department (CSE), Vignan's Institute ofInformation Technology, Vishakapatnam, India, harsha888@gmail.com
Tedlapu Drakshayani, Computer Science Department (CSE), Vignan's Institute ofInformation Technology, Vishakapatnam, India, tedlapu2002@gmail.com
Ranguri Charith, Computer Science Department (CSE), Vignan's Institute ofInformation Technology, Vishakapatnam, India charithranguri@gmail.com

Abstract:
Approaches to detecting and addressing heart conditions stand as critical milestones toward enhancing patient outcomes and preventative care strategies. This research introduces an innovative "Heart Disease Detection using XGB-Classifier and Failure Prediction using Gradient Boosting," a comprehensive web-based application designed to bridge the gap between early detection of cardiac conditions and the subsequentvulnerability to heart failure. Utilizing the UCI(University of California, Irvine) Heart
Disease Dataset, our system empowers users to assess their heart disease risk through a user-friendly interface, offering insights based on a range of clinical and lifestyle factors. Upon confirming the existence of cardiac conditions, the application further classifies the type of the disease and guides users through stages of risk and provides tailored information on precautions, symptoms, potential causes, and recommendations for professional consultation. Unique to this system is the integration of a subsequent heart failure predictionmodule, which, despite the exclusion of the 'time' feature due to its impracticality for new users, leverages advanced machine learning models—XG Boost and Gradient Boosting—topredict the likelihood of heart failure with impressive accuracies of 85% and 93%, respectively. Our findings indicate significant potential for machine learning applications in personalized healthcare, providing a scalable and effective tool for heart disease management. This paper contributes to the field by outlining the creation and validation of amultifaceted predictive system, underscoring the importance of accessibility and user-centricdesign in healthcare technologies.

Keywords :
Heart Disease Prediction, Machine Learning, XG Boost and Gradient Boosting,Personalised.

I. INTRODUCTION

Heart conditions remain a primary contributor to both illness and death worldwide, presenting considerable hurdles for medical systems and impacting the well-being of countless individuals. The WHO(World Health Organization) reports that heart conditions account for approximately 12 million fatalities around the globe annually[2].Timely identification and precise forecasting of heart condition stages are crucial for appropriate intervention, management, and enhancing patient outcomes. It also reducing the physical and financial burden on individuals and healthcare institutions[1].Despite progress in medical technology and healthcare, there is an urgent demand for novel, accessible, and

effective solutions for predicting cardiac conditions. Contemporary advances in computational intelligence, specifically within the realms of machine learning and AI, demonstrate significant promise in revolutionizing healthcare provisions, especially concerning thediagnosis and forecasting of diseases. These technologies offer the possibility of harnessingcomplex clinical data to construct forecasting tools that can support medical professionals indecision-making processes and enhance patient care. This paper introduces a novel Hear
Disease Detection using XGB-Classifier and Failure Prediction System using Gradient Boosting that leverages machine learning techniques to offer a comprehensive solution for early identification and prediction of cardiac conditions. The system employs the XG Boostclassifier for disease prediction and a Gradient Boosting-based model for failure prediction,with a notable exclusion of the "time" feature to ensure applicability to new users. Our approach is distinct in its stage-wise prediction of heart disease types, integration of user- centric information on precautions, causes, symptoms, and recommended consultations, andthe inclusion of a failure prediction feature to assess patient prognosis. The motivation behindthis research is twofold. First, to address the gap in accessible and user-friendly tools for heart disease prediction that can empower patients with knowledge and support healthcare providers in early detection efforts. Second, to evaluate the effectiveness of machine learningstrategies in handling complex clinical data and forecasting health outcomes with high accuracy.

II. LITERATURE REVIEW

Heart conditions remains a leading reason of death globally, with various factors contributingto its onset and progression. Traditional models for forecasting heart conditions havedepended on statistical methods and risk factor analysis, focusing on age, sex, hypertension indicators, cholesterol measurements, diabetes presence, and smoking habits[3]. Whileeffective to a degree, these models often lack the precision and personalization necessary for prompt identification and proactive management. The authors expanded the research datasetby incorporating two additional variables: dietary fat intake and smoking behaviour. To make predictions, they employed data exploration techniques, encompassing Tree-based Models, Bayesian Classifiers and Neural Architectures, and conducted analyses on a heart disease database[3][4].In another research, the team centered on creating a prediction model for heartrelated conditions utilizing machine learning methods. They employed data utilizing the dataset from the Cleveland heart study, featuring 303 cases and 17 variables, which was employed from the UCI Machine Learning Repository[1][5]. Several ML algorithms have shown promise in heart disease forecasting, highlighting the distinct advantages and challenges of each approach. Another study leveraged multiple machine learning models, conducting thorough comparisons with their ensemble versions based on accuracy, specificity, and sensitivity. This approach aimed to determine the most effective model for clinical applications. The system and model developed were grounded on the 'Cardiovascular
Heart Disease' dataset, accessible publicly on Kaggle[6]. Ischemic Heart Disease (IHD) serves as a primary underlying cause of heart failure (HF) [7][8] and is associated with highermortality rates. In a study involving 306 individuals diagnosed with ischemic heart disease (IHD), it was found that 64.1% progressed to heart failure (HF). Notably, the proportion of female patients was approximately double that of male patients[9]. A different research effort highlighted the use of a composite dataset encompassing demographic details, incorporatingchest X-ray images, and clinical diagnostics as well as symptoms from 100 patients with heart failure. This comprehensive approach was employed to develop predictive models based on Random Forest (RF) alongside Logistic Regression (LR) techniques, both of whichreported a prediction accuracy rate of 93%[7][10]. Despite the advances brought by ML in heart disease prediction, challenges remain. Data privacy, ethical considerations, and the need for large, annotated datasets are ongoing concerns. This research adds to the existing literature by implementing an XG Boost classifier for the initial detection of heart disease, followed by a Gradient Boosting model for mortality risk assessment. Unique to this study is the stage-wise prediction of heart disease types and the comprehensive integration of user-centric information, filling a gap in the current literature on personalized and accessible heartdisease management tools.

III. PROPOSED SYSTEM & METHODOLOGY

The objective of this research was to devise an advanced, interactive system capable ofaccurately

predicting heart disease and evaluating the likelihood of heart failure amongpatients, employing a user-centric web interface for real-time health assessments. This section delineates the methodology adopted, encompassing data collection, model development, and the distinctive attributes of the suggested system.

A. Data Collection:

Data for this study were meticulously curated from two primary sources: The dataset for Heart Disease was sourced from the UCI ML Repository and Heart failure risk assessment dataset was obtained from Kaggle's dataset of Heart Failure Clinical Records. Both datasetsunderwent rigorous preprocessing, including cleaning, normalization, and missing value imputation, to ensure data integrity and reliability for model training. A noteworthy aspect of our approach was the exclusion of the 'time' variable from the heart failure dataset, motivated by the consideration that new users would not have corresponding values,potentially skewing predictions.

The forecasting of likelihood of heart disease occurrence segment of our study leverages a dataset curated from the UCI ML Repository, renowned for its comprehensive compilation of clinical, demographic, and diagnostic variables pertinent to heart disease. This dataset encompasses an array of characteristics containing age, gender, chest discomfort categories, baseline hypertension levels, cholesterol measurements, fasting blood glucose indicators, finding from resting electrocardiograms, peak heart rate achieved, induction of chest pain byphysical activity, ST segment depression resulting from exercise compared to inactivity, inclination of the exercise ST elevation, count of significant vessels identified through fluoroscopy, and the detection of thal(defect type). Each feature within this dataset has beenmeticulously selected based on its established correlation with heart disease, as documentedin medical research and clinical practice. The diversity and clinical relevance of these features ensure a robust foundation for crafting a potent predictive system for heart disease,aimed at facilitating early detection and intervention.

Sno	Attributes	Description
1	Age	Age in years
2	Sex	Male or Female
3	Cp	Chest pain type
4	Thestbps	Resting blood pressure
5	Chol	Serum cholesterol
6	Restecg	Resting electrographic results
7	Fbs	Fasting blood sugar
8	Thalach	Max. heart rate achieved
9	exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment
12	Ca	No. of major vessels colored
13	Thal	Defect type

Fig-1: Heart Disease Prediction Dataset

For the heart failure risk assessment, our study utilizes a dataset available on Kaggle's platform, specifically designed to assess the chance of heart failure events. Unlike the heartdisease prediction dataset, this collection focuses on clinical and physiological markers thatare critical in assessing the severity and prognosis of existing heart conditions. Characteristics encompass individual age, identification of anaemia, creatinine kinase level,diabetes condition, cardiac output rate, increased blood pressure, platelet quantity, creatinine

concentration in serum, sodium levels in serum, gender, and smoking history. Notably, we made a strategic decision to exclude the 'time' variable from our analysis. This choice was predicated on the rationale that new users, the primary beneficiaries of this predictive system,would not possess historical data, rendering the 'time' variable irrelevant and potentially misleading in their context. This dataset's

composition reflects a targeted approach to capturing the multifaceted nature of heart failure risk, laying the groundwork for a nuanced predictive model that transcends mere detection to offer actionable insights into heart failureprogression.

S.no.	**Features**
1	Age
2	Anaemia
3	High Blood Pressure
4	Creatinine Phosphokinase-CPK
5	Diabetes
6	Ejection Fraction
7	Sex
8	Platelets
9	Serum Creatinine
10	Serum Sodium
11	Smoking
12	Time-Follow Up Period
13	Death Event (Target)

Fig-2: Heart Failure Prediction Dataset

By judiciously selecting and curating data from the esteemed UCI ML Repository and Kaggle's rich repository of Heart Failure Clinical Records, we have established a robust dataset foundation that encompasses a wide spectrum of variables critical to heart disease and failure prediction.

B. Data Preprocessing and Model Development:

The datasets were processed to maintain data integrity and consistency. This stage involvedcleaning (e.g., handling missing values), normalization (to standardize the scale of numerical features), and transforming categorical variables into numeric format through encoding techniques. The goal was to ready the datasets for effective ML analysis, minimizing potential biases and improving the models' predictive performance. In the effort of developing a strong heart disease detection and prediction system, we employed a comparative evaluation of different machine learning techniques, leveraging their unique strengths and evaluating their performance across two datasets: one for heart diseasedetection employed from the UCI ML Repository, and another for heart failure risk assessment obtained from Kaggle's dataset of Clinical Records for Heart Failure.

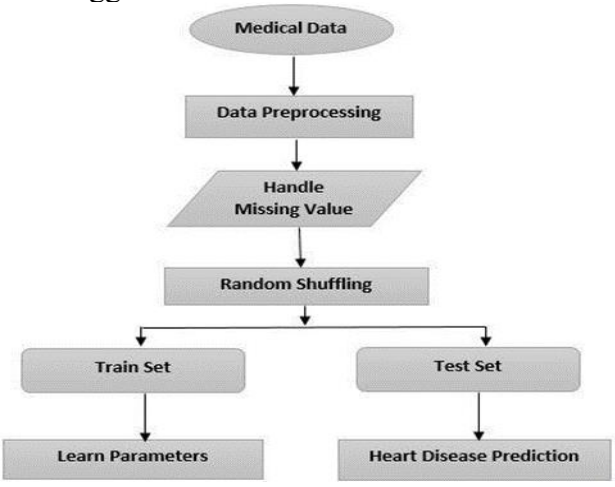


Fig-3: Framework

The algorithms evaluated included:

- **Logistic Regression:** A initial model for dichotomous outcome prediction, knownfor its simplicity and interpretability.
- **Decision Tree Classifier:** A model employing non-linear decision boundaries to segment the dataset into branches to form a tree structure, making it easy to interpret.
- **Random Forest Classifier:** A method that aggregates several decision trees to enhance classification accuracy rate and control over-fitting.
- **K-Neighbors Classifier (KNN):** A non-parametric approach in which data points are classified according to the most common category within their neighbourhood, utilising the concept of proximity for categorisation.
- **Gradient Boosting Classifier:** A composite approach that construct models in a sequential manner, with each subsequent model addressing and correcting the inaccuracies of the predecessors.
- **XG Boost Classifier:** An advanced, optimized distributed gradient boosting framework known for its high efficiency, adaptability, and portability.
- **AdaBoost Classifier:** Another composite approach that integrate numerous weak orbasic learners to construct a robust classifier.
- **Gaussian Naive Bayes:** A probabilistic method that employ s Bayes' theorem underthe 'naïve' presumption that all pairs of features are independent of each other.
- **Support Vector Machine (SVM):** A sturdy classifier capable in multidimensional contexts, particularly efficient when the dimensionality surpasses the sample count.

Each algorithm was rigorously tested using a combination of cross-validation techniques andgrid search for hyperparameter optimization, aiming to balance accuracy with computational efficiency. The evaluation metrics focused on accuracy, considering the critical nature of both detecting heart disease accurately and predicting heart failure risk reliably[11][12].

C. User Interface:

A significant innovation of this study is the creation of an extensive online platform that notonly supports the forecasting of heart disease and heart failure risk but also provides users with detailed, stage-specific health information. The interface includes sections such as 'Predict', where users input personal health data; 'About', offering a comprehensive examination of the project; and 'Details', offering detailed perspectives on the datasets and models utilized. A unique feature is the system's capability to classify heart disease into stages and offer corresponding health advice, including symptoms, causes, precautions, and recommended specialist consultations. This stage-wise classification and the dual predictioncapability for both heart disease and heart failure risk underscore the system's holistic approach to cardiovascular health assessment.

In conclusion, the methodology adopted for this research integrates rigorous data preprocessing, careful model selection based on a comprehensive evaluation framework, andthe creation of an intuitive web-based platform. The system's unique features, including its stage-wise disease classification and comprehensive health information provision, set it apartas an innovative tool in the domain of heart health prediction and management.

IV. RESULTS AND DISCUSSIONS

This study embarked on a mission to upgrade the detection and forecasting of heart diseaseand heart failure risk using sophisticated algorithmic methods. Through a meticulous comparison of multiple algorithms, the XG Boost Classifier emerged as the top-performing approach for heart disease detection, achieving an accuracy of 85%. Concurrently, the Gradient Boosting Classifier was identified as the optimal strategy for assessing heart failure risk potential, with a commendable accuracy of 93%. These results not only validatethe effectiveness of ensemble learning methods in medical informatics but also underscore the potential of predictive analytics in preemptive healthcare strategies.

The deployment of the XG Boost Classifier for heart disease detection signifies a noteworthy advancement in diagnostic accuracy. This model's ability to handle a variety ofdata types, manage missing values, and its robustness against overfitting contributes significantly to its superior performance. An accuracy of 85% indicates a high level of reliability, essential for clinical applications

where the early detection of heart disease can drastically alter patient outcomes. However, it's important to consider the equilibrium between sensitivity and specificity in this scenario; guaranteeing that the model not only identifies most patients with the disease but also minimizes false positives is crucial for practical use.

The selection of the Gradient Boosting Classifier for heart failure prediction underscores themodel's capacity to leverage complex interactions between features to improve prediction accuracy. Securing an accuracy rate of 93% emphasizes the potential of machine learning in prognostic evaluations, offering healthcare practitioners valuable insights into patient riskprofiles. This model's predictive power can facilitate early intervention strategies, ultimately improving patient quality of life and survival rates. However, it's critical to acknowledge thechallenges of generalizing these results across diverse patient populations, emphasizing the need for further validation studies.

Implementing these predictive algorithms in clinical settings could transform heart disease management and heart failure, offering a more personalized and proactive approach to patient care. By enabling prompt identification and risk evaluation, healthcare providers cantailor treatment plans more effectively, potentially reducing the incidence of severe outcomes and associated healthcare costs. Moreover, these findings highlight the significance of interdisciplinary partnership between data scientists and healthcare experts to refine and
adapt predictive models for real-world clinical environments.

V. CONCLUSION AND FUTURE SCOPE

This study set out to tackle the vital issue of early identification and risk forecasting for heartdisease and heart failure through advanced machine learning approaches. Through a rigorousevaluation of various algorithms, we have identified the XG Boost Classifier as the superiormodel for identifying heart disease with an accuracy rate of 85%, and the Gradient BoostingClassifier as the most optimal for predicting heart failure risk, achieving an accuracy of 93%.These results underscore the capability of ensemble learning approaches in markedly improving diagnostic accuracy and risk assessment in the realm of cardiovascular diseases. The implementation of such predictive models in clinical settings could revolutionize the approach to cardiovascular healthcare, shifting from a reactive to a more preventive and personalized strategy. By leveraging these models, healthcare practitioners can detect at-riskpatients sooner, tailor interventions more accurately, and ultimately improve patient outcomes. Moreover, our decision to exclude the 'time' variable in the heart failure forecasting model highlights the importance of model adaptability to practical clinical applications, particularly for newly diagnosed patients. Our study adds to the expanding literature on leveraging the machine learning for healthcare, offering in-depth understanding of the use of data-driven approaches for disease prediction and management. It also demonstrates the significance of cross-disciplinary teamwork in developing tools that are notonly technically sound but also practically applicable in healthcare settings. Looking forward, several avenues for future research emerge from our study. First, exploring the integration of additional predictive variables and data sources could refine and optimise thepredictive capabilities of the systems. Second, conducting extensive validation studies acrossdiverse patient populations will be crucial in assessing the generalizability of the findings. Third, the development of accessible and user-friendly interfaces for these predictive modelscould facilitate their adoption in clinical practice, making advanced predictive analytics moreaccessible to healthcare professionals.

In conclusion, this study represents an important step forward in the employment of ML approaches for heart disease forecasting and management. By utilizing the capability of sophisticated algorithms to improve prompt identification and risk evaluation, we move closer to a future where personalized and preventive healthcare can become a reality for patients around the world. Our work lays the groundwork for future innovations in this domain aspiring to offer valuable contributions toward the progression of medical carethrough technological innovations.

REFERENCES

[1]Bhatt, C. M., et al. "Effective heart disease prediction using machine learning techniques." Algorithms, vol. 16, no. 2, 2023, Article 88.

[2]Al-Janabi, M. I., et al. "Machine learning classification techniques for heart disease prediction: A

review." Applied Science Private University.
[3]Ordonez, C. "Improving heart disease prediction using constrained association rules." Technical Seminar Presentation, University of Tokyo, 2004.
[4]David, H., & Belcy, S. A. "Heart disease prediction using data mining techniques." ICTACT Journal on Soft Computing, 2018.
[5]Shah, D., et al. "Heart disease prediction using machine learning techniques." SN Computer Science, vol. 1, 2020, p. 345.
[6]Shorewala, V. "Early detection of coronary heart disease using ensemble techniques." Informatics in Medicine Unlocked, vol. 26, 2021, Article 100655.
[7]Zhou, D., et al. "Risk prediction of heart failure in patients with ischemic heart disease using network analytics and stacking ensemble learning." BMC Medical Informatics and Decision Making, vol. 23, 2023, Article 99.
[8]Klein, L., & Gheorghiade, M. "Coronary artery disease and prevention of heart failure."Mayo Clinic Proceedings, vol. 88, no. 5, 2004, pp. 1209–1235.
[9]Abdissa, S.G. "Predictors of incident heart failure in a cohort of patients with ischemic heart disease." Pan African Medical Journal, vol. 35, 2020, pp. 1-12.
[10] Rammal, H.F., & Z.A. "Heart Failure Prediction Models using Big Data Techniques." International Journal of Advanced Computer Science and Applications, vol. 9, 2018
[11] Doppala BP, NagaMallik Raj S, Stephen Neal Joshua E, Thirupathi Rao N (2021) Automatic determination of harassment in social network using machine learning.
[12] NagaMallik Raj, S., Neeraja, S., Thirupathi Rao, N., Bhattacharyya, D. (2023). Multitask Deep Learning Model for Diagnosis and Prognosis of the COVID-19 Using Threshold-Based Segmentation with U-NET and SegNet Classifiers. In: Sisodia, D.S., Garg,L., Pachori, R.B. Tanveer, M. (eds) Machine Intelligence Techniques for Data Analysis andSignal Processing. Lecture Notes in Electricall Engineering, vol 997. Springer, Singapore.