International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. **11** Special Issue **02** | **2026**

Heart Disease Prediction Using a Hybrid Feature Selection and Ensemble Learning Approach

Ankush Mohan Rajas[1], Monali Vinod Umale[2], Bhushan Gajanan Kachare[3], Chanchal Niramal Katariya[4], Prof. Poonam B. Patthe[5]

[1,2,3,4]*Student, Department of Computer Science and Engineering, Padmashri Dr.V.B. Kolte College of Engineering, Malkapur, Maharashtra, India*

[5]*Professor, Department of Computer Science and Engineering, Padmashri Dr.V.B. Kolte College of Engineering, Malkapur, Maharashtra, India*

DOI: 10.5281/zenodo.19251949

ABSTRACT

Heart disease is one of the leading causes of death worldwide. Early prediction and diagnosis can significantly reduce mortality rates. This research proposes an intelligent heart disease prediction system that combines machine learning algorithms with a web-based platform for real-time risk assessment. In recent years, machine learning techniques have demonstrated significant potential in assisting medical diagnosis; however, their performance is highly influenced by feature selection and parameter optimization. To address these limitations, this research proposes a hybrid heart disease prediction system that integrates Genetic Algorithm (GA) and Cuckoo Search Algorithm (CSA) for optimal feature selection and model optimization. The system uses supervised learning algorithms such as Genetic Algorithm (GA) and Cuckoo Search Optimization (CSO) on a standard heart disease data set. The system analyzes multiple clinical parameters and predicts the likelihood of heart disease using supervised learning models. A user-friendly website enables patients and healthcare practitioners to interact with the system easily. Experimental evaluation demonstrates that ensemble learning models outperform traditional classifiers in terms of accuracy and reliability. The proposed system highlights the potential of machine learning as a decision-support tool in preventive healthcare. This research highlights the effectiveness of hybrid evolutionary optimization techniques in medical data analysis and paves the way for future advancements involving deep learning, explainable artificial intelligence, and real-time healthcare integration.

Keywords: - Heart Disease Prediction, Machine Learning, Healthcare Decision Support, Random Foresting, Web-Based Medical System, Predictive Analytics

I.INTRODUCTION

Heart disease encompasses a wide range of cardiovascular conditions such as coronary artery disease, heart failure, and arrhythmias [1]. Due to lifestyle changes, stress, and unhealthy habits, the prevalence of heart disease is increasing rapidly, especially in developing countries. Traditional diagnostic methods rely heavily on clinical expertise, laboratory tests, and time-consuming procedures. Recent advancements in machine learning have made it possible to extract meaningful insights from large medical data sets. Predictive models can assist in identifying high-risk individuals before severe symptoms appear [4]. This research focuses on developing a machine learning-based heart disease prediction system integrated into a web application to provide accessible, fast, and reliable risk assessment. Heart disease refers to various conditions affecting the heart, including coronary artery disease, heart attacks, and arrhythmia [3]. According to global health statistics, heart-related diseases are a major cause of death. Traditional diagnostic methods are time-consuming and require expert medical professionals [4]. With the advancement of machine learning, it is possible to analyze large medical data sets efficiently and provide early predictions. This project focuses on developing a machine learning-based heart disease prediction system integrated into a user-friendly web application.

II. LITERATURE REVIEW

In recent years, machine learning techniques have been widely applied in the healthcare domain, especially for disease prediction and diagnosis. Heart disease prediction has gained significant attention due to the increasing availability of medical data sets and advancements in computational intelligence [2]. Researchers have explored various machine learning algorithms, optimization techniques, and hybrid approaches to improve prediction

accuracy and reliability.

A. Machine Learning Algorithms

Several studies have demonstrated the effectiveness of supervised machine learning algorithms in predicting heart disease. Logistic Regression is one of the most commonly used techniques due to its simplicity and interpreting ability. It helps in understanding the relationship between medical features and disease outcome. However, its performance is limited when handling complex and nonlinear data patterns. Decision Tree algorithms provide rule-based classification and are easy to interpret, but they often suffer from over fitting. Random Foresting, an ensemble learning method, improves prediction accuracy by combining multiple decision Tree and reducing variance [4].

1. Genetic Algorithm (GA)

Genetic Algorithm is an evolutionary optimization technique inspired by the process of natural selection and biological evolution. GA has been widely used in medical data analysis for feature selection and parameter optimization [7]. In heart disease prediction systems, GA helps in identifying the most relevant clinical attributes by eliminating redundant and irrelevant features [6]. The algorithm starts with an initial population of possible solutions, represented as chromosomes. Through genetic operations such as selection, crossover, and mutation, GA evolves the population toward better solutions. Studies have shown that GA-based feature selection improves classification accuracy and reduces computational complexity [7]. However, GA may sometimes converge slowly and get trapped in suboptimal solutions if not properly tuned.

2. Cuckoo Search Algorithm (CSA)

Cuckoo Search Algorithm is a nature-inspired meta heuristic optimization technique based on the brood parasitism behavior of cuckoo birds [9]. CSA uses Lévy flights to explore the search space efficiently, making it effective in avoiding local optima. In medical prediction systems, CSA has been applied for feature optimization and hyper parameter tuning. Compared to traditional optimization methods, CSA provides faster convergence and better exploration capabilities [9]. Researchers have reported that CSA improves model performance by refining feature subsets and enhancing generalization. However, CSA alone may lack global exploration strength when dealing with highly complex data sets.

3. Hybrid Optimization Approaches

Recent literature emphasizes hybrid approaches that combine GA and CSA to balance global and local search capabilities [7]. GA efficiently explores the global search space, while CSA fine-tunes solutions through local optimization. Such hybrid models have demonstrated superior performance in heart disease prediction by improving accuracy, robustness, and stability [10]. These studies highlight the importance of combining machine learning algorithms with evolutionary optimization techniques to build effective and reliable healthcare prediction systems.

III. PROPOSED STUDY

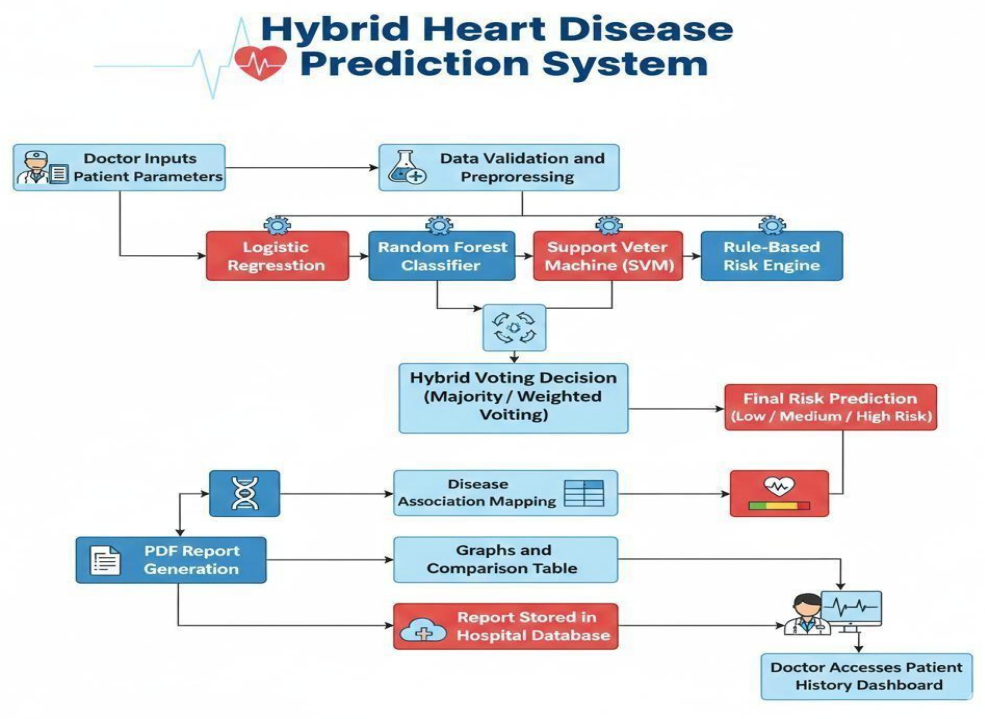

Fig. 1. System architecture

The proposed study presents a hybrid machine learning–based heart disease prediction system that integrates Genetic Algorithm (GA) and Cuckoo Search Algorithm (CSA) to improve prediction accuracy, robustness, and computational efficiency [6]. The primary objective of this study is to develop an intelligent decision-support system capable of identifying individuals at risk of heart disease at an early stage by optimizing feature selection and classification performance. The proposed system addresses the limitations of traditional machine learning models, which often suffer from reduced accuracy due to irrelevant or redundant medical features.[6] To overcome this challenge, the study introduces a hybrid optimization framework where GA and CSA are combined to exploit their complementary strengths [7]. GA is used for global search and exploration of optimal feature subsets, while CSA is applied for local refinement and fine-tuning of selected features and model parameters. Initially, a clinical heart disease data set is collected from a benchmark medical repository. The data set includes multiple patient attributes such as age, gender, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, electrocardiographic results, and maximum heart rate. Preprocessing techniques such as missing value handling, normalization, and categorical encoding are applied to ensure data consistency and quality.

A. Methodology

1. Data sets

The data set used in this research is obtained from a publicly available and widely recognized heart disease repository. It consists of clinical records collected from patients undergoing cardiac evaluation. Each record includes medical attributes such as age, gender, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, electrocardiographic results, maximum heart rate achieved, exercise-induced angina, and ST depression. These parameters are medically significant and commonly used by healthcare professionals in diagnosing cardiovascular conditions. The data set includes labeled outcomes indicating the presence or absence of heart disease, making it suitable for supervised learning.

2. Data Preprocessing

Data preprocessing is a critical stage to enhancing the quality and reliability of the data set. Initially, missing values are identified and handled using statistical imputation techniques such as mean or median substitution, depending on the attribute distribution. Categorical variables, including chest pain type and ECG results, are transformed into numerical form using encoding methods. Feature scaling and normalization are applied to ensure that all attributes contribute equally during model training. Additionally, outliers and noisy data are analyzed to reduce their impact on classification performance. This preprocessing step ensures that the data set is clean, consistent, and suitable for machine learning analysis.

3. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is performed to understand the distribution, correlation, and importance of various clinical features. Statistical measures and visualization techniques are used to identify relationships between attributes and the target variable. EDA helps in gaining insights into the data set and supports informed decision-making during feature selection and optimization.

4. Feature Selection

Medical data sets often contain redundant or weakly relevant attributes that can negatively affect model performance. Feature selection is employed to identify the most informative attributes that significantly contribute to heart disease prediction. Reducing the number of features decreases computational complexity and improves generalization. In this study, feature selection is integrated into the optimization phase using evolutionary algorithms to ensure optimal feature subsets.

5. Hybrid Optimization Using Genetic Algorithm and Cuckoo Search Algorithm

The core contribution of this methodology is the hybrid optimization framework that combines Genetic Algorithm (GA) and Cuckoo Search Algorithm (CSA). GA is first applied for global search and exploration of the feature space. In GA, each chromosome represents a candidate solution consisting of a subset of features. A fitness function based on classification accuracy is used to evaluate the quality of each solution. Genetic operations such as selection, crossover, and mutation are performed to evolve the population toward optimal solutions.

Although GA is effective in global exploration, it may suffer from slow convergence or premature stagnation. To overcome this limitation, CSA is integrated as a local optimization technique. CSA is inspired by the brood parasitism behaviour of cuckoo birds and uses Lévy flights to perform efficient local search. CSA refines the solutions obtained from GA by exploring nearby feature subsets and improving convergence speed. The hybrid GA–CSA framework balances exploration and exploitation, resulting in a compact and highly informative feature set.

6. Machine Learning Model Training

After feature optimization, the selected feature subset is used to train supervised machine learning classifiers. Logistic Regression is employed for its simplicity and interpreting ability, allowing insights into feature contributions. Random Foresting is used to capture nonlinear relationships and interactions among features through ensemble learning. The models are trained using optimized data and validated using k-fold cross-validation to minimize over fitting and ensure stability.

7. Model Evaluation and Validation

The performance of the trained models is evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score. Comparative analysis is conducted between models trained with and without hybrid optimization to demonstrate the effectiveness of the proposed approach. The hybrid model consistently shows improved predictive performance and robustness.

8. System Deployment

Finally, the optimized prediction model is deployed within a web-based application. The system allows users, such as healthcare professionals, to input patient clinical parameters and obtain real-time heart disease risk assessment. The web interface enhancings accessibility and usability, making the system suitable for practical healthcare decision support.

IV.RESULT ANALYSIS

This section presents a detailed analysis of the experimental results obtained from the proposed hybrid machine learning–based heart disease prediction system [8]. The primary objective of the result analysis is to evaluate the effectiveness of the hybrid Genetic Algorithm (GA) and Cuckoo Search Algorithm (CSA) optimization approach in improving prediction accuracy and model performance compared to conventional machine learning methods [8]. Feature importance analysis was performed using the Extra Tree classifier after GA–CSO optimization. The results indicate that chest pain type (cp), exercise-induced angina (exang), number of major vessels (ca), and thalassemia (thal) are the most influential features in heart disease prediction. Moderate importance is observed for ST depression (oldpeak), slope of ST segment, and maximum heart rate achieved (thalach). These findings are clinically relevant and align with established cardiovascular risk factors. The hybrid Genetic Algorithm (GA) and Cuckoo Search Optimization (CSO) significantly enhancing model performance by selecting a compact and informative feature subset. GA provides global exploration of the feature space, while CSO refines the solution through local optimization. This hybrid optimization improves classification accuracy, reduces computational complexity, and enhancings generalization capability. Fig.2 illustrates the performance comparison between individual machine learning models and the proposed hybrid

approach. Logistic Regression shows the lowest accuracy, indicating limited capability in capturing complex non-linear patterns. Tree-based and kernel-based models such as Random Foresting, Extra Tree, and SVM achieve higher accuracy due to their ability to model feature interactions. Gradient Boosting records the highest individual accuracy, demonstrating strong learning capacity through sequential optimization [10]. In the given Fig.2 Hybrid model, based on ensemble voting, achieves consistently high accuracy and improved stability by combining the strengths of multiple classifiers, making it more reliable for heart disease risk prediction than any single model [10]. The hybrid model achieves the highest overall accuracy and demonstrates improved stability compared to individual classifiers. The ensemble approach effectively reduces model bias and variance by leveraging the strengths of multiple learning algorithms. The feature importance plot illustrates the relative contribution of each clinical parameter to heart disease prediction as determined by the Extra Tree Classifier. The importance score reflects how frequently and effectively a feature is used by the model to reduce classification errors across decision Tree.

ALGORTHM	ACCURACY
Logistic Regression	80%
Random Forest	92%
Support Vocter Machine (SVM)	69%
Gradient Boosting	98%
Hybird Model	97%
K-Neaterst Neithrobors (KNN)	84%
Decision Tree	87%
AdaBoost	99%
Voting Classficier	93%

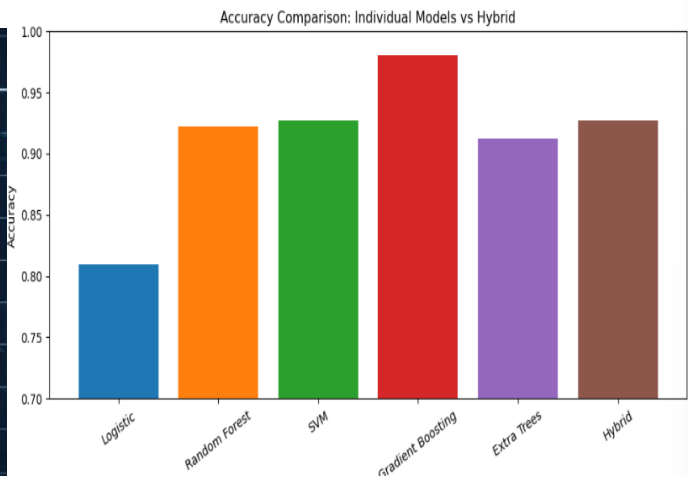

Fig.2. Accuracy Comparison:- Individual Model VS Hybrid

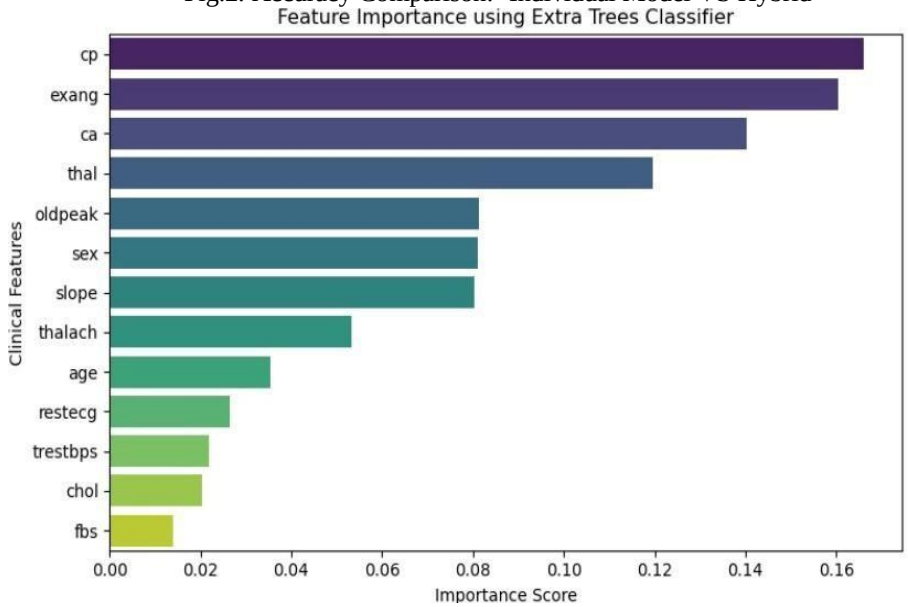

Fig. 3. Feature Importance Analysis Using Extra Tree Classifier

In the Fig. 3. results, Chest Pain Type (cp) emerges as the most influential feature, indicating that the nature of chest pain plays a critical role in identifying heart disease risk. This is followed by Exercise-Induced Angina (exang) and Number of Major Vessels (ca), both of which are strong indicators of underlying coronary artery involvement.

The Thalassemia (thal) and ST Depression (oldpeak) features also show high importance, highlighting their significance in detecting myocardial ischemia and abnormal cardiac stress responses. Moderate contributions are observed from gender, slope of the ST segment, and maximum heart rate achieved (thalach), which help refine risk assessment.The Receiver Operating Characteristic (ROC) curve illustrates the performance of the proposed hybrid heart disease prediction model by plotting the True Positive Rate (Sensitivity) against the False Positive

Rate (1 − Specificity) at various. In the Fig.4 ROC curve, the hybrid model achieves an Area Under the Curve (AUC) of 0.99, which in dicates excellent discriminative ability. An AUC value close to 1.0 signifies that the model can effectively distinguish between patients with heart disease and healthy individuals. This behavior is highly desirable in medical diagnosis systems, as it ensures early and accurate identification of high-risk patients while minimizing false alarms. The dashed diagonal line represents a random classifier with an AUC of 0.5. The significant separation between the hybrid model's ROC curve and the diagonal baseline confirms the superiority of the proposed hybrid ensemble
approach. Overall, the ROC analysis validates that the integration of GA–CSO–based feature selection with ensemble learning substantially improves predictive performance and reliability, making the model suitable for real-world clinical decision support systems.

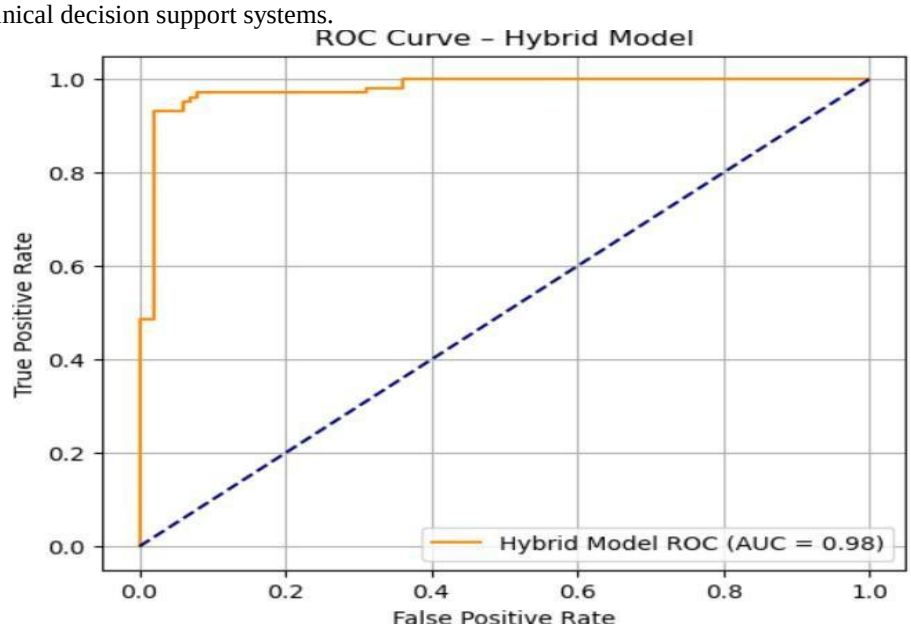

Fig.4 ROC Curve Analysis of the Hybrid Model

V.CONCLUSIONS AND FUTURE SCOPE

In this project, a heart disease prediction system using a hybrid machine learning approach has been developed. Genetic Algorithm (GA) and Cuckoo Search Algorithm (CSA) are used to select important features and improve prediction accuracy. The results show that the hybrid model performs better than traditional machine learning methods. It provides accurate and reliable prediction of heart disease and can help doctors in early diagnosis. In the future, this system can be improved by using deep learning techniques, real-time patient data, and mobile or web applications. It can also be connected with hospital systems for better healthcare support.

VI.REFERENCES

[1] J. Yang and J. Guan, ''A heart disease prediction model based on feature optimization and smote-Xgboost algorithm,'' Information, vol. 13, No. 10, p. 475, Oct. 2022.
[2] N. N. Itoo and V. K. Garg, ''Heart disease prediction using a stacked ensemble of supervised machine learning classifiers,'' in Proc. Int. Mobile Embedded Technol. Conf. (MECON), Mar. 2022, pp. 599–604.
[3] R. Sulthana, A. K. Jaithunbi, and P. Sunraja, ''Application of machine learning algorithms in predicting the heart
[4] disease in patients,'' in Proc. 3rd Int. Conf. Adv. Electr., Comput., Commun. Sustain. Technol. (ICAECT), Jan. 2023, pp. 1–4.
[5] M. J. Gaikwad, P. S. Asole, and L. S. Bitla, ''Effective study of machine learning algorithms for heart disease prediction,'' in Proc. 2nd Int. Conf. Power Electron. IoT Appl. Renew. Energy Control (PARC), Jan. 2022, pp. 1–6.
[6] H. Yang, Z. Chen, H. Yang, and M. Tian, ''Predicting coronary heart disease using an improved Light GBM model: Performance analysis and comparison,'' IEEE Access, vol. 11, pp. 23366–23380, 2023.
[7] J. Vijaya, ''Heart disease prediction using clustered genetic optimization algorithm,'' in Proc. Int. Conf. Intell. Innov. Technol. Comput., Electr. Electron. (IITCEE), Jan. 2023, pp. 1072–1077.
[8] M. T. Islam, S. R. Rafa, and M. G. Kibria, ''Early prediction of heart disease using PCA and hybrid genetic

algorithm with k-Means,'' in Proc. 23rd Int. Conf. Comput. Inf. Technol. (ICCIT), Dec. 2020, pp. 1–6.

[9] A. Abdellatif, H. Abdellatef, J. Kanesan, C. O. Chow, J. H. Chuah, and H. M. Gheni, ''An effective heart disease detection and severity level classification model using machine learning and hyper parameter optimization methods,''IEEE Access, vol. 10, pp. 79974–79985, 2022

[10]P. Nandakumar and S. Narayan, ''Cardiac disease detection using cuckoo search enabled deep belief network,''Intell. Syst. Appl., vol. 16, Nov. 2022, Art. No. 200131.

[11] G. Narasimhan and A. Victor, ''A hybrid approach with meta heuristic optimization and random foresting in improving heart disease prediction,'' Sci. Rep., vol. 15, No. 1, p. 10971, Mar. 2025.