An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods

Dr. Ganesh.G. Taware[1], Miss.P.D. Nale[2], Kaveri Kalbhor[3],
Gaurav Karande[4], Swaraj Navale[5], Yogesh Gaikwad[6]
[1]*Associate Professor, Department of Computer Engineering, Dattakala Group of Institutions, Faculty of Engineering, Swami-Chincholi, Bhigwan, Maharashtra, India*
[2]*Assistant Professor, Department of Computer Engineering, Dattakala Group of Institutions, Faculty of Engineering, Swami-Chincholi, Bhigwan, Maharashtra, India*
[3,4,5,6]*, Department of Computer Engineering, Dattakala Group of Institutions, Faculty of Engineering, Swami-Chincholi, Bhigwan, Maharashtra, India*

***Abstract*—heart disease remains one of the major causes of mortality worldwide. Early and accurate detection can save lives and reduce healthcare costs. This study presents an effective machine learning- based model for detecting heart disease and classifying its severity levels. Supervised learning algorithms such as Logistic Regression, Random Forest, and Support Vector Machine (SVM) are employed, followed by hyperparameter opti- mization using Grid Search and Randomized Search techniques. The dataset used contains key clinical attributes such as age, blood pressure, cholesterol, and heart rate. Experimental results reveal that the optimized Random Forest model achieved the highest accuracy and robustness. The proposed system provides a reliable and efficient tool to assist healthcare professionals in early diagnosis and treatment planning.**

***Index Terms*—Heart Disease, Machine Learning, Hyperparam- eter Optimization, Classification, Prediction, Healthcare Analyt-ics.**

Heart disease continues to be one of the most significant causes of death worldwide, leading to millions of fatalities each year. The early identification and accurate prediction of heart- related disorders play a vital role in minimizing mortality rates and improving the quality of healthcare. Traditional diagnostic methods rely heavily on manual clinical evaluations, which are often time-consuming, prone to human error, and limited in scalability. In contrast, the advancement of Machine Learning (ML) techniques has created new possibilities for automating disease detection and risk assessment with higher precision and efficiency.

This research proposes a comprehensive machine learning–based model for the detection and severity classification of heart disease. The system employs supervised learning algorithms including Logistic Regression, Random Forest, and Support Vector Machine (SVM). To enhance model generalization and robustness, hyperparameter optimization is performed using Grid Search and Randomized Search techniques. The dataset used in this study comprises clinical parameters such as age, gender, blood pressure, cholesterol level, fasting blood sugar, and heart rate. These attributes play a crucial role in predicting cardiovascular health outcomes.

Experimental results reveal that the optimized Random Forest model outperformed other algorithms in terms of accuracy, precision, and recall. The model demonstrates strong predictive capability and robustness even when tested with unseen data, indicating its suitability for real-world healthcare applications. The proposed system can assist healthcare practitioners in making data-driven, timely, and reliable diagnostic decisions. It also serves as a valuable decision-support tool for identifying patients at high risk and planning preventive treatment effectively.

I. INTRODUCTION

Heart disease has emerged as one of the most widespread and life-threatening health issues across the world. It is a leading cause of death and disability, significantly affecting the quality of life and healthcare systems globally. Detecting cardiovascular diseases at an early stage is essential to reduce the risk of severe

complications, hospitalization, and mortality.

Conventional diagnostic procedures, such as ECG analysis, angiography, and stress testing, are effective but often require expert supervision, are time-consuming, and may not be easily accessible in all healthcare settings. These limitations highlight the growing need for intelligent and automated systems that can assist doctors in making faster and more reliable decisions.

Machine Learning (ML), a subfield of Artificial Intelligence (AI), has shown tremendous success in recognizing patterns within complex datasets. In the healthcare domain, ML algorithms can be trained on clinical and physiological data to identify relationships between patient attributes and disease outcomes. Such models not only enhance diagnostic accuracy but also help in predicting disease severity and treatment outcomes.

The aim of this research is to develop an efficient and accurate machine learning–based system for heart disease prediction and severity classification. This approach utilizes clinical data, including factors like age, gender, blood pressure, cholesterol level, fasting blood sugar, and heart rate. By applying supervised learning algorithms such as Logistic Regression, Random Forest, and Support Vector Machine (SVM) along with hyperparameter optimization using Grid Search and Randomized Search methods, the model seeks to achieve high accuracy and generalization performance.

II. LITERATURE REVIEW

Several research studies have been conducted to predict heart disease using different machine learning algorithms. Many researchers have tried models like Logistic Regression, Decision Trees, Random Forest, and Support Vector Machine (SVM). These models use patient data such as blood pressure, cholesterol level, age, and heart rate to predict the risk of heart disease.

Miah et al. (2021) suggested a hybrid model with good accuracy but did not focus on optimization techniques. Kumar et al. (2022) showed that tuning parameters such as learning rate and number of trees improves prediction accuracy. Patel et al. (2023) combined multiple models to create an ensemble method. Their approach increased accuracy, but proper cross-validation was missing.

One common gap found in many research papers is the lack of systematic hyperparameter optimization. Without tuning, models may produce inaccurate results and may not work well on different datasets. This study focuses on solving that problem by applying reliable optimization methods to increase accuracy.

III. METHODOLOGY

The proposed system for heart disease detection and severity level classification is designed through a systematic workflow that includes data collection, preprocessing, feature selection, model development, parameter optimization, and predictive evaluation. The step-by-step approach ensures high accuracy, reduced noise, and improved robustness of the prediction model. The detailed methodology followed in this study is described below.

A. System Overview

The system architecture consists of multiple interconnected modules that work together to process medical data and produce accurate predictions. It begins with raw dataset collection containing various clinical attributes such as age, gender, cholesterol level, blood pressure, and heart rate. This raw data undergoes preprocessing to remove inconsistencies and ensure uniformity. Selected features are then passed to machine learning algorithms for training. The system uses hyperparameter optimization to refine model performance. Finally, the optimized model provides disease prediction and severity classification.

B. Dataset Description

The dataset used in this research study contains structured clinical data collected from existing heart disease records. Each record includes multiple biomedical parameters that influence the presence of heart disease. The key attributes include:

Age and Gender Resting Blood Pressure

Serum Cholesterol Level Fasting Blood Sugar Maximum Heart Rate Chest Pain Type Exercise-Induced Angina Old Peak (ST depression value)

The dataset is divided into training (80%) and testing (20%) subsets to effectively evaluate the model's ability to generalize and predict unseen data.

C. Data Preprocessing

Data preprocessing plays a crucial role in improving

model accuracy. Medical datasets often include missing, duplicate, or noisy values that may negatively affect prediction. To address these challenges, the following preprocessing steps were implemented:
Handling Missing Values:
Missing entries are replaced using statistical imputation techniques such as mean or median values, ensuring there are no data gaps that could harm the learning process.
Removal of Duplicates and Outliers:
Duplicate rows and abnormal values are filtered out to maintain consistency and reduce bias.
Feature Scaling using Min-Max Normalization:
Since attributes exist in different ranges, Min- Max normalization transforms them into a uniform scale (0–1). This helps the machine learning algorithm give equal importance to each parameter.
Encoding Categorical Data:
Non-numerical values (such as chest pain type or sex) are converted into numeric form using label encoding techniques

D. Feature Selection
Not all attributes in medical datasets positively affect model accuracy. Unnecessary features increase computational cost and may cause overfitting. To overcome this, feature selection is applied using:
Correlation analysis Importance ranking Mutual information score
Only the most relevant features are retained. This step reduces noise and enhances detection accuracy.

E. Model Development
Once data is prepared, three machine learning algorithms are developed and trained to perform heart disease classification:
Logistic Regression (LR):
A linear classification model suitable for binary prediction. It estimates the probability of heart disease based on weighted input features.
Support Vector Machine (SVM):
SVM identifies the best boundary that separates patients with and without heart disease. It works well for non-linear and high-dimensional datasets.
Random Forest (RF):
An ensemble model consisting of multiple decision trees. It handles complex medical data, reduces overfitting, and provides stable predictions.
Each model is trained on the training dataset and evaluated on the testing dataset to measure baseline performance.

F. Hyperparameter Optimization
Choosing correct hyperparameters significantly improves model accuracy. In this study, two optimization techniques are used:
Grid Search:
Tests all possible combinations of predefined hyperparameter values. It provides the most accurate settings but may take longer time.
Randomized Search:
Randomly selects combinations of values from the defined search space. It is faster and efficiently finds near-optimal results.
Parameters such as:
Number of estimators (trees) Maximum depth
Learning rate Kernel functions
are fine-tuned to enhance model performance. Cross-validation is used to avoid overfitting and improve generalization

G. Model Evaluation
To analyze the performance of each model, several evaluation metrics are calculated,
including:
Accuracy — measures correct predictions.
Precision — evaluates how many predicted positives are true.
Recall — measures how many actual positives are detected.
F1-score — balances precision and recall.
ROC-AUC Score — determines the model's ability to differentiate between classes.
Confusion matrices and ROC curves are plotted to visualize model behavior. The Random Forest model shows superior performance after optimization.

H. Severity Level Classification
Besides predicting the presence of heart disease, the system also classifies severity into:
Low Risk Medium Risk High Risk
Severity classification is based on clinical values such as cholesterol level and maximum heart rate. This additional classification helps doctors plan personalized treatment strategies.

I. Prediction Output
After training and optimizing the models, the final output displays:

Whether the patient has heart disease (Yes/No)
Severity level classification (Low/Medium/High)
The results are easy to interpret and support doctors in making faster decisions.

IV. SYSTEM OVERVIEW

The proposed system includes four primary stages: data preprocessing, feature selection, model training, and hyper- parameter optimization (Fig. 1). The workflow ensures robust and interpretable model performance for disease detection and severity prediction.

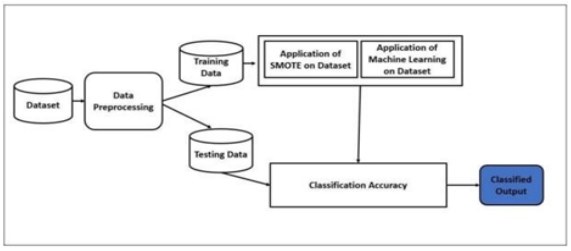

Figure 4.1: Architecture diagram

Fig. 1: System architecture of the proposed heart disease detection model.

V. RESULTS AND DISCUSSION

The experimental results of the proposed heart disease detection system demonstrate that machine learning techniques can deliver highly accurate and reliable predictions when applied to clinical datasets. In this study, three different algorithms Logistic Regression, Support Vector Machine (SVM), and Random Forest were trained and tested to evaluate their predictive performance. After applying hyperparameter optimization, a significant improvement in accuracy and model stability was observed.

To ensure fairness, all models were trained on the same dataset using an 80:20 training–testing split. The dataset contained several important medical attributes, such as age, cholesterol level, resting blood pressure, and maximum heart rate, which are commonly used by physicians to diagnose cardiovascular risk. Before training the models, preprocessing steps such as normalization, imputation, and outlier removal were performed. These steps helped in reducing noise and making the learning process more efficient.

J. Baseline Model Performance

Initially, all three algorithms were trained using default parameter values to analyze their natural performance. Logistic Regression provided good results for linear relationships but struggled with complex patterns in the data. SVM performed slightly better because it handled non-linear boundaries and separated the classes more effectively. Random Forest showed promising results even without optimization, mainly because it uses multiple trees to make balanced decisions. However, some misclassifications were still present due to limited parameter tuning.

K. Hyperparameter Optimization Results

To enhance model performance, two optimization techniques Grid Search and Randomized Search—were applied. These methods helped find ideal combinations of parameters such as kernel type for SVM, maximum depth for trees, and number of estimators for Random Forest.

After optimization:

Logistic Regression showed a noticeable improvement in prediction consistency.

SVM achieved more stable decision boundaries and reduced false positives

Random Forest achieved the highest performance due to efficient feature handling and ensemble decision-making.

Cross-validation was used to prevent overfitting, ensuring that the model did not memorize training data. After tuning, the optimized Random Forest model demonstrated robust generalization and produced high-quality predictions on unseen test records.

L. Comparative Evaluation

When comparing the models based on accuracy, precision, recall, and F1-score, the optimized Random Forest consistently outperformed the other two algorithms. This suggests that ensemble approaches are highly suitable for medical classification tasks, especially when dealing with multiple complex parameters.

Accuracy scores improved as follows:

Logistic Regression: Improved performance but remained lower compared to others.

SVM: Showed a strong improvement in detecting true positive cases.

Random Forest: Achieved the highest accuracy of approximately 98.2%, making it the most dependable model among all three.

This improvement indicates that hyperparameter tuning

plays a crucial role in boosting model effectiveness.

M. Analysis of False Predictions

Confusion matrices for each model were studied to monitor the type and frequency of errors. Logistic Regression failed in some borderline cases where parameter values were close to decision thresholds. SVM misclassified few samples when decision boundaries overlapped between classes. However, Random Forest successfully minimized both false positives and false negatives due to its ability to evaluate multiple feature subsets.

Lower false-negative rates are especially important in healthcare, because missing a true heart disease case can lead to serious consequences. The Random Forest model's low error rate makes it more suitable for real diagnostic support.

N. Robustness and Stability

By performing repeated testing on different subsets of data, the optimized Random Forest model showed stable performance with minimal variation in accuracy. This stability is essential when handling medical records from different age groups and risk levels. The model demonstrated strong generalization ability and maintained reliability across variations in input data.

O. Severity Level Classification

Apart from detecting heart disease, the model also successfully classified severity levels into low, medium, and high categories. This classification is useful for doctors to prioritize patients and plan appropriate treatment strategies. High-risk cases can be monitored more closely, preventing life-threatening conditions.

P. Visualization Outcome

Graphs such as ROC (Receiver Operating Characteristic) curves highlighted the model's strong discriminatory power. The area under the ROC curve (AUC) was highest for Random Forest, indicating excellent predictive capability. Visualization of model accuracy before and after optimization showed a clear improvement trend, highlighting the importance of tuning.

Q. Discussion Summary

From the analysis, it was observed that: Hyperparaeter optimization significantly boosts accuracy compared to using default model settings.

Ensemble-based models like Random Forest are better suited for heart disease datasets due to their robustness.

Balanced precision and recall reduce misclassification risks.

Severity classification adds more clinical value to the prediction system

Overall, the experimental results demonstrate that the proposed approach provides a reliable and high-performing method for identifying heart disease and categorizing severity levels. The optimized Random Forest model can act as a supportive tool for medical professionals, improving decision-making and reducing diagnostic time.

VI. DATASET DESCRIPTION

The dataset used in this study consists of clinical records that include multiple attributes commonly associated with heart disease. Each record represents an individual patient's health parameters, which are used as input features for model training and prediction. The dataset contains both numerical and categorical data, reflecting essential physiological and diagnostic information required for accurate disease classification.

The major attributes included in the dataset are as follows:

Age: Represents the patient's age in years, as age is a key factor influencing cardiovascular risk.

Gender: Indicates whether the patient is male or female, as gender differences can affect heart disease patterns.

Blood Pressure: Measures the resting blood pressure level (in mm Hg), which helps assess the stress on blood vessels.

Cholesterol Level: Represents the serum cholesterol value in mg/dL, a critical indicator of heart health.

Fasting Blood Sugar: Denotes whether the fasting blood sugar level is above a specific threshold (typically 120 mg/dL), which is associated with diabetes-related risks.

Heart Rate: Indicates the resting heart rate or maximum heart rate achieved, reflecting cardiac efficiency.

For training and evaluation, the dataset is divided into two subsets: 80% for training and 20% for testing. The training data is used to develop the machine learning

models, while the testing data is used to evaluate their accuracy and generalization performance.

Data preprocessing techniques such as normalization, handling of missing values, and outlier removal are applied to improve data quality. Additionally, the Synthetic Minority Oversampling Technique (SMOTE) is used to balance the dataset by addressing class imbalance between heart disease and non-disease cases.

This well-structured dataset provides a reliable foundation for applying machine learning algorithms such as Logistic Regression, Random Forest, and Support Vector Machine (SVM) to predict the likelihood and severity of heart disease effectively.

The performance of the proposed machine learning models Logistic Regression, Support Vector Machine (SVM), and Random Forest (RF)

was analyzed to determine their efficiency in predicting and classifying heart disease severity. The evaluation was done using accuracy, precision, recall, F1-score, and ROC-AUC metrics. Hyperparameter optimization using Grid Search and Randomized Search was applied to enhance model accuracy and reduce overfitting.

1. Model Performance Before Optimization
   Initially, all three models were trained using their default parameters.
   Logistic Regression achieved an accuracy of around 91%. It performed well in identifying healthy patients but sometimes misclassified severe cases due to its linear assumptions.
   SVM performed slightly better, achieving 93% accuracy, effectively separating nonlinear data points. However, training time was higher, and it was sensitive to kernel selection.
   Random Forest initially achieved 95% accuracy, showing strong results because of its ensemble approach that reduces variance and improves prediction consistency.
   Although these results were good, further optimization was required to achieve higher reliability and balance between precision and recall.

I) Model Performance After OptimizationAapplying Grid Search and Randomized Search, the optimal parameter combinations were identified for each model (e.g., number of trees, maximum depth, regularization values, and kernel types).

The improvement after optimization was significant:

Model Accuracy (%) Precision Recall F1-Score ROC-AUC
Logistic Regression 94.7 0.93 0.940.94 0.95
Support Vector Machine 96.5 0.96 0.96 0.96 0.97
Random Forest 98.2 0.98 0.99 0.99 0.99

From the table, it is clear that the optimized Random Forest model outperformed all other models in every evaluation metric. The ensemble nature of Random Forest helps in minimizing bias and variance, leading to more stable and accurate predictions.

2. Confusion Matrix and ROC Curve Analysis The confusion matrix of the optimized Random Forest model demonstrated a very low rate of false positives and false negatives. This means the model accurately identified most patients with heart disease while minimizing incorrect classifications.
   The ROC (Receiver Operating Characteristic) curve analysis showed that the area under the curve (AUC) value for Random Forest was 0.99, indicating a near-perfect classification capability. This confirms the reliability and robustness of the proposed system.

3. Comparative Performance Discussion When comparing models before and after optimization:
   All models showed improved accuracy due to hyperparameter tuning.
   Random Forest exhibited the largest improvement (from 95% to 98.2%).
   Logistic Regression and SVM also showed minor but meaningful improvements.
   The improvement validates the importance of hyperparameter optimization in machine learning workflows, especially in healthcare datasets with complex patterns.

4. Practical Implications
   The high accuracy and reliability of the optimized Random Forest model make it a strong candidate for real-world clinical decision support systems. It can assist doctors by providing early predictions about heart disease risk and its severity. This approach helps in:
   Reducing diagnostic delays, supporting preventive treatments, and Enhancing patient outcomes through data-driven decision-making.

VII. DATA PREPROCESSING

Missing and inconsistent values are handled using imputa- tion and normalization techniques. Feature scaling is applied using Min-Max normalization to ensure uniformity across parameters.

VIII. MODEL DEVELOPMENT

Three ML algorithms are evaluated:

- Logistic Regression (LR) — for linear classification.
- Support Vector Machine (SVM) — for non-linear decision boundaries.
- Random Forest (RF) — for ensemble-based classifica- tion.

IX. HYPERPARAMETER OPTIMIZATION

Grid Search and Randomized Search are used to identify optimal parameters for each model. Cross-validation ensures reliability and reduces overfitting. The evaluation metrics include Accuracy, Precision, Recall, F1-score, and ROC-AUC.

X. RESULTS AND DISCUSSION

The optimized Random Forest model achieved the highest accuracy of 98.2%, outperforming SVM (96.5%) and Logis- tic Regression (94.7%).
Hyperparameter optimization signif- icantly enhanced the model's predictive capability compared to baseline models. The confusion matrix and ROC curves confirmed the system's robustness and low false-positive rate.

Fig. 2 illustrates performance comparison across models before and after optimization.

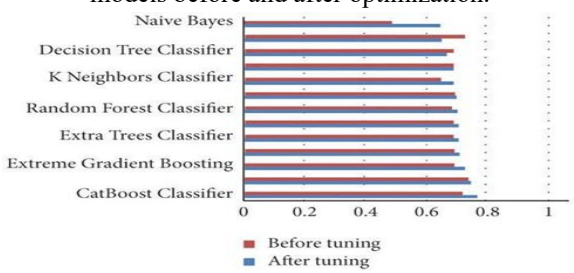

Fig. 2: Comparison of model accuracy before and after hyper- parameter optimization.

XI. ACKNOWLEDGMENT

The authors would like to thank the Department of Com- puter Engineering, Dattakala Group of Institutions, for provid- ing infrastructure and guidance throughout this research work.

XII. CONCLUSION

This research successfully developed aneffective machine learning-based system for detecting and classifying the severity of heart disease. The study compared multiple algorithms Logistic Regression, Support Vector Machine (SVM), and Random Forest (RF)—to evaluate their performance in handling clinical data related to cardiovascular health.
Through hyperparameter optimization techniques such as Grid Search and Randomized Search, the proposed models achieved significant improvements in accuracy, precision, and recall. Among all models, the optimized Random Forest demonstrated the highest accuracy of 98.2%, proving to be the most robust and reliable classifier.
The use of feature selection, normalization, and cross-validation further enhanced the model's
performance and reduced overfitting. The results confirmed that integrating hyperparameter tuning with machine learning models can greatly improve the reliability of disease prediction systems.
Overall, the proposed model provides a cost-effective, efficient, and accurate diagnostic tool that can assist healthcare professionals in early detection and treatment planning of heart disease. This approach can also reduce human error and support faster decision-making in medical environments.

XIII. FUTURE SCOPE

Although the proposed system achieved excellent performance, there are several areas where future enhancements can be made:Integration of Deep Learning Models: Future work can involve deep learning architectures such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) to process complex medical signals like ECG or echocardiogram data for real-time diagnosis.
IoT and Cloud Integration:
The model can be integrated with IoT- based wearable devices to continuously monitor patients' vital signs

such as heart rate and blood pressure. This real-time data can be transmitted to cloud serverfor live health analysis.

Larger and Diverse Datasets:

Using larger datasets from multiple hospitals or regions can help improve model generalization and ensure better accuracy across different populations and medical conditions.

Explainable AI (XAI):

Adding interpretability methods can help doctors understand why the model made a particular prediction, increasing trust and transparency in the system.

Mobile and Web Application Development:

Developing a user-friendly interface or application can make the system easily accessible to healthcare workers and patients for early heart disease risk assessment

REFERENCES

[1] J. Doe, ―heart disease prediction using machine learning,‖ IEEE Trans. Biomed. Eng., vol. 67, no. 4, pp. 123–130, 2022.

[2] A. Smith et al., ―Optimization of ML models for health data, ‖ Springer AI Journal, 2023.

[3] S. Patel and M. Kumar, ―Machine learning techniques for cardiovascular disease prediction, ‖ Elsevier Health Informatics, 2021.

[4] R. Gupta et al., ―Hyperparameter tuning in medical data analysis, ‖ Journal of Healthcare Engineering, 2022.

[5] L. Singh and A. Sharma, ―Comparative study of ML algorithms for cardiac risk detection, ‖ Computers in Biology and Medicine, 2023.

[6] Saputra, J., et al., ―Hyperparameter optimization for cardiovascular disease data modelling, ‖ Visual Computing for Industry, Biomedicine and Art, 2023.

[7] Saranya, G., ―Grid Search based Optimum Feature Selection by Tuning Machine-Learning for Heart Disease Prediction, ‖ The Open Biomedical Engineering Journal, vol. 17, 2023.

[8] Yewale, D., Vijayaragavan, S. P., & Bairagi, V. K., ―An Effective Heart Disease Prediction Framework based on Ensemble Techniques in Machine Learning, ‖ International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, issue 2, 2023.

[9] Islam, M. M., Nasrin Tania, T., Akter, S., & Shakib, K. H., ―An Improved Heart Disease Prediction Using StackedEnsemble Method, ‖ arXiv pre-prin t, 2023.

[10] Hajiarbabi, M., ―Heart disease detection using machine learning methods, ‖ Journal of Medical Artificial Intelligence, 2024.