Adaptive Ensemble Feature Selection and Genetic Algorithm-Tuned Ensemble Model for Robust Heart Disease Prediction

## Abstract

## Keywords

## Introduction

Cardiovascular diseases (CVDs), particularly heart disease, are among the leading causes of mortality worldwide, accounting for an estimated 17.9 million deaths annually, according to the World Health Organization. Early detection and prevention of heart disease are essential to reducing the global burden of this condition. However, timely diagnosis remains a significant challenge due to the complex interplay of risk factors, including age, lifestyle, genetic predisposition, and physiological metrics. With the increasing availability of medical datasets and the rise of artificial intelligence (AI), machine learning (ML)-based models offer promising solutions for accurate and early heart disease prediction (Maini et al., 2019).

A key factor influencing the performance of ML models is the quality of the input features. Many publicly available medical datasets contain redundant, irrelevant, or noisy features, which can degrade prediction accuracy and increase computational costs. Feature selection (FS), therefore, plays a crucial role in identifying the most informative features that contribute to the prediction task. Traditional FS methods—whether filter-based, wrapper-based, or embedded—offer varying advantages, but relying on a single technique often leads to suboptimal results. Ensemble Feature Selection (EFS) methods, which combine multiple FS approaches, have emerged as a powerful solution to enhance robustness and generalizability (Noroozi et al., 2023).

Another critical challenge in heart disease prediction is hyperparameter tuning of classification models. Manual tuning or grid search methods are computationally intensive and may fail to explore the optimal parameter space. In contrast, Genetic Algorithms (GAs), inspired by evolutionary processes, offer an efficient approach for optimizing hyperparameters in high-dimensional spaces. Studies have shown that GA-based optimization significantly improves the performance of classifiers like Random Forest, Logistic Regression, and Gradient Boosting in medical applications (Aleem et al., 2022), (Yewale & Vijayaragavan, 2024).

Furthermore, real-world medical data often suffer from class imbalance, where the number of healthy samples vastly outnumbers diseased cases. This imbalance can skew model training, leading to poor recall or sensitivity in detecting heart disease. Adaptive resampling techniques like SMOTE, Tomek Links, and ADASYN have proven effective in addressing this issue (Baviskar et al., 2021).

In this research, we propose a comprehensive framework for heart disease prediction that combines adaptive ensemble feature selection, genetic algorithm-based model tuning, ensemble classification, and explainability analysis. The framework is tested across multiple global datasets, including those from the UCI and Kaggle repositories, to ensure model robustness, interpretability, and cross-dataset generalizability. Additionally, the model is evaluated under noise conditions to simulate real-world sensor inaccuracies and to assess robustness (Brahimi & Feradji, 2024).

Key contributions of this study include:

Integration of multiple global heart disease datasets with harmonized features.

A robust ensemble-based feature selection strategy combining filter, wrapper, and embedded methods.

GA-driven hyperparameter tuning of diverse ensemble classifiers.

Adaptive resampling techniques to handle varying levels of class imbalance.

A final ensemble model using soft voting and stacking strategies for improved prediction performance.

SHAP-based interpretability and robustness evaluation through noise injection.

Cross-dataset generalization analysis to validate model transferability.

Through this comprehensive approach, the study aims to develop a scalable, explainable, and high-performance predictive model that can assist clinicians in early and accurate detection of heart disease, ultimately contributing to better preventive healthcare.

## Related Work

Numerous studies have explored the application of feature selection and genetic algorithms to improve heart disease prediction accuracy. These works collectively underscore the importance of selecting optimal features and tuning classifiers to enhance diagnostic performance.

(Aleem et al., 2022) demonstrated the effectiveness of evolutionary algorithms—specifically Genetic Algorithm (GA) and Particle Swarm Optimization (PSO)for feature selection, showing that GA yielded the best improvement in classification accuracy when combined with classifiers such as Naïve Bayes, SVM, and Random Forest.

Baviskar et al. (2021) utilized GA and PSO in conjunction with deep learning models (RNN, LSTM) to enhance heart disease prediction. Their results confirmed that feature optimization significantly improves classification metrics, with LSTM+PSO achieving 93.5% accuracy (Baviskar et al., 2021)

Yewale and Vijayaragavan (2024) proposed a GA-driven feature optimization approach and achieved up to 96.72% accuracy using models like Logistic Regression and Gradient Boosting, reinforcing the value of metaheuristic techniques in medical prediction systems (Yewale & Vijayaragavan, 2024)

Similarly, Brahimi and Feradji (2024) introduced the Rival Genetic Algorithm (RGA), which enhanced prediction performance by introducing competitive selection and dynamic mutation strategies. Their model achieved over 94% accuracy on standard heart disease datasets (Brahimi et al., 2024).

Noroozi et al. (2023) evaluated sixteen feature selection techniques, concluding that both filter and evolutionary methods could significantly improve predictive accuracy—although performance varied depending on classifier type and dataset characteristics (Noroozi et al., 2023).

El-Shafiey et al. (2021) applied a hybrid RF-GA feature selection approach and reported a classification accuracy of 95.6%, validating the utility of GA in selecting the most informative attributes(El-Shafiey et al., 2021).

Although these studies have made valuable contributions, they typically focus on either individual feature selection methods or singular classification techniques. In contrast, our proposed methodology integrates a diverse ensemble of feature selection methods (filter, embedded, and wrapper), addresses class imbalance adaptively, and applies a Genetic Algorithm for tuning multiple ensemble classifiers. Furthermore, our approach includes cross-dataset validation, noise robustness testing, and SHAP-based explain ability offering a comprehensive framework for heart disease prediction.

## Methodology

## Results

## Conclusion

## References