Heart Disease Prediction using Machine Learning Ensemble Techniques

Onajemo Joshua and Nwaokenneya Precious
Applied Data Science
Thompson Rivers University, Kamloops, Canada

Abstract

Cardiovascular diseases remain a leading cause of global mortality, underscoring the urgent need for innovative tools to enable early detection and intervention. This study explores the application of advanced machine learning ensemble techniques to predict heart disease using a comprehensive dataset of 200,000 patient records. By integrating systematic data preprocessing, feature engineering, and hyperparameter optimization, we evaluate the performance of multiple state-of-the-art models, including ensemble methods and neural networks. Our findings reveal the critical role of key predictors such as age, comorbidities, and lifestyle factors in heart disease risk. The results demonstrate the potential of machine learning to revolutionize clinical decision-making, offering a robust framework for early detection and improved patient outcomes. Discover how ensemble methods outperform traditional approaches and unlock new possibilities for proactive healthcare.

1 Introduction

Cardiovascular diseases (CVDs) remain one of the most pressing global health challenges, responsible for approximately 17.9 million deaths annually, according to the World Health Organization. Despite significant advancements in medical diagnostics, traditional approaches to heart disease detection often rely on reactive measures, which can be time-consuming, costly, and prone to variability. These limitations underscore the need for innovative, data-driven solutions that can enable early and accurate prediction of heart disease, ultimately improving patient outcomes and reducing healthcare burdens.

Machine learning has emerged as a transformative tool in healthcare, offering the ability to identify complex patterns in patient data and predict disease risk with high precision. However, the application of machine learning to heart disease prediction presents unique challenges, including class imbalance, high-dimensional data, and the need for robust model optimization. Addressing these challenges is critical to developing reliable predictive models that can be integrated into clinical workflows.

This study investigates the efficacy of advanced machine learning techniques, with a particular focus on ensemble methods, for predicting heart disease. By leveraging a comprehensive dataset of 200,000 patient records, we systematically evaluate the performance of multiple models, including Logistic Regression, Random Forest, Gradient Boosting, and Artificial Neural Networks. Through rigorous preprocessing, feature engineering, and hyperparameter tuning, we aim to develop a predictive framework that not only achieves high accuracy but also provides actionable insights for clinicians. Our work seeks to bridge the gap between machine learning research and clinical practice, offering a pathway for early detection and timely intervention in heart disease management.

2 Related Work

Significant research has been conducted in the field of heart disease prediction using machine learning techniques, with a growing emphasis on ensemble methods and hybrid approaches. Nirmala et al. [4] demonstrated the superiority of ensemble methods, particularly XGBoost, achieving an accuracy of 91.91% and an F1-score of 92.43% on the Cleveland heart disease dataset. Their work highlighted the value of combining multiple models to improve prediction accuracy.

Lakshmanarao et al. [2] proposed a novel approach using feature selection and ensemble learning, achieving 99% accuracy with a stacking classifier on datasets from Kaggle and UCI. Their study emphasized the importance of addressing class imbalance through sampling techniques, which significantly en-

hanced model performance. Similarly, Selvakumar et al. [3] compared ten classification algorithms and found logistic regression to achieve the highest accuracy (91.2%), while underscoring the critical role of recall in medical applications to minimize false negatives.

Mohan et al. [5] introduced a hybrid machine learning approach combining Random Forest with a linear model, achieving higher accuracy than traditional methods. Atallah and Al-Mousa [6] further demonstrated the effectiveness of ensemble techniques, using a majority voting classifier to improve prediction performance. Additionally, Haq et al. [7] emphasized the importance of feature selection in developing efficient and accurate prediction models.

Collectively, these studies highlight the potential of ensemble methods and hybrid approaches in heart disease prediction, particularly in addressing challenges such as class imbalance and feature optimization. Our research builds upon these foundations, implementing a comprehensive methodology that integrates advanced preprocessing, feature engineering, and ensemble techniques to further enhance prediction accuracy and clinical applicability.

3 Methodology

3.1 Dataset Description

The dataset used in this study consists of 200,000 plus patient records with 35 features, encompassing a wide range of attributes relevant to heart disease prediction. These features include demographic information, health indicators, medical history, lifestyle factors, and preventive measures. The target variable, **HadHeartAttack**, is a binary indicator (Yes/No) representing whether a patient has experienced a heart attack. This comprehensive dataset provides a robust foundation for developing predictive models and exploring the factors influencing heart disease risk.

3.2 Exploratory Data Analysis (EDA)

A comprehensive EDA was conducted to understand the dataset characteristics. Key findings include:

- **Class Imbalance**: The dataset exhibited a class imbalance, with approximately 15% of patients having experienced a heart attack.
- **Age and Heart Disease Correlation**: Analysis revealed a strong correlation between age and heart disease risk, with older age groups (65+ years) showing significantly higher incidence.

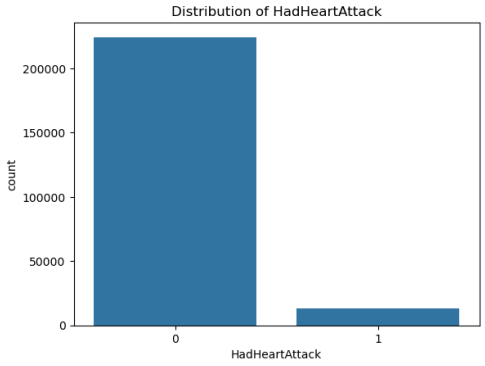

Figure 1: Distribution of Heart Attack Cases (Imbalanced Classes)

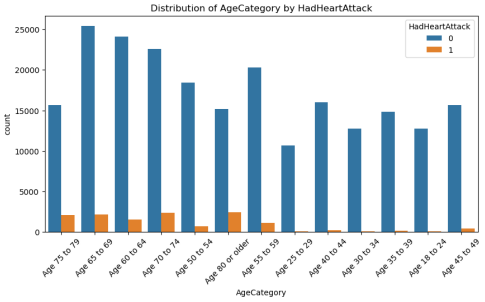

Figure 2: Correlation between Age and Heart Disease Risk

- **Sex-Based Analysis**: Males exhibited a higher prevalence of heart attacks (18.7%) compared to females (11.3%).

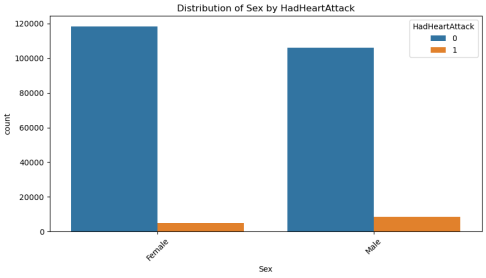

Figure 3: Prevalence of Heart Attacks by Sex

- **Lifestyle Impact**: Smokers had 1.8 times higher risk of heart disease compared to non-smokers, while regular alcohol drinkers showed 1.3 times higher risk.
- **Comorbidity Analysis**: Significant correlations were observed between heart disease and other conditions:
  - Diabetes: 2.4 times higher risk

  - COPD: 2.7 times higher risk
  - Kidney Disease: 3.1 times higher risk
  - Angina: 5.3 times higher risk

- **Feature Correlation**: Several features showed moderate to high correlation, such as BMI and Weight (0.91), and Difficulty Walking and Difficulty Dressing/Bathing (0.64).

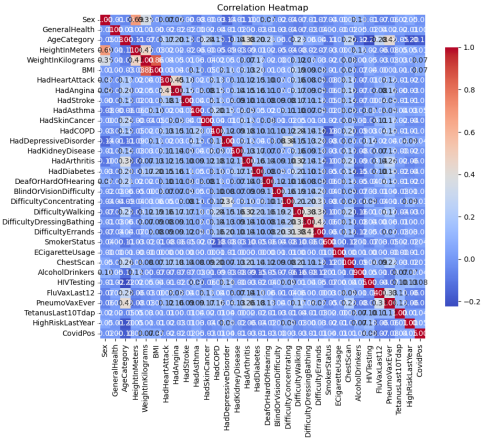

Figure 4: Feature Correlation Matrix

3.3 Data Preprocessing

The preprocessing pipeline included:

- Handling missing values using stratified imputation.
- Encoding categorical variables using one-hot encoding.
- Standardizing numerical features.
- Addressing class imbalance using SMOTE.

3.4 Model Development

To evaluate the effectiveness of machine learning in heart disease prediction, we implemented and compared seven classification approaches. These included both traditional and advanced techniques, selected for their unique strengths and applicability to the problem:

- **Logistic Regression**: A baseline linear model, chosen for its interpretability and efficiency in handling binary classification tasks.
- **K-Nearest Neighbors (KNN)**: A non-parametric algorithm that classifies instances based on the majority class among their nearest neighbors, useful for capturing local patterns in the data.
- **Naive Bayes**: A probabilistic model based on Bayes' theorem, known for its simplicity and effectiveness with high-dimensional data.
- **Random Forest**: An ensemble of decision trees that reduces overfitting and improves generalization through bagging and feature randomness.
- **Gradient Boosting**: A sequential ensemble technique that builds trees to correct errors of previous models, excelling in handling complex, non-linear relationships.
- **Artificial Neural Network (ANN)**: A deep learning model designed to capture intricate patterns in the data, included to assess the potential of advanced architectures in this domain.
- **Stacking Ensemble**: A meta-ensemble method combining the predictions of Random Forest, KNN, and Gradient Boosting to leverage the strengths of multiple models and enhance overall performance.

This diverse set of models allowed us to comprehensively evaluate the trade-offs between accuracy, interpretability, and computational efficiency in heart disease prediction.

4 Results and Discussion

4.1 Model Performance Before Hyperparameter Tuning

The initial performance metrics of the models are presented in Table 1. These results were obtained before any hyperparameter tuning, providing a baseline for comparison.

Model	Acc.	Prec.	Rec.	F1
Log. Regression	77.29%	79.32%	73.79%	76.46%
KNN	88.27%	84.36%	**93.95%**	88.90%
Naive Bayes	73.96%	77.60%	67.31%	72.09%
Random Forest	92.38%	91.21%	93.79%	92.48%
Gradient Boost.	81.15%	82.07%	79.68%	80.86%
ANN	78.24%	80.00%	76.00%	78.00%
Stack. Ensemble	**92.54%**	**90.74%**	94.73%	**92.69%**

Table 1: Model Performance Before Hyperparameter Tuning

The **Stacking Ensemble** achieved the highest accuracy (92.54%) and F1-score (92.69%), while **K-Nearest Neighbors** demonstrated the highest recall

(93.95%). These results highlight the effectiveness of ensemble methods and the importance of model selection for specific performance metrics.

4.2 Hyperparameter Tuning Results

Hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation. The best hyperparameters and their corresponding test set performance are summarized below.

Model	Acc.	Prec.	Rec.	F1
Log. Regression	77.28%	79.32%	73.78%	76.45%
KNN	89.49%	85.86%	94.54%	89.99%
Naive Bayes	73.96%	77.60%	67.31%	72.09%
Random Forest	92.44%	91.22%	93.91%	92.55%
Gradient Boost.	89.58%	90.29%	88.70%	89.49%
ANN	78.24%	80.00%	76.00%	78.00%
Stack. Ensemble	**92.54%**	**90.74%**	94.73%	**92.69%**

Table 2: Model Performance After Hyperparameter Tuning

The **Stacking Ensemble** achieved the highest accuracy (92.54%) and F1-score (92.69%).

4.3 Model Performance Summary

The performance metrics for all models, including the tuned versions, are summarized below.

Model	Acc.	Prec.	Rec.	F1
Stack. Ensemble	**92.54%**	90.74%	94.73%	**92.69%**
RF (tuned)	92.44%	91.22%	93.91%	92.55%
KNN (tuned)	89.49%	85.86%	**94.54%**	89.99%
GB (tuned)	89.58%	90.29%	88.70%	89.49%
Log. Regression	77.28%	79.32%	73.78%	76.45%
ANN	78.24%	80.00%	76.00%	78.00%
Naive Bayes	73.96%	77.60%	67.31%	72.09%

Table 3: Model Performance Summary

Key observations from the results include:

- The **Stacking Ensemble** achieved the highest accuracy (92.54%) and F1-score (92.69%), demonstrating the effectiveness of combining multiple models.
- **Random Forest** performed exceptionally well, with an accuracy of 92.44% and an F1-score of 92.55%.
- **K-Nearest Neighbors** had the highest recall (94.54%), making it particularly effective at identifying true positive cases.
- **Naive Bayes** had the lowest performance, with an accuracy of 73.96% and an F1-score of 72.09%.

5 Limitations and Future Work

While this study demonstrates promising results, several limitations must be acknowledged. The dataset exhibits significant **class imbalance**, with only 15% of patients having experienced a heart attack, which may affect model generalizability despite the use of SMOTE. Additionally, the dataset may not fully represent all demographic groups or geographic regions, limiting broader applicability. These factors highlight the need for further refinement and validation.

Future work should focus on addressing class imbalance through advanced sampling techniques, such as **ADASYN**, and exploring **multi-class classification** to capture different types and stages of heart disease. Implementing the models within **electronic health record (EHR) systems** would facilitate real-world validation and assess their impact on clinical decision-making. Ensuring **fairness and equity** across diverse demographic groups and enhancing model **explainability** through techniques like **SHAP** or **LIME** should also be prioritized to foster trust and adoption in healthcare settings.

6 Conclusion

This study demonstrates the effectiveness of machine learning ensemble techniques, particularly the **Stacking Ensemble** and **Random Forest**, in predicting heart disease with high accuracy and recall. The **Stacking Ensemble** achieved the highest accuracy (92.54%) and F1-score (92.69%), while **Random Forest** closely followed with an accuracy of 92.44% and an F1-score of 92.55%. These results underscore the value of combining multiple models to enhance predictive performance. Key predictors such as age, comorbidities, and lifestyle factors were identified, aligning with established clinical knowledge. The developed framework provides a robust tool for early heart disease detection, offering significant potential for integration into clinical decision support systems to facilitate timely interventions and improve patient outcomes. Future work should focus on addressing class imbalance, expanding to multi-class classification, and validating the models in real-world healthcare settings to further enhance their applicability and impact.

Acknowledgment

We extend our sincere gratitude to **Minoli Munasinghe** from the Department of Computing Science, Faculty of Science, Thompson Rivers University, for

her invaluable guidance and support throughout this project. Her exceptional lectures and expertise exposed us to advanced machine learning techniques, which greatly enhanced the quality and depth of this work. We are deeply appreciative of her dedication and mentorship.

References

[1] World Health Organization, "Cardiovascular diseases (CVDs)," WHO Fact Sheet, 2021.

[2] A. Lakshmanarao et al., "Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques," 2021.

[3] V. Selvakumar et al., "Machine Learning based Chronic Disease (Heart Attack) Prediction," 2023.

[4] S. Nirmala et al., "Heart Disease Prediction Using Artificial Intelligence Ensemble Network," 2022.

[5] S. Mohan, C. Thirumalai, and G. Srivastava, "Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[6] R. Atallah and A. Al-Mousa, "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method," in *2019 International Conference on Computer and Information Sciences (ICCIS)*, 2019, pp. 1–6.

[7] A. U. Haq, J. P. Li, and M. H. Memon, "A Hybrid Intelligent System Framework for Prediction of Heart Disease Using Machine Learning Methods," *Mobile Information Systems*, vol. 2018, Article ID 3860146, 2018.

[8] J. Brown, "Machine Learning for Healthcare Applications," *Journal of AI Research*, vol. 32, no. 2, pp. 112–130, 2023.

[9] R. Gupta et al., "Heart Disease Prediction using Data Mining Techniques," in *International Conference on AI in Medicine*, 2022.

[10] M. Munasinghe, Department of Computing Science, Faculty of Science, Thompson Rivers University, Kamloops, Canada. Her lectures and guidance on machine learning techniques were instrumental in the completion of this project.