Improving Heart Attack Prediction with Feature Selection and Stacking Ensemble Method

Sachin Girawale, Janhavi Kahare, Aneesh Kanhere

Under the Guidance of Dr. K . Rajeshwari

Pimpri Chinchwad College of Engineering, Pune -411044, India

Abstract: Heart attacks, or myocardial infarctions, remain a significant global health concern, contributing to substantial morbidity and mortality. Early detection and prevention are paramount in reducing the burden of cardiovascular diseases, necessitating the development of effective predictive models. This paper presents a comprehensive exploration of machine learning techniques for heart attack prediction, aiming to enhance preventive healthcare strategies. Leveraging a diverse range of classifiers, including logistic regression, decision trees, random forests, support vector machines and gradient boosting, we evaluate their performance in predicting heart attack occurrences. Through rigorous analysis and comparison, we highlight the potential of machine learning in accurately assessing an individual's likelihood of experiencing a heart attack based on demographic, lifestyle, and health-related factors. Additionally, we discuss the contribution of heart attack prediction to healthcare analytics, emphasizing its role in proactive management, personalized medicine, resource optimization, patient empowerment, and ongoing research efforts in cardiovascular medicine. Our findings underscore the significance of machine learning in shifting healthcare delivery from reactive to proactive, ultimately leading to improved health outcomes and resource utilization.

Keywords: Heart attack prediction, machine learning, preventive healthcare, cardiovascular disease, healthcare analytics

1. INTRODUCTION

Heart attacks, or myocardial infarctions, represent acute and often life-threatening events within the spectrum of cardiovascular diseases. Globally, heart attacks stand as a leading cause of mortality, underscoring the critical need for accurate risk prediction and early intervention strategies [1]. Despite advancements in medical science, heart attacks continue to pose significant challenges, necessitating innovative approaches for timely detection and preventive measures.

Coronary artery disease, characterized by the narrowing or blockage of coronary arteries, serves as the primary precursor to heart attacks, making it imperative to identify individuals at heightened risk [2]. The ability to predict an individual's susceptibility to a heart attack not only facilitates proactive management but also offers opportunities to mitigate the disease's devastating consequences.

In recent years, the integration of machine learning techniques has emerged as a promising avenue for enhancing heart attack risk prediction. Machine learning algorithms, fueled by vast datasets and computational power, hold the potential to discern complex patterns and relationships within heterogeneous patient data. Leveraging this technology, researchers endeavor to develop robust predictive models capable of accurately assessing an individual's likelihood of experiencing a heart attack.

## In this context, the motivation to explore heart attack prediction transcends mere academic inquiry. It represents a proactive stance in combating a formidable public health challenge, aiming to empower individuals, healthcare providers, and policymakers with actionable insights for preventive healthcare.

By harnessing the predictive power of machine learning, we aspire to shift the paradigm from reactive to proactive healthcare delivery, ultimately saving lives and alleviating the burden on healthcare systems.

This study embarks on a journey to explore the efficacy of machine learning models in predicting heart attacks, leveraging comprehensive patient data encompassing demographic, lifestyle, and health-related factors. Through rigorous analysis and evaluation, we seek to delineate the landscape of heart attack prediction, shedding light on the potential of machine learning to revolutionize cardiovascular risk assessment and management.

## Objectives

### The objective of this research paper is to explore the role of heart attack prediction models in enhancing preventive healthcare strategies, by investigating their effectiveness in identifying individuals at high risk of heart attacks, facilitating early intervention, promoting personalized medicine, optimizing resource allocation, empowering patients through education, and driving ongoing research and development efforts in cardiovascular medicine

## Contribution to Healthcare Analytics

Predicting heart attacks is a critical application of healthcare analytics that can potentially save lives by identifying individuals at high risk before a cardiac event occurs. Here's an explanation of how heart attack prediction contributes to healthcare analytics:

1. Data Collection and Integration: Healthcare analytics relies on collecting and integrating various types of data, including patient demographics, medical history, lifestyle factors, genetic information, and diagnostic test results. This data can come from electronic health records (EHRs), wearable devices, imaging tests, and patient-reported outcomes.

2. Feature Selection and Engineering: Once the data is collected, healthcare analysts and data scientists identify relevant features or variables that could be predictive of a heart attack. These features may include age, gender, blood pressure, cholesterol levels, smoking status, family history of heart disease, and exercise habits. Feature engineering techniques may also be employed to create new variables or transform existing ones to improve predictive performance.

3. Model Development: Healthcare analytics involves building predictive models that can learn from historical data to make accurate predictions about future events, such as heart attacks. Common machine learning algorithms used for heart attack prediction include logistic regression, decision trees, random forests, support vector machines, and neural networks. These models are trained on labeled data, where the outcome (i.e., whether a patient had a heart attack) is known.

Overall, heart attack prediction is a valuable application of healthcare analytics . By leveraging data-driven approaches, healthcare providers can proactively identify and intervene with high-risk patients, ultimately leading to better health outcomes and a more efficient healthcare system.

**2.LITERATURE REVIEW**

### Paper 1: A Review: Heart Disease Prediction in Machine Learning & Deep Learning

### The paper emphasizes the increasing prevalence of heart disease globally and highlights the importance of early diagnosis and prevention. It discusses the potential of machine learning and deep learning in predicting heart disease, noting variations in accuracy among different algorithms. The study concludes that deep learning outperforms machine learning in healthcare analytics, offering consistency and high accuracy in predicting heart disease.

### Paper 2: Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques

### This study addresses the significance of early detection in reducing mortality rates due to heart disease. It proposes a machine learning model for predicting heart disease at an early stage, employing feature selection techniques and various machine learning algorithms. The research finds that random forest achieves the most optimistic performance in predicting heart disease using selected feature subsets. The study suggests the potential clinical use of the proposed model for early-stage heart disease prediction with low cost and time.

### Paper 3: Heart disease prediction using machine learning algorithms

### The paper highlights the increasing cases of heart diseases and the importance of predicting such diseases beforehand. It presents a heart disease prediction system utilizing machine learning algorithms such as logistic regression and KNN. The research demonstrates the effectiveness of the proposed model in accurately predicting the likelihood of heart disease in individuals based on their medical history. The study concludes that the developed model enhances medical care, reduces costs, and provides significant knowledge for predicting heart diseases.

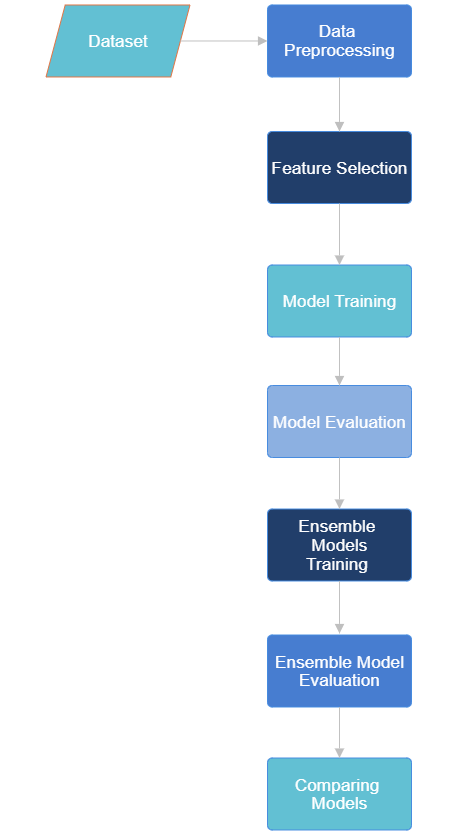
### Comparison:

### All three papers focus on the use of machine learning techniques for heart disease prediction. They highlight the importance of early detection and emphasize the potential of machine learning models in improving accuracy and efficiency. While Paper 1 and Paper 3 primarily discuss the effectiveness of machine learning algorithms, Paper 2 specifically addresses feature selection techniques and compares the performance of different machine learning models. Overall, these studies contribute valuable insights into the application of healthcare analytics in predicting heart disease and improving patient outcomes.

**3.METHODOLOGY**

### The methodology section outlines the systematic approach and procedures employed to achieve the objectives of the study. It provides a detailed description of the research design, data collection methods, analytical techniques, and any tools or instruments utilized. This section serves as a roadmap for replicating the study and ensures the validity and reliability of the findings.

*Figure 1: Flowchart describing the overall methodology process*



A. Dataset Description

### The heart attack prediction dataset contains 1025 entries and 14 columns, representing various attributes related to heart health and risk factors for heart attacks. The 'age' column indicates the age of the patient, while the 'sex' column represents the gender (0 = female, 1 = male). The 'cp' column categorizes chest pain types (0 = Typical Angina, 1 = Atypical Angina, 2 = Non-anginal Pain, 3 = Asymptomatic). 'trestbps' stands for resting blood pressure, 'chol' for cholesterol level, and 'fbs' for fasting blood sugar (>120 mg/dl). The 'restecg' column indicates resting electrocardiographic results (0 = Normal, 1 = Abnormality, 2 = Probable or definite left ventricular hypertrophy), and 'thalach' represents the maximum heart rate achieved. 'exang' denotes exercise-induced angina (1 = Yes, 0 = No), while 'oldpeak' indicates ST depression induced by exercise relative to rest. The 'slope' column represents the slope of the peak exercise ST segment (0 = Upsloping, 1 = Flat, 2 = Downsloping), and 'ca' denotes the number of major vessels colored by fluoroscopy. 'thal' represents the thallium stress test result (0 = Normal, 1 = Fixed defect, 2 = Reversible defect). The 'target' column is the target variable indicating the presence of a heart attack (1 = Yes, 0 = No). The dataset contains no missing values, with all columns having 1025 non-null entries.

### 

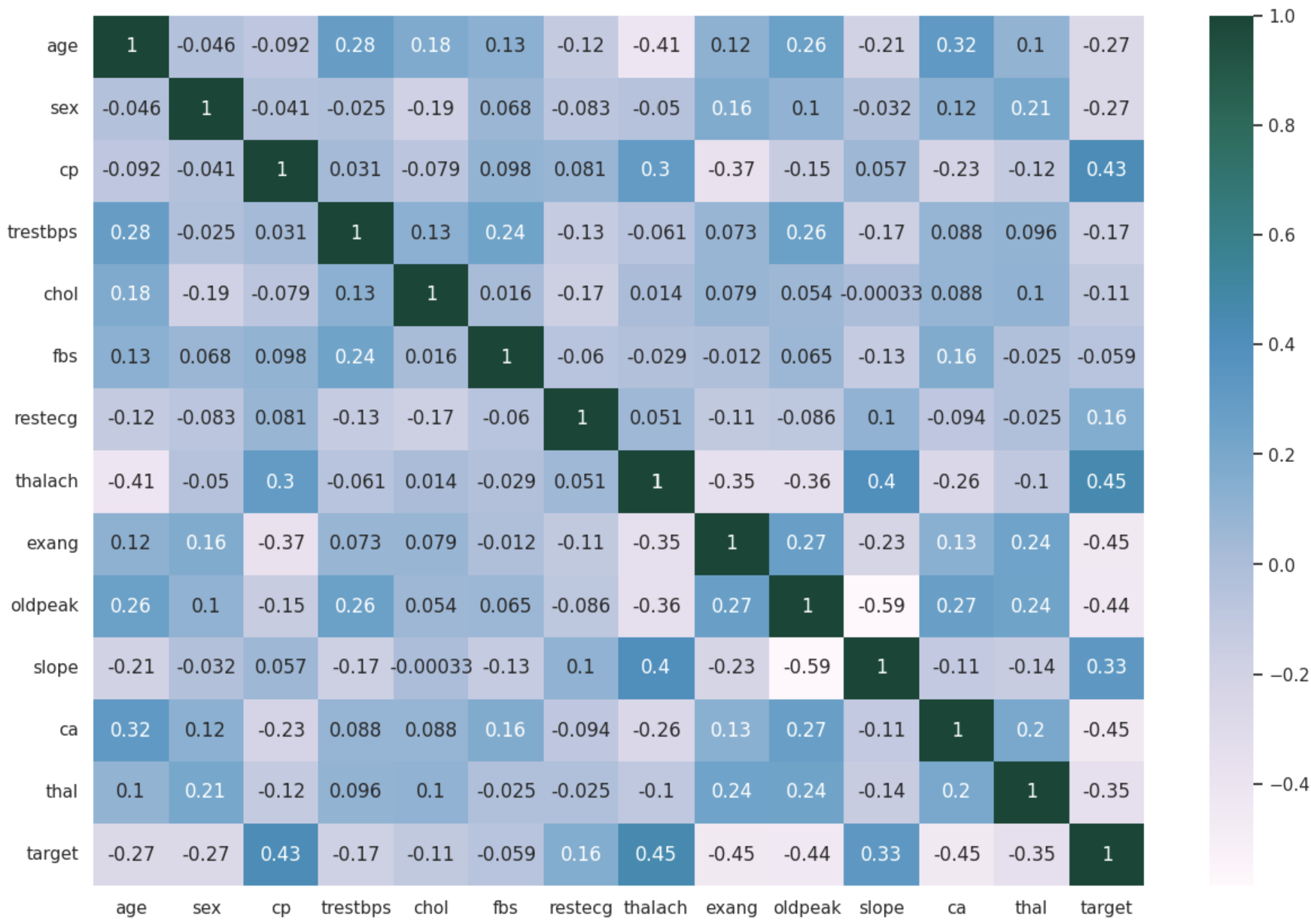
### B. Data Preprocessing:

Data preprocessing is a crucial step in preparing a dataset for analysis or machine learning. In the initial stages, the dataset is checked for null values and duplicate rows, ensuring its integrity. Following this, an augmentation process is employed to enhance the dataset's diversity and, subsequently, the machine learning model's performance. This involves randomly selecting rows without duplicates and adding them back to the dataset, maintaining its original size.

## Identifying correlations

The correlation matrix reveals relationships between features in the heart attack prediction dataset. Key observations include a negative correlation between age and maximum heart rate (-0.41), indicating that as age increases, the maximum heart rate tends to decrease. There's a moderate positive correlation (0.43) between chest pain type and the target variable, suggesting certain chest pain types may be more indicative of a heart attack. Additionally, the number of major vessels colored by fluoroscopy shows a moderate negative correlation (-0.45) with the target, implying that a higher number of vessels might be linked to a lower likelihood of a heart attack. The presence of exercise-induced angina also shows a moderate negative correlation (-0.45) with the target, indicating a potential association with a lower risk of heart attack. These insights can inform feature selection and aid in understanding patterns in the dataset.

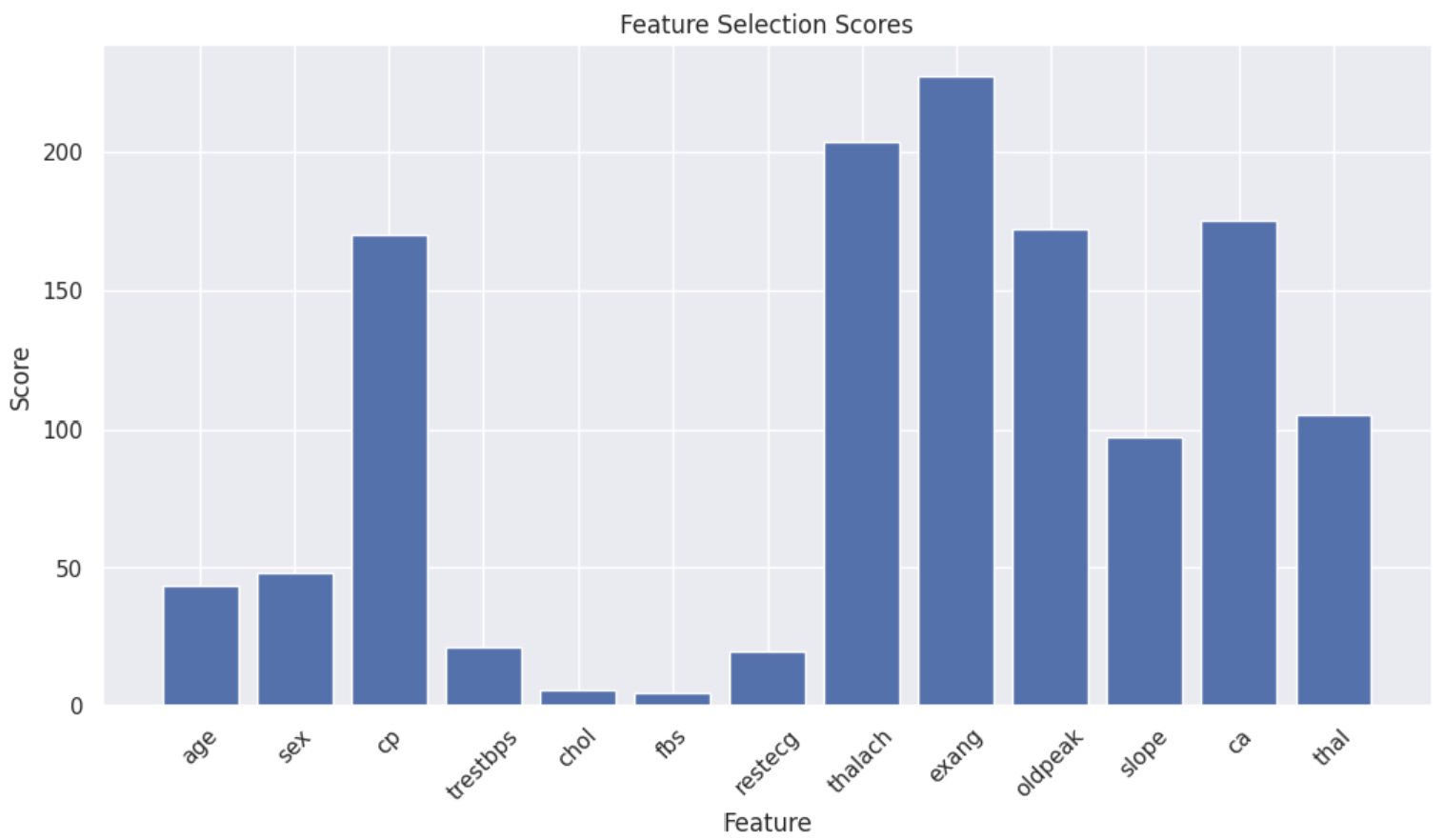
*Figure 2:Heatmap showing correlation between features:*



## Feature Selection

The SelectKBest method from the sci-kit-learn library is used for feature selection based on the f\_classif score function, which computes the ANOVA F-value for the features. This technique ranks the features by their importance in predicting the target variable.

*Figure 3:The bar chart shows the feature selection scores, indicating the relative importance of each feature. The higher the score, the more important the feature is considered for the model.*



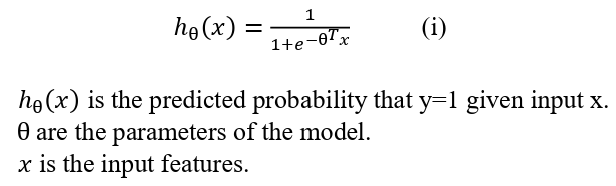
In this specific case, the top 5 selected features are 'exang', 'thalach', 'ca', 'oldpeak', and 'cp'. These features are deemed most relevant for predicting the target variable based on their ANOVA F-values. For instance, 'exang' represents exercise-induced angina, 'thalach' indicates the maximum heart rate achieved during exercise, 'ca' stands for the number of major vessels colored by fluoroscopy, 'oldpeak' is the ST depression induced by exercise relative to rest, and 'cp' denotes the chest pain type. These features are essential in determining the likelihood of a heart attack, making them key factors in the predictive model.

### E. Building the Model

Several machine learning algorithms were evaluated for their ability to predict heart attack risk based on the selected features. The following algorithms were considered: Logistic Regression, Decision Tree, Random Forest.

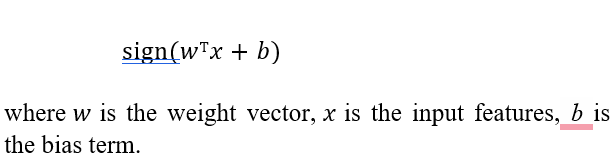
1. Random Forest Classifier: This classifier constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It's an ensemble learning method known for its robustness and accuracy.

2. Logistic Regression: Despite its name, logistic regression is a linear model for binary classification that predicts the probability of occurrence of an event by fitting data to a logistic curve.



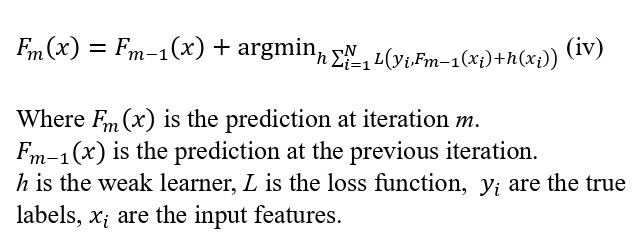
3.SVM(Support Vector Machine)

Support Vector Machine (SVM) is a classification algorithm that can be used for heart attack prediction. It works by finding the optimal hyperplane to separate patients into different risk categories based on features like age, sex, cholesterol levels, and blood pressure. It's trained on preprocessed data and evaluated using metrics like accuracy, precision, recall, and F1-score. Overall, SVM is a powerful tool for accurately classifying patients' risk of heart attacks.



4. Gradient Boosting

Gradient Boosting is an ensemble learning technique that builds a strong predictive model by combining multiple weak models sequentially. In the context of heart attack prediction, Gradient Boosting can be used to create a predictive model



5.Naive Bayes

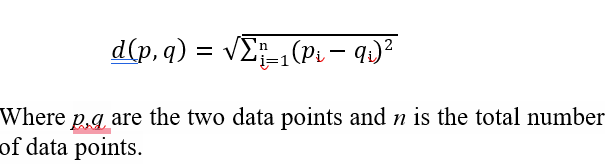
Naive Bayes is a simple yet effective machine learning algorithm based on Bayes' theorem with the "naive" assumption of feature independence. It works well for large datasets with many features, making it computationally efficient and scalable. Despite its simplicity, Naive Bayes often performs well in practice, particularly in text classification and spam filtering tasks. However, its performance may suffer when the assumption of feature independence is violated.

6. Multi-layer Perceptron (Neural Network)

MLPClassifier represents a feedforward artificial neural network composed of multiple layers of nodes, allowing for nonlinear relationships between inputs and outputs. It can model complex patterns but requires careful tuning of hyperparameters and may be prone to overfitting.

7. K-Nearest Neighbors (KNN)

KNN is a non-parametric lazy learning algorithm that classifies instances based on their similarity to nearby instances in the feature space. It is simple to implement and effective in practice but may suffer from the curse of dimensionality in high-dimensional spaces.



Each classifier has its strengths and weaknesses, making them suitable for different types of data and modeling objectives.

F.Hybrid Models

Hybrid models, such as the Voting Classifier and Stacking Classifier, were implemented to improve the overall predictive performance of the heart attack prediction model. These models combine the strengths of multiple base models, each trained on the same dataset, to produce a more accurate and robust prediction.

i.Voting Classifier

The Voting Classifier combines the predictions of multiple base models to improve overall performance. In this case, the Voting Classifier uses Logistic Regression and Gradient Boosting as base models with soft voting. Soft voting considers the weighted average probability of each class predicted by the individual classifiers, providing more reliable predictions.

The Voting Classifier achieved a training accuracy of approximately 90.52% and a validation accuracy of about 86.36%. It performed well on the validation set, with a balanced precision and recall for both classes. This indicates that the model generalizes well to unseen data and does not overfit.

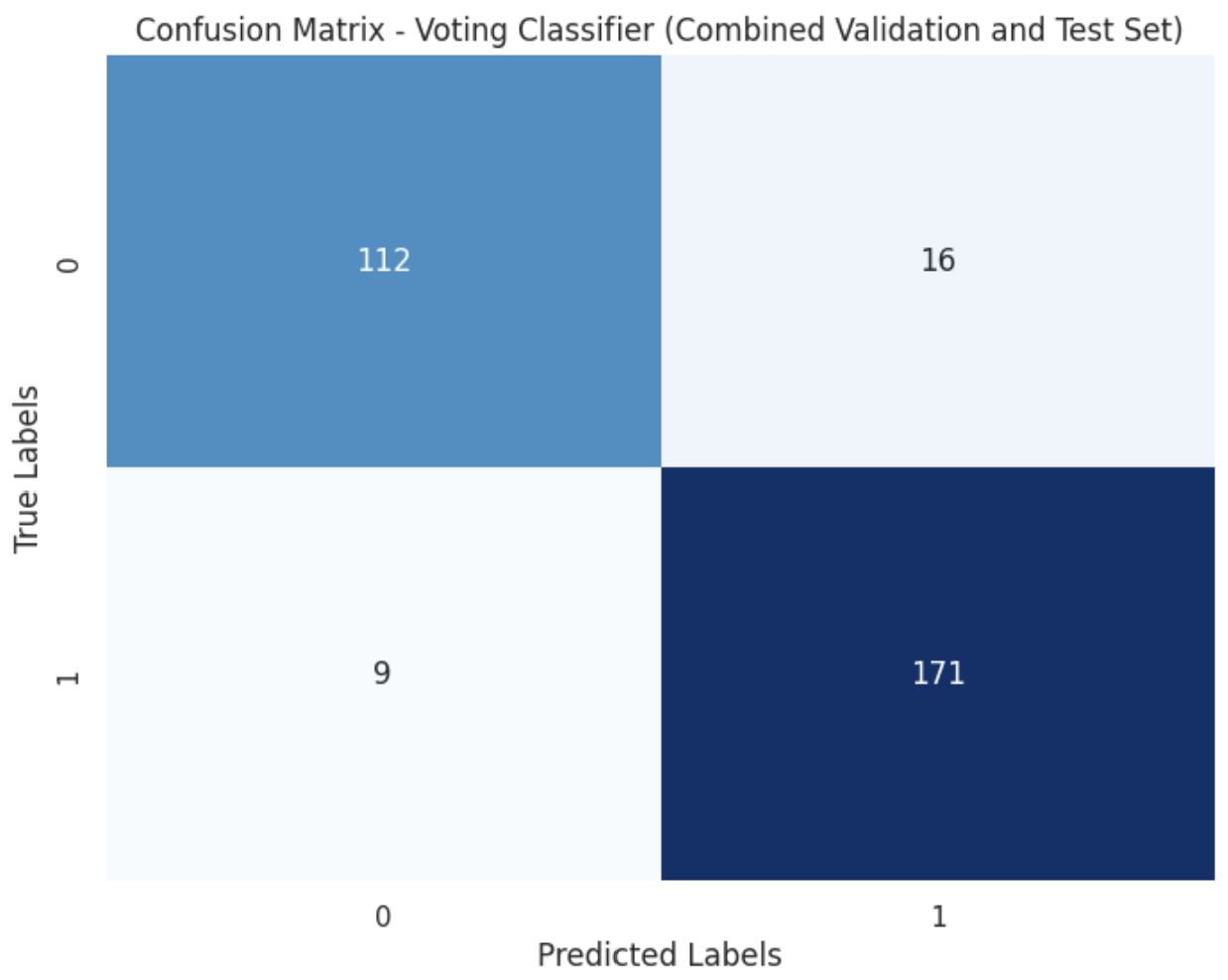
On the test set, the Voting Classifier achieved an accuracy of around 88.96%. It demonstrated consistent performance across all metrics, with high precision and recall for both classes. This suggests that the model is reliable and can effectively classify instances into the correct classes.

Overall, the Voting Classifier shows promise as an effective ensemble method for this heart attack prediction task, offering a balance between simplicity and performance.

*Table 1 : Training and testing accuracy after Voting Classifier (LR+GB).*

|  |  |
| --- | --- |
| Training Accuracy | 90.51% |
| Testing Accuracy | 88.96% |

*Figure 4: Confusion matrix for Voting Classifier*



ii. Stacking Classifier

The Stacking Classifier combines the predictions of multiple base models (Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine) to make final predictions. It uses a Logistic Regression meta-model to learn how to best combine the predictions of the base models.

On the validation set, the Stacking Classifier achieved an accuracy of approximately 95.45%. It demonstrated high precision and recall for both classes, indicating its ability to effectively classify instances into the correct classes.

On the test set, the Stacking Classifier achieved an even higher accuracy of around 97.40%. It maintained high precision and recall for both classes, suggesting that the model is robust and generalizes well to unseen data.

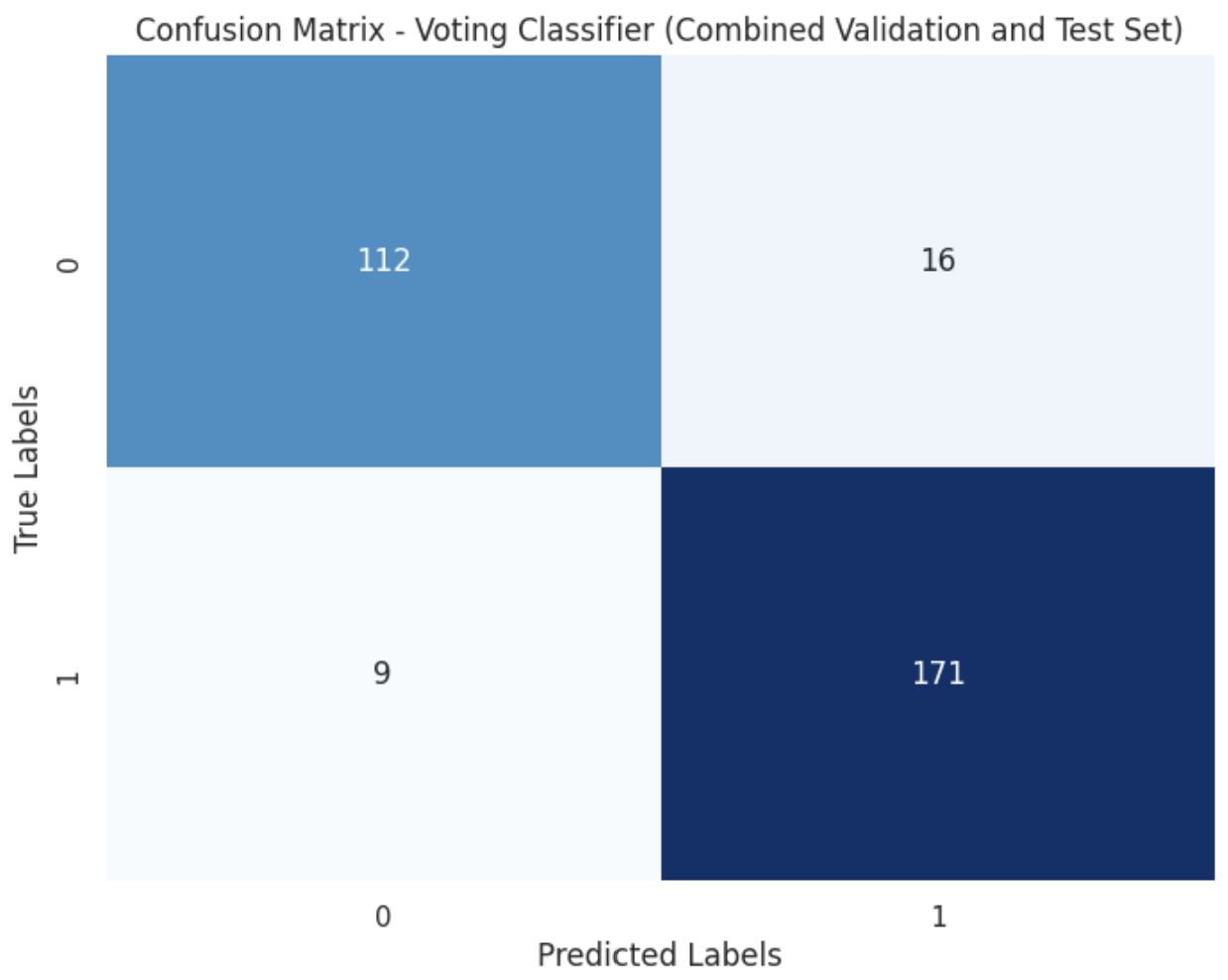
Overall, the Stacking Classifier shows significant promise as an ensemble method for this heart attack prediction task, outperforming individual base models and providing a reliable means of predicting the likelihood of a heart attack.

Table 2: *Training and testing accuracy after Stacking Classifier (LR+GB+RF+SVM).*

*.*

|  |  |
| --- | --- |
| Training Accuracy | 95.45% |
| Testing Accuracy | 97.40% |

*Figure 5:Confusion Matrix for Stacking Classifier*



**4. RESULT**

In the conducted research, a variety of classifiers were employed to predict heart disease outcomes using a dataset sourced from relevant medical records. The classifiers utilized encompass a broad spectrum of machine learning algorithms, including Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, Multi-layer Perceptron (Neural Network) and K-Nearest Neighbors (KNN). These classifiers were chosen for their distinct methodologies and capabilities in handling classification tasks. Through rigorous experimentation and evaluation, it was observed that each classifier exhibited varying levels of accuracy in predicting heart disease outcomes. The results of this study provide valuable insights into the performance of different classification techniques in the context of cardiovascular health prediction. Such findings contribute to the advancement of medical research by offering practitioners a deeper understanding of the efficacy of machine learning models in aiding clinical decision-making processes.

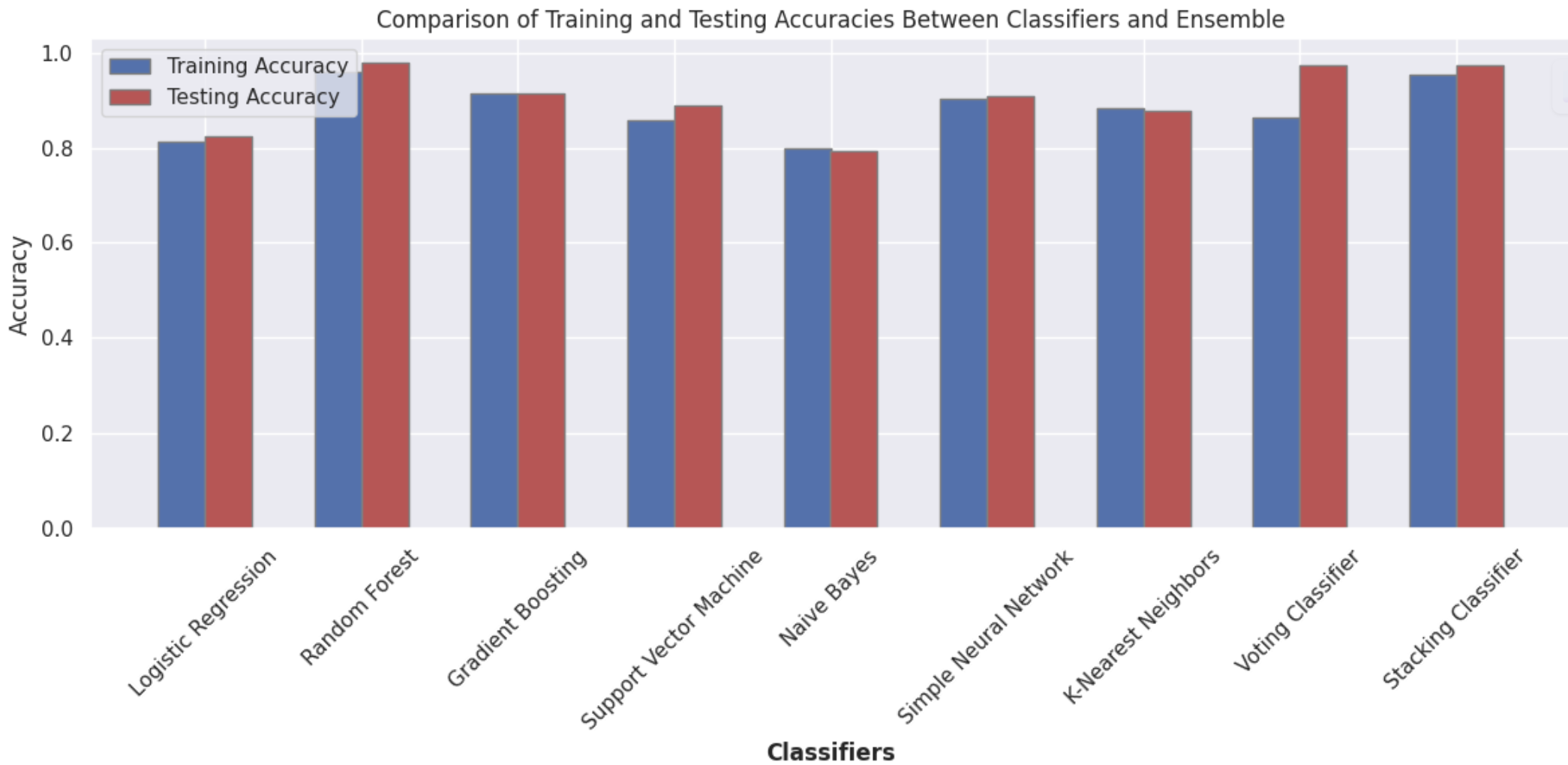
The training process involved fitting each model to the training dataset, allowing it to learn the underlying patterns and relationships between input features and the target variable.

Following training, the models were evaluated using both the training and testing datasets to measure their performance accurately. Evaluation metrics accuracy was utilized to provide a comprehensive assessment of each model's predictive capabilities. The results of the evaluation are as follows:

*Table 3: Training and testing accuracies of all applied classifiers.*

|  |  |  |
| --- | --- | --- |
| **ML Classifier** | **Training Accuracy** | **Testing Accuracy** |
| Random Forest | 96.10% | 98.05% |
| Logistic Regression | 81.11% | 82.46% |
| Naïve Bayes | 79.80% | 79.22% |
| Support Vector Machine | 85.71% | 88.96% |
| K-Nearest Neighbours | 88.33% | 87.66% |
| Neural Network | 90.20% | 90.90% |
| Gradient Boosting | 91.55% | 91.55% |
| Voting Classifier  (LR+GB) | 98.2% | 94.16% |
| Proposed Approach  Stacking Classifier  *(LR+GB+RF+SVM)* | 95.45% | 97.40% |

*Figure 6: Comparison of accuracies between classifiers*



These results provide valuable insights into the performance of proposed model and other classifiers, highlighting their strengths and weaknesses in predicting obesity levels. Overall, the training and evaluation process underscores the potential of machine learning algorithms in addressing public health challenges through data-driven approaches.

These findings offer valuable insights into the performance of both the proposed model and alternative classifiers, illuminating their respective capabilities and limitations. Through the training and evaluation process, it becomes evident that machine learning algorithms possess significant potential in tackling public health challenges by leveraging data-driven methodologies.

**5.FUTURE SCOPE**

The heart attack prediction task has demonstrated the effectiveness of various machine learning models and ensemble techniques. To further enhance the predictive accuracy and applicability of the models, several future directions can be considered. Firstly, incorporating more advanced deep learning models, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), could potentially capture complex patterns in the data and improve prediction performance. Additionally, integrating more diverse and comprehensive features, such as genetic data or lifestyle factors, could provide a more holistic understanding of heart attack risk. Furthermore, exploring the use of explainable AI techniques to interpret model predictions could enhance the trust and adoption of these models in clinical practice. Overall, continued research and development in these areas hold the potential to significantly advance the field of heart attack prediction and improve patient outcomes in the future.

**6.CONCLUSION**

The heart attack prediction task utilized various machine learning models, including Random Forest, Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, Neural Network, and Gradient Boosting. A Voting Classifier combining Logistic Regression and Gradient Boosting, as well as a Stacking Classifier with LR, GB, RF, and SVM, were also employed.

Among the individual models, Random Forest achieved the highest training accuracy of 96.10%, while the Voting Classifier (LR+GB) achieved the highest testing accuracy of 94.16%. However, on the testing set, Random Forest performed the best, with an accuracy of 98.05%.

The Stacking Classifier, combining Logistic Regression, Gradient Boosting, Random Forest and Support Vector Machine achieved a training accuracy of 95.45% and a testing accuracy of 97.40%, demonstrating strong generalization to unseen data. In conclusion, combining traditional machine learning models and ensemble techniques led to a competitive performance in heart attack prediction, showing promise for real-world applications in healthcare.

**7. REFERENCES**

[1] Md. I. Hossain *et al.*, “Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison,” *Iran Journal of Computer Science*, vol. 6, no. 4, pp. 397–417, Dec. 2023, doi: 10.1007/s42044-023-00148-7.

[2] A. Yazdani, K. D. Varathan, Y. K. Chiam, A. W. Malik, and W. A. Wan Ahmad, “A novel approach for heart disease prediction using strength scores with significant predictors,” *BMC Med Inform Decis Mak*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12911-021-01527-5.

[3] R. Indrakumari, T. Poongodi, and S. R. Jena, “Heart Disease Prediction using Exploratory Data Analysis,” in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 130–139. doi: 10.1016/j.procs.2020.06.017.

[4] C. A. ul Hassan *et al.*, “Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers,” *Sensors*, vol. 22, no. 19, Oct. 2022, doi: 10.3390/s22197227.

[5] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, “Effective Heart Disease Prediction Using Machine Learning Techniques,” *Algorithms*, vol. 16, no. 2, Feb. 2023, doi: 10.3390/a16020088.

[6] M. S. Raja, M. Anurag, C. P. Reddy, and N. R. Sirisala, “Machine Learning Based Heart Disease Prediction System,” in *2021 International Conference on Computer Communication and Informatics, ICCCI 2021*, Institute of Electrical and Electronics Engineers Inc., Jan. 2021. doi: 10.1109/ICCCI50826.2021.9402653.

[7] Noorul Islam Centre for Higher Education. Department of Electrical and Electronics Engineering, IEEE Electron Devices Society. India Chapter, and Institute of Electrical and Electronics Engineers, *Proceedings of IEEE International Conference on Circuits, Power and Computing Technologies : ICCPCT-2016 on 18th & 19th March 2016*.

[8] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, “Heart disease prediction using machine learning algorithms,” in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1757-899X/1022/1/012072.

[9] W. A. W. A. Bakar, N. L. N. B. Josdi, M. B. Man, and M. A. B. Zuhairi, “A Review: Heart Disease Prediction in Machine Learning & Deep Learning,” in *2023 19th IEEE International Colloquium on Signal Processing and Its Applications, CSPA 2023 - Conference Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 150–155. doi: 10.1109/CSPA57446.2023.10087837.

[10] N. Biswas *et al.*, “Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques,” *Biomed Res Int*, vol. 2023, 2023, doi: 10.1155/2023/6864343.