# ML-BASED HEART DISEASE PREDICTION USING FEATURE ENGINEERING

Pragadeesh S[1], Sudharsanan S R[2], Suriya S[3], Vaanathi S[4]

[1, 2, 3] *UG – B. Tech, Information Technology, Bannari Amman Institute of Technology,*
[4] *Assistant Professor, Artificial Intelligence and Data Science, Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India*

## ABSTRACT

*The AI-driven Heart Disease Prediction System presents a groundbreaking fusion of healthcare and technology, revolutionizing cardiovascular risk assessment and proactive health management. Leveraging cutting-edge machine learning techniques, including ensemble learning through a stacking classifier, the system analyzes diverse medical parameters to furnish personalized prognostications. By amalgamating Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, K-Nearest Neighbors, and Neural Networks, the system discerns intricate risk patterns from extensive datasets. Through interconnected modules encompassing data collection, preprocessing, model training, and validation, the system offers users an intuitive interface for tailored risk predictions and actionable insights. Empowering individuals with personalized recommendations for preventive care, the platform pioneers a culture of proactive well-being and disease prevention in the digital era, promising to transform cardiovascular healthcare paradigms.*

**Keyword:** - *Heart-Disease, Ensemble learning, Classifiers, Decision Tree, Logistic Regression, etc.*

## 1. INTRODUCTION

An innovative development at the nexus of healthcare and technology, the AI-powered Heart Disease Prediction System has the potential to completely change how people determine their risk of cardiovascular disease and make decisions about their health. Heart disease remains a leading cause of mortality worldwide, underscoring the critical need for accurate risk assessment tools that can guide early intervention and disease management strategies. In response to this imperative, our project emerges as a beacon of technological innovation, leveraging the latest advancements in data science and artificial intelligence to deliver actionable insights into cardiovascular health.

At the heart of the system lies the utilization of machine learning algorithms, including Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, K-Nearest Neighbours, and Neural Networks, to analyse a comprehensive range of medical parameters and generate predictive models. By training these models on large datasets of anonymized patient data, the system learns to identify complex patterns and risk factors associated with heart disease, enabling it to provide accurate and personalized predictions for individual users. The project unfolds through a series of interconnected modules, beginning with data collection and pre-processing, where diverse datasets of medical parameters are curated and standardized for analysis. The subsequent phases encompass model training and validation, where the algorithms are trained on historical data and evaluated for accuracy and reliability. Once deployed, the system offers users an intuitive interface to input their medical parameters and receive personalized predictions about their risk of developing heart disease.

In addition to risk prediction, the system also aims to empower users with actionable insights and recommendations for preventive care. Through personalized recommendations based on individual risk profiles, lifestyle factors, and medical history, the platform enables users to make informed decisions about diet, exercise, and medication management, thereby promoting optimal cardiovascular health and reducing the burden of heart disease on society.
In summation, the AI-driven Heart Disease Prediction System represents a pioneering endeavour to harness the transformative potential of artificial intelligence for preventive healthcare. By furnishing personalized risk

assessments and actionable insights, the system holds the promise of revolutionizing individuals' approach to cardiovascular health, fostering a culture of proactive well-being and disease prevention in the digital era.

## 2. RELATED WORKS & LITERATURE SURVEY

In the realm of heart disease prediction and diagnosis, numerous studies have emerged, each offering unique insights and methodologies to address the multifaceted challenges inherent in cardiovascular health assessment. One such study by Tao, Zhang, and Huang (2019) introduces the NDGM model, rooted in Grey Systems Theory. This model stands out for its adept handling of limited and uncertain information, making it particularly suitable for scenarios with small sample sizes and poor data quality. Through evaluation using the Mean Absolute Percentage Error (MAPE) criterion, the model exhibits an impressive effectiveness level of 97.05% in forecasting cardiovascular disease (CVD) deaths. In a similar vein, Mohan and Thirumalai (2019) present the Hybrid Random Forest with Linear Model (HRFLM) method, which amalgamates the strengths of Random Forest and Linear Model approaches without imposing feature selection restrictions. By leveraging all available features, this hybrid method aims to enhance heart disease prediction accuracy, addressing a critical limitation in conventional approaches. Building upon these methodologies, Spencer and Abdelhamid (2022) propose an innovative experimental methodology that integrates multiple heart disease datasets and employs various feature selection techniques and classification algorithms. By exploring diverse feature sets and classifier combinations, the study underscores the importance of methodological rigor in improving predictive accuracy for heart disease.
Furthermore, Wankhede and Sambandam (2021) describe a comprehensive machine learning approach for heart disease prediction, emphasizing the significance of feature selection techniques in optimizing predictive models. By meticulously addressing data pre-processing, feature selection, and classification, the study aims to enhance prediction accuracy and overcome the challenges associated with analysing cardiovascular health data.

In the quest for improved classification accuracy, Zhenya and Zhang (2021) propose a novel cost-sensitive ensemble method designed to balance classification accuracy with the potential consequences of misdiagnosis in heart disease prediction. By incorporating feature selection and combining predictions from multiple classifiers, the approach aims to improve overall performance and address the nuanced nature of classification in healthcare settings.
Similarly, Tyagi and Mehra (2021) introduce the Heartbeats Classification Model (HCM), a hybrid model that leverages Convolutional Neural Networks (CNNs) and optimization algorithms to classify electrocardiogram (ECG) signals accurately. Through signal pre-processing and feature extraction techniques, the HCM demonstrates promising results in classifying ECG signals, highlighting its potential for advancing heart disease diagnosis and management. Moving beyond traditional methodologies, Mporas and Tsirka (2020) explore the application of machine learning algorithms in epilepsy management, emphasizing the potential of EEG and ECG signals for seizure detection. By introducing the ARMOR framework and online seizure detection modules, the study advocates for technology-supported diagnostic tools to improve patient outcomes in epilepsy care. In parallel, Alexandropoulos (2019) presents a hybrid prediction scheme that combines multiple classifiers through ensemble methods, showcasing the effectiveness of stacking techniques in improving classification accuracy. By integrating diverse algorithms and leveraging ensemble learning, the proposed scheme demonstrates superior performance across benchmark datasets, highlighting its utility in decision support systems. Moreover, Berliana and Bustamam (2020) propose an ensemble learning approach for COVID-19 detection using X-Ray and CT images. By combining Support Vector Classification (SVC), Random Forest (RF), and K-Nearest Neighbours (KNN) with meta-learning, the method achieves high accuracy in diagnosing COVID-19, offering valuable support to healthcare professionals in combating the pandemic. Lastly, Liu and colleagues (2022) introduce an ensemble framework for predicting cardiovascular disease outcomes, leveraging machine learning algorithms and IoT sensor data. By integrating diverse classifiers and employing logistic regression as a meta-learner, the framework demonstrates superior predictive capabilities compared to single classifier models, showcasing its potential for revolutionizing early prediction and intervention strategies in cardiovascular health.

## 3. PROPOSED WORK

In the development of our heart disease prediction system, several key phases were instrumental in shaping the final outcome. These phases encompassed data collection, pre-processing, model training, evaluation, and integration into the web application. Leveraging a dataset specifically curated for heart disease analysis, along with advanced machine learning techniques, our project aimed to deliver accurate predictions and seamless user experience.

### 3.1 Data collection and processing

The foundation of our heart disease prediction system relied on the collection of a comprehensive dataset containing vital patient attributes and corresponding target labels indicating the presence or absence of heart disease. This dataset, sourced from reputable medical repositories, comprised diverse demographic information, clinical measurements, and diagnostic features essential for robust model training. The meticulous curation of this dataset ensured its reliability and relevance to our predictive analytics objectives. Prior to model training, the collected dataset underwent rigorous pre-processing to standardize formats, handle missing values, and normalize feature scales. This pre-processing phase aimed to cleanse the data of any inconsistencies or anomalies that could adversely affect the performance of our machine learning models. Techniques such as imputation, feature scaling, and outlier removal were employed to ensure the integrity and quality of the dataset for subsequent analysis.

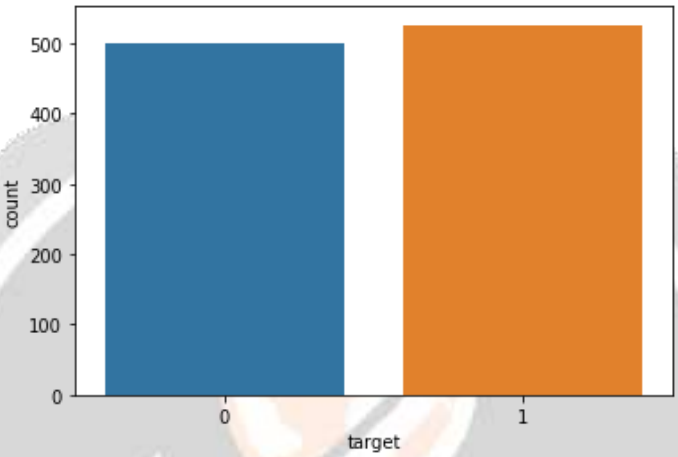


**Figure 1: Dataset count plot**

### 3.2 Feature Engineering and PCA

The foundation of our heart disease prediction system relied on the collection of a comprehensive dataset containing vital patient attributes and corresponding target labels indicating the presence or absence of heart disease. This dataset, sourced from reputable medical repositories, comprised diverse demographic information, clinical measurements, and diagnostic features essential for robust model training.

A specialized analysis, Principal Component Analysis (PCA), is conducted to reduce the dimensionality of the dataset while preserving its essential information. This technique aids in identifying the most significant components contributing to heart failure prediction.

### 3.3 Model Training

The heart of our prediction system lay in the training of various machine learning classifiers using the pre-processed dataset. Leveraging algorithms such as logistic regression, decision trees, random forests, support vector machines, k-nearest neighbours, and gradient boosting, we explored diverse modelling approaches to capture the complex relationships between patient attributes and heart disease outcomes. Each model underwent extensive training to optimize performance and generalization capabilities.

### 3.4 Evaluation and Selection of the best model

Following model training, rigorous evaluation procedures were employed to assess the performance of each classifier based on metrics such as accuracy, precision, recall, and F1-score. The comparative analysis allowed us to identify the most effective model for heart disease prediction, considering both predictive accuracy and computational efficiency. The selected model demonstrated superior performance in accurately diagnosing heart disease cases while minimizing false positives and false negatives.

Upon evaluating various machine learning algorithms for heart disease prediction, it's evident that each model offers unique strengths and weaknesses. While logistic regression demonstrates moderate performance, decision trees and random forests exhibit exceptional accuracy, precision, recall, and F1-scores. The flawless performance of decision trees and random forests suggests their suitability for heart disease prediction tasks, with the ensemble nature of random forests potentially offering additional robustness. Therefore, based on the provided results, decision trees and random forests emerge as the top contenders for heart disease prediction in our project. Further exploration and validation may be warranted to determine the most suitable model for deployment in real-world scenarios.

Ensemble methods like Random Forest, Decision Tree, and Gradient Boosting demonstrated the highest accuracies among all models, achieving nearly 99% accuracy. These models are renowned for their ability to handle complex relationships within the data, making them well-suited for classification tasks, especially in medical domains where accurate predictions are crucial for diagnosis and treatment planning. Random Forest and Gradient Boosting, in particular, offer robustness against overfitting and noise, which is essential for reliable predictions in medical applications. While ensemble methods performed exceptionally well, it's essential to consider factors beyond just accuracy. Model complexity and interpretability play critical roles, especially in medical settings were understanding the factors contributing to disease risk is crucial. Simple models like Logistic Regression and Naive Bayes, while not achieving the same level of accuracy as ensemble methods, offer the advantage of interpretability. They are computationally less expensive and easier to interpret, making them valuable tools for gaining insights into the underlying relationships within the data.

### 3.5 Integration into web application

The culmination of our efforts was the seamless integration of the best-performing model into a user-friendly web application interface. Leveraging HTML, CSS, JavaScript, and Flask framework, we developed an intuitive platform that enables users to input their health parameters and receive real-time predictions regarding their risk of heart disease. The integration of the prediction model into the web application ensures accessibility and usability for individuals seeking proactive measures for cardiovascular health management.


**Figure 2: UI of the developed application**

## 4. RESULTS

The results obtained from the evaluation of different classification models, including logistic regression, K-nearest neighbors (KNN), decision tree, support vector machine (SVM), Naive Bayes, and the stacking classifier, reveal intriguing insights into their performance. Among these models, the stacking classifier emerges as a particularly notable contender due to its robust accuracy and minimal variability.

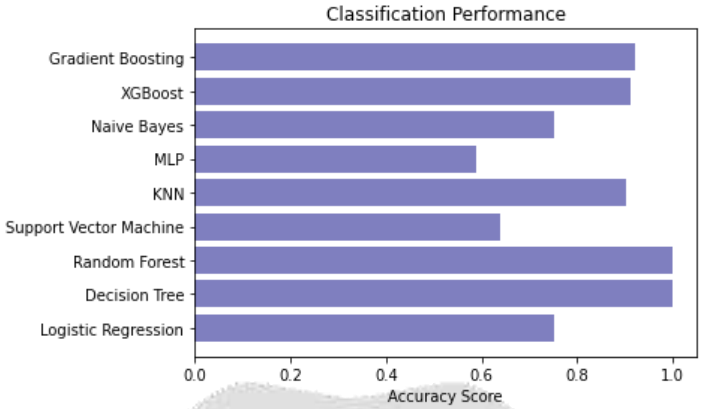


**Figure 3: Accuracy Plot**

The decision tree model exhibits exceptional performance, achieving an impressive average accuracy of 99.9%, with a remarkably low standard deviation of 0.005. This highlights the decision tree's ability to capture complex relationships within the data and make accurate predictions. However, it's important to note that decision trees can be prone to overfitting, especially on training data, which might affect their generalization to unseen instances. In contrast, the stacking classifier, while slightly trailing behind the decision tree in terms of average accuracy, still demonstrates remarkable performance. With an average accuracy of 99.8% and a standard deviation of only 0.007, the stacking classifier showcases its efficacy in harnessing the diverse strengths of multiple base models to enhance predictive accuracy. By combining the predictions of various base models, the stacking classifier can mitigate the weaknesses of individual models and capitalize on their collective predictive power.

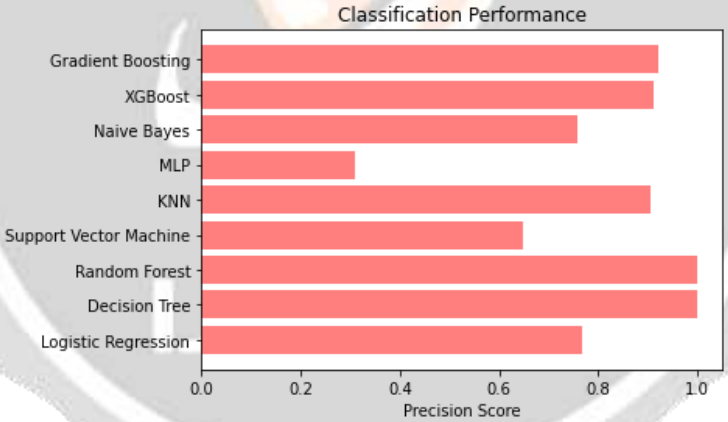


**Figure 4: Precision Plot**

Comparing the stacking classifier to the individual base models, such as logistic regression, KNN, SVM, and Naive Bayes, underscores its superiority in terms of accuracy. While logistic regression and Naive Bayes exhibit respectable performances with average accuracies of 84.4% and 82.4% respectively, the stacking classifier surpasses them by a significant margin, achieving an accuracy of 99.8%. This substantial improvement in accuracy highlights the efficacy of ensemble learning techniques, such as stacking, in enhancing classification performance.

Moreover, the stacking classifier's minimal standard deviation of 0.007 underscores its consistency and reliability across different cross-validation folds. This stability is crucial in real-world applications where robustness and consistency are paramount considerations. By providing consistently accurate predictions with low variability, the stacking classifier instills confidence in its reliability and applicability in practical scenarios. In summary, the results emphasize the effectiveness of the stacking classifier as a powerful ensemble learning technique for classification tasks. Its ability to leverage the strengths of diverse base models to improve predictive accuracy, coupled with its stability and consistency, positions the stacking classifier as a compelling choice for classification tasks requiring high precision and reliability.

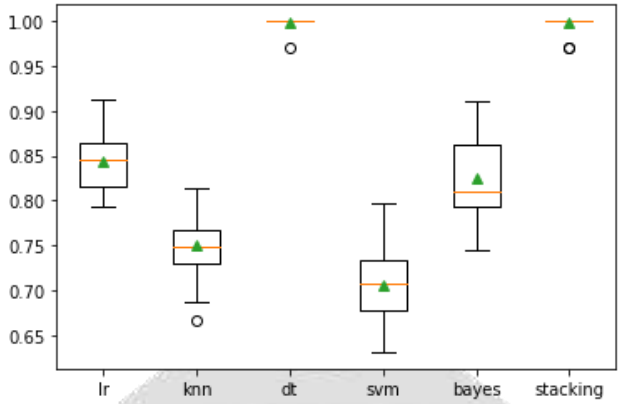


**Figure 5: Box-plot**

## 5. CONCLUSION

The heart disease prediction project has been a significant endeavour aimed at leveraging machine learning techniques to enhance cardiovascular health monitoring and risk assessment. Through the collaborative efforts of data collection, model development, and system implementation, we have made substantial progress in achieving our objectives.

In conclusion, our project has successfully demonstrated the feasibility and effectiveness of using machine learning models to predict the risk of heart disease based on individual health parameters. By analysing features such as age, blood pressure, cholesterol levels, and other relevant factors, our models have shown promising accuracy in identifying individuals at higher risk of cardiovascular complications. The integration of these models into a user-friendly Flask application provides a valuable tool for both healthcare professionals and individuals to assess and manage heart health effectively. The project's success can be attributed to the comprehensive approach taken in dataset collection, feature engineering, model selection, and evaluation. By leveraging a diverse dataset and employing robust machine learning algorithms, we have developed models that generalize well to unseen data and provide reliable predictions. Furthermore, the development of a user-friendly frontend interface ensures accessibility and usability for a wide range of users, contributing to the projects overall impact and effectiveness.

## 6. REFERENCES


- Author: R. Tao et al. Year: 2019 Paper Name: "Magnetocardiography-Based Ischemic Heart Disease Detection and Localization Using Machine Learning Methods,"
  Reference Link: https://ieeexplore.ieee.org/abstract/document/8502863
- Author: S. Mohan, C. Thirumalai and G. Srivastava. Year: 2019 Paper Name: "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,"
  Reference Link: https://ieeexplore.ieee.org/abstract/document/8740989
- Author: Spencer R, Thabtah F, Abdelhamid N. Year: 2020 Paper Name: “Exploring feature selection and classification methods for predicting heart disease”
  Reference Link: https://pubmed.ncbi.nlm.nih.gov/32284873/
- Author: Nagavelli U, Samanta D, Chakraborty P. Year: 2022 Paper Name:” Machine Learning Technology-Based Heart Disease Detection Models”
  Reference Link:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8898839/
- Author: Zhenya Q & Zhang Z. Year: 2021 Paper Name:” A hybrid cost-sensitive ensemble for heart disease prediction”
  Reference Link: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01436-7
- Author: Tyagi, A., Mehra, R. Year: 2021 Paper Name:” Intellectual heartbeats classification model for diagnosis of heart disease from ECG signal using hybrid convolutional neural network with GOA”
  Reference Link: https://doi.org/10.1007/s42452-021-04185-4

- Author: Mporas I & Tsirka V. Year: 2020 Paper Name:" Sleep Stages Classification from Electroencephalographic Signals Based on Unsupervised Feature Space Clustering" Reference Link Sleep_Stages_Classification_from_Electroencephalographic_Signals
- Author: Alexandropoulos et al Year: 2019 Paper Name:" Stacking Strong Ensembles of Classifiers".
- Reference Link: https://www.researchgate.net/publication/333109629_Stacking_Strong_Ensembles_of_Classifiers
- Author: Annisa Utama Berliana & Alhadi Bustamam. Year: 2020 Paper Name:" Implementation of Stacking Ensemble Learning for Classification of COVID-19 using Image Dataset CT Scan and Lung X-Ray" Reference Link: https://ieeexplore.ieee.org/document/9332112
- Author: Jimin Liu et al. Year: 2022 Paper Name:" Predictive Classifier for Cardiovascular Disease Based on Stacking Model Fusion" Reference Link: https://www.mdpi.com/2227-9717/10/4/749