Ryerson University
Department of Electrical, Computer, & Biomedical Engineering
Faculty of Engineering & Architectural Science

Heart Disease Prediction Using Hybrid Machine Learning Algorithms

by

Karishab Sharma, Alby Thekkemuriyil, Farzaan Rahim, Saikot Paul

Report Summarizing Design Activities - COE 70B

Electrical & Computer Engineering Capstone Design Project

Toronto Metropolitan University, 2025

Acknowledgments

We would like to express our sincere gratitude to our advisor, Dr. Truman Yang, whose expertise and unwavering support have been invaluable throughout both semesters of our project. His guidance and encouragement pushed us to challenge conventions and confidently address the complexities of developing and deploying hybrid machine-learning techniques. Dr. Yang's openness to our innovative ideas allowed us to explore unique solutions while ensuring our perspectives were integrated into the project's progress.

We also extend our heartfelt thanks to the authors and researchers whose work formed the foundation of our study. Their contributions, particularly in hybrid machine-learning methods and data processing technologies, provided us with crucial insights and shaped our approach. The wealth of knowledge from the academic community served as both guidance and inspiration, driving our efforts to tailor existing solutions to the specific needs of this project.

Furthermore, we would like to acknowledge the important role of GitHub in our research. The availability of open-source platforms and datasets on GitHub significantly enriched our research framework. These tools advanced our technical progress and sparked ideas that helped refine and expand on our project's core principles.

Finally, we deeply appreciate every team member's dedication and hard work. From conducting detailed research to analyzing literature and refining our methodology, the team's commitment was integral to our success. This semester's advancements build upon our earlier accomplishments, and we look forward to the continued impact of our efforts as we complete this transformative journey.

Certification of Authorship

We certify that the content of this report is entirely original, and we prepared it solely as a group. Any references to the text, figures, and tables have been properly cited where applicable. We affirm that the provided report follows the guidelines stated in the report template presented by Dr. Naimul Khan.

Karishab Sharma	Karishab Sharma	04-05-2025
Signature	**Student A**	**Date**
A Thekkemuriyil	Alby Thekkemuriyil	04-05-2025
Signature	**Student B**	**Date**
FR	Farzaan Rahim	04-05-2025
Signature	**Student C**	**Date**
S.P.	Saikot Paul	04-05-2025
Signature	**Student D**	**Date**

Table of Contents

Abstract

Heart disease remains a critical global health issue, driving the need for innovative solutions that can effectively assess and mitigate risk. This project focuses on the development of a hybrid machine-learning system to predict heart disease risk by analyzing key clinical and demographic data, such as age, blood pressure, and cholesterol levels. By integrating multiple machine-learning models, the approach aims to enhance accuracy and reliability while addressing common challenges like bias and overfitting. Designed with accessibility in mind, the system provides actionable insights through a user-friendly interface, supporting informed decision-making for both healthcare professionals and individuals. In this phase of the project, the hybrid prediction system was implemented and refined, employing Logistic Regression and Random Forest models. Extensive testing and evaluation on diverse datasets ensured robust performance across key metrics, demonstrating the reliability of this approach. Strategic preprocessing techniques, such as sampling optimization and imbalance handling, further enhanced the system's precision and adaptability. Design improvements focused on creating an intuitive interface that delivers clear predictions, ensuring ease of use in real-world healthcare settings. Additionally, the exploration of real-time data integration and cloud-based deployment highlighted the potential for scalability and widespread adoption. The outcomes of this research validate the efficacy of hybrid machine-learning techniques in heart disease prediction. The system not only demonstrated strong predictive performance but also proved its practical applicability, offering a reliable tool for early intervention and risk management. Future efforts will focus on expanding dataset diversity, optimizing model architecture, and integrating continuous monitoring through wearable devices. These advancements aim to refine the system further, ensuring it meets healthcare standards and addresses the needs of both clinical environments and individual users. This research lays a strong foundation for scalable and impactful heart disease prediction technology. By blending cutting-edge machine-learning methods with practical usability, the project aspires to contribute meaningfully to preventing and managing one of the most pressing health challenges worldwide.

Keywords: Heart Disease Prediction, Hybrid Machine Learning, Ensemble Models, Clinical Data, Healthcare Applications.

Introduction & Background

Heart disease continues to be a leading global health challenge, accounting for 1 in 5 deaths worldwide and highlighting the urgent need for effective tools to assess and reduce risk [2]. Advances in machine learning have empowered systems to evaluate patient data and identify potential health risks with increased precision. This research project focuses on developing a hybrid machine-learning-based prediction system to identify and mitigate the risks associated with heart disease. The primary objective is to create a robust model that evaluates critical metrics, such as age, blood pressure, cholesterol, and other health indicators, to detect individuals at greater risk of cardiovascular complications.

To overcome common challenges in predictive modeling, including bias, overfitting, and reliability, this project adopts a hybrid approach by combining conventional and advanced machine-learning techniques. Hybrid methodologies are designed to leverage the strengths of multiple algorithms, mitigating the limitations of single-model approaches while enhancing predictive performance. The research involves a systematic analysis of existing hybrid and machine-learning models to evaluate their strengths and weaknesses, ensuring the selection of an optimized framework for heart disease prediction.

The proposed system provides actionable insights, enabling individuals and healthcare professionals to make informed decisions regarding cardiovascular health. It features an intuitive, user-friendly interface that allows users to input their health data and receive detailed risk assessments, complete with explanations of contributing factors. This dual utility ensures accessibility for personal use and practical applicability in clinical settings.

Building on the progress from the initial phase, this semester focuses on implementing and refining the hybrid model. Logistic Regression and Random Forest were identified as the primary algorithms, offering a balance between accuracy and interpretability. Rigorous testing on diverse datasets has confirmed the system’s reliability across performance metrics. Additionally, design enhancements have been made to ensure usability, accessibility, and scalability, positioning the system as a practical tool for real-world healthcare applications.

Although heart disease prediction has been extensively explored, hybrid techniques remain underutilized. This research addresses that gap by advancing hybrid machine-learning approaches, aiming to provide a reliable, deployable solution for assessing risk and monitoring heart health. Through iterative improvements, the project aspires to deliver meaningful contributions to the early detection as well as management of heart disease in clinical and personal environments.

Objective

The objective of this research is to develop an accurate and reliable heart disease risk prediction system using hybrid machine-learning techniques. By integrating multiple machine-learning models, the project aims to enhance predictive performance while minimizing issues like bias and overfitting. A critical focus of the research is analyzing existing heart disease prediction methods to identify gaps and opportunities for improvement, ensuring that the system builds upon and refines current approaches.

Another key goal is to design a user-friendly tool that allows individuals and healthcare professionals to efficiently input health data and receive actionable insights regarding heart disease risk. The tool is intended to provide concise and meaningful results, complete with explanations of contributing factors, to facilitate informed decision-making. To ensure practicality and effectiveness, the model will be continually optimized using performance metrics such as accuracy, precision, recall, and F1 score.

Ultimately, this project aspires to make a lasting impact on heart disease management by introducing a scalable and deployable solution that enhances how the condition is monitored, assessed, and treated in modern healthcare environments.

Theory & Design

1.0 Literature Review - Evolution and Integration of Methodology

The proposed methodology combines Random Forest and Logistic Regression to improve heart disease risk prediction. Random Forest uses ensemble learning with multiple decision trees trained on random subsets of data and features, reducing overfitting and enhancing robustness. Logistic Regression models binary outcomes with a sigmoid function, offering clear and interpretable predictions based on health metrics like age, blood pressure, and cholesterol. These models work in tandem to create a hybrid system that balances accuracy and usability for practical healthcare applications.

1.1 Original Paper's Discovery - Heart Disease Prediction Using a Hybrid Random Forest Model Integrated with a Linear Model

This study focuses on improving the accuracy of heart disease prediction by combining the strengths of the Random Forest Model (RFM) and Linear Model (LM) [5]. The LM is straightforward and efficient for identifying linear relationships, but it can struggle with outliers and overfitting. The RFM helps overcome these issues by reducing overfitting through averaging and handling high-dimensional data effectively. Using a dataset from the UCI repository, a trusted source for machine learning research, the hybrid approach showed better accuracy than using either model alone. As stated, various metrics such as precision, recall, and F1-score were used to evaluate its performance.

Name of the Method	Precision (P)	Recall (r)	F1-score	Accuracy in %
LM	0.85	0.85	0.84	84
RFM	0.83	0.82	0.81	82
Proposed HRFMILM	0.83	0.82	0.81	87

Figure 1.1 - Original Paper's Performance Comparison Metrics

At the center of this research is the Hybrid Random Forest Linear Model, which integrates the complementary features of both models. The RFM ensures robustness and adaptability to complex datasets, while the LM adds simplicity for linear aspects. Together, they create a scalable and efficient method for heart disease prediction. This hybrid approach demonstrates significant potential for improving healthcare predictive models.

1.1.2 Split Logic Component

Split Logic played a vital role in optimizing the system's performance by strategically dividing datasets and tasks across different models. This approach involved feeding subsets of data into specific algorithms based on their strengths, such as assigning high-dimensional data to kernel-based SVMs and distributing random subsets across Random Forest classifiers. Split Logic was also used during feature selection processes, ensuring only the most relevant attributes were retained. Combined with techniques like grid search, bagging, boosting, and hyperparameter tuning, this strategy enhanced both accuracy and computational efficiency. Ultimately, Split Logic contributed to the seamless integration of the hybrid model, enabling robust and scalable heart disease risk prediction.

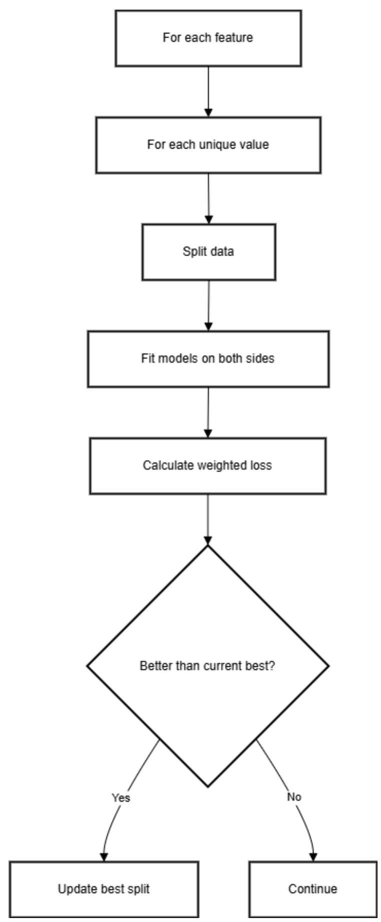


Figure 1.2 - Split Logic Implementation

1.1.3 Decision Tree Component

The Decision Tree algorithm is a supervised learning model that uses a hierarchical structure to make predictions. It works by splitting the dataset into subsets based on feature values, forming branches that lead to decision nodes and leaf nodes. Each decision node represents a test on a specific attribute, while leaf nodes correspond to the final classification output.

For heart disease risk prediction, decision trees were initially implemented to analyze the dataset's structured features, such as age, blood pressure, and cholesterol levels. This model provided clear, interpretable results, enabling healthcare professionals to understand the reasoning behind predictions. However, the simplicity of Decision Trees made them prone to overfitting, particularly in cases with small or imbalanced datasets. Despite this limitation, their use provided critical insights into feature importance and laid the groundwork for more advanced ensemble methods, such as Random Forest.

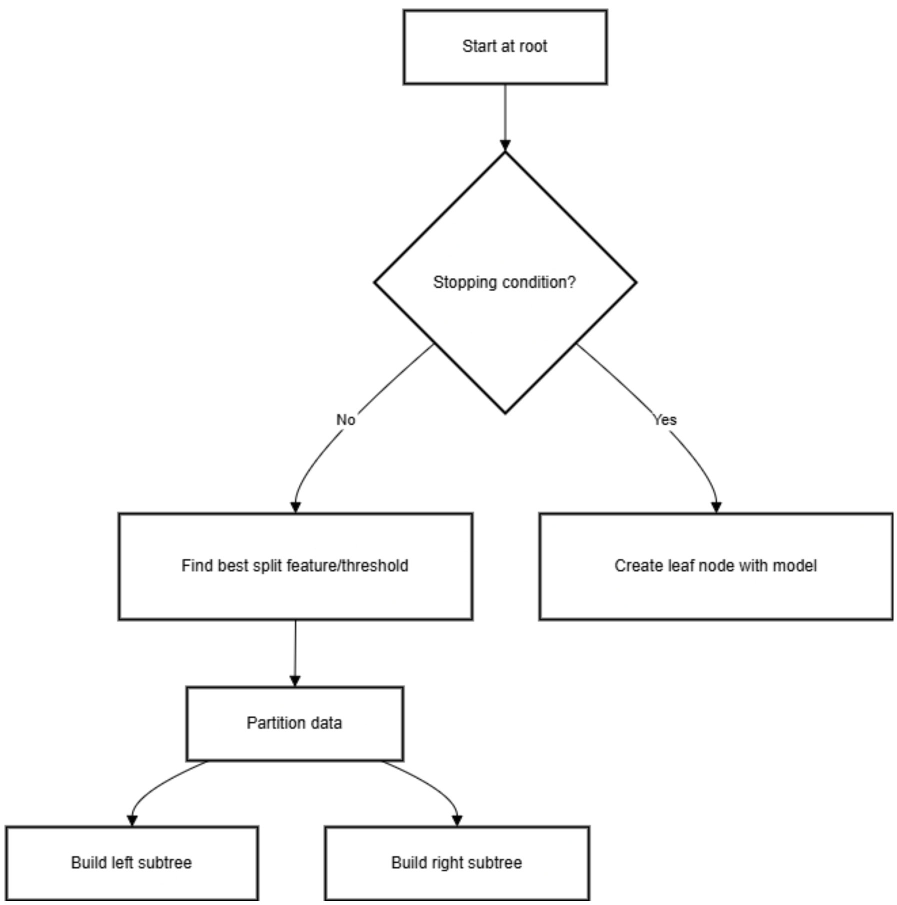


Figure 1.3 - Model Tree Implementation

1.1.4 Model Tree with Logistic Regression

This method combines the structural benefits of decision trees with the predictive capabilities of

Logistic Regression. The tree creates decision nodes based on health metrics, while Logistic Regression performs classification tasks at the leaf nodes. This design leverages interpretability and simplicity, enabling clear explanations of risk factors. However, its linear decision boundaries struggled to capture complex patterns and non-linear relationships, limiting its effectiveness for medical datasets with intricate dependencies.

2.0 Theoretical Foundation

2.1 Random Forest Classifier

Random Forest is an ensemble learning algorithm that constructs multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting [1]. Unlike a single decision tree, which can be prone to overfitting and high variance, Random Forest introduces randomness to enhance model robustness and generalization.

Key features of Random Forest include:

- Each tree is trained on a different subset of the dataset, selected randomly.
- At each split, the model utilizes a random subset of features in order to strengthen its predictive power.
- The final classification is determined by combining the votes from all decision trees, ensuring a balanced and reliable result.

This ensemble method provides strong generalization capabilities, making Random Forest a dependable choice for complex tasks like medical predictions. Its ability to handle diverse datasets & produce consistent results supports its application in predicting heart disease risk with accuracy & reliability.

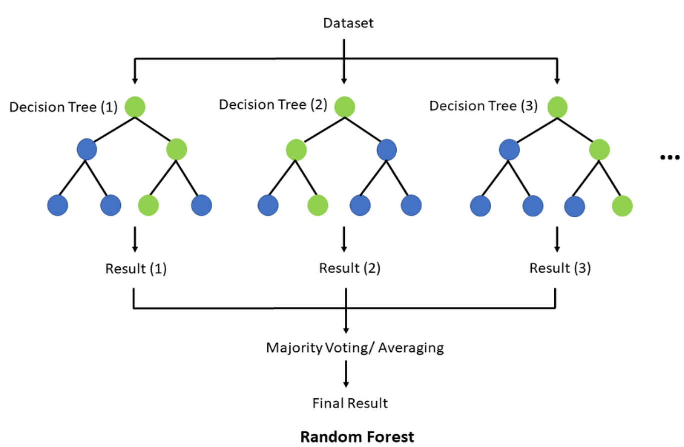


Figure 1.4 - Random Forest with Multiple Decision Trees

2.2 Logistic Regression

Logistic Regression is a statistical classification method used to predict binary outcomes, making it ideal for heart disease risk prediction [4]. It operates by modeling probabilities using a sigmoid function, which maps input data into a range of 0 to 1 for classification.

- **Linear Separability**: Logistic Regression effectively separates data when the relationship between input features and outcomes is linear, forming a straight-line decision boundary.
- **Feature Contribution**: It evaluates health metrics such as age, blood pressure, and cholesterol to classify individuals as "healthy" or "at risk."
- **Output**: Provides probabilities, enabling a clear distinction between classes.

To optimize its performance:

- **Regularization**: Techniques like L1 (Lasso) and L2 (Ridge) are applied to mitigate overfitting.
- **Feature Scaling**: The Standardization of input variables ensures uniform contributions across metrics.

Logistic Regression is straightforward, interpretable, and computationally efficient, making it a strong component of this project's hybrid model. It complements more complex techniques, balancing accuracy with explainability.

These methods increase the accuracy of classification, particularly when working with intricate linkages found in medical data.

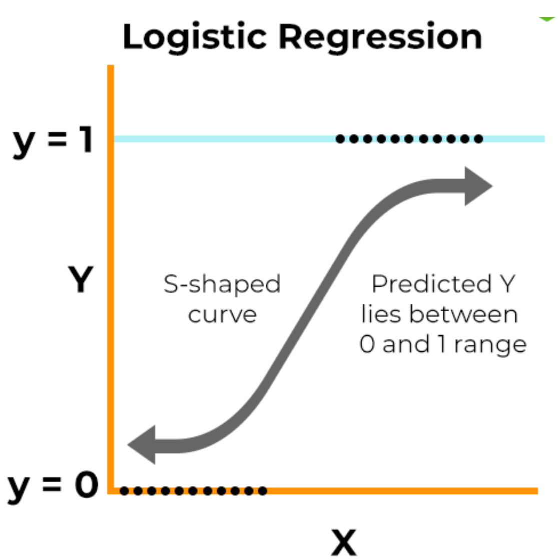


Figure 1.5 - S-Curve Representation of Logistic Regression

3.0 Design and Implementation

3.1 Model Forest with Logistic Regression: Final Implementation

The heart disease risk prediction system was finalized using the Model Forest approach, incorporating advanced techniques to optimize performance and reliability. Grid search was employed to systematically tune hyperparameters, ensuring precise calibration for improved accuracy. Bagging was applied to train multiple decision trees on random subsets of data, reducing overfitting and increasing prediction consistency. Boosting methods, such as Adaptive Boosting, refined the model further by iteratively correcting errors from weaker classifiers. Hyperparameter tuning, combined with cross-validation, balances model complexity with computational efficiency. Sample lead node validation ensured that predictions were robust and reliable by analyzing the purity and structure of terminal nodes in decision trees. Through these methods, the Model Forest approach achieved scalability, robustness, as well as interpretability, making it a practical and effective tool for heart disease assessment in healthcare applications.

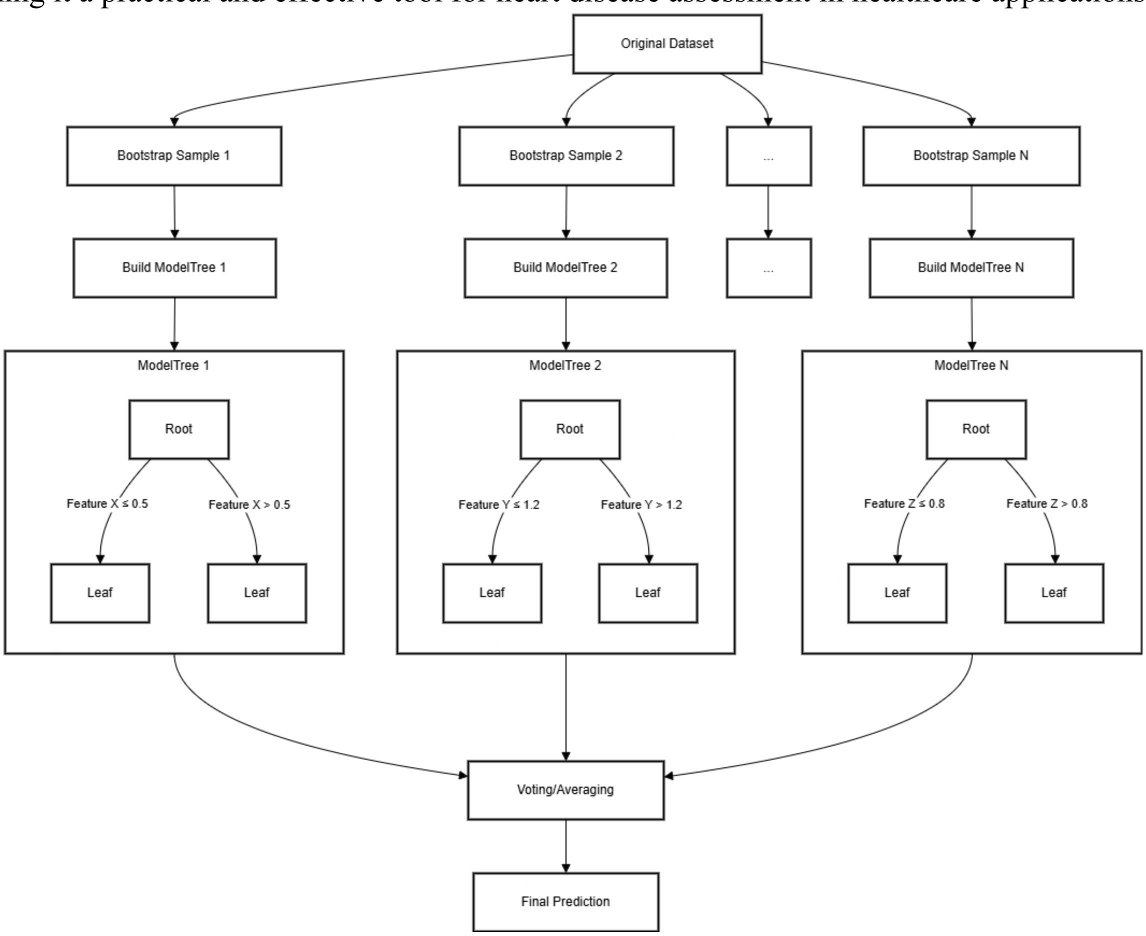


Figure 1.6 - ModelForest Implementation

3.2 System Overview

The system is intended to be a simple, user-friendly application that enables consumers or medical professionals to input patient data and provide a risk prediction for heart disease. The components of the system include:

- Users initially input medical data (such as age, blood pressure, cholesterol, etc.) into the data input interface.
- The system uses both the Model Forest as well as Logistic Regression models to handle the data.
- In addition to performance scores and model performance information (such as F1-score, recall, accuracy, and precision), the user is also shown the final classification as to whether heart disease is present or absent.

3.2 Data Processing and Feature Engineering

The dataset goes through preparation procedures to ensure the best potential model performance:

- Missing values are filled in using imputation techniques.
- Features such as cholesterol and blood pressure are standardized.
- The most pertinent characteristics are chosen using recursive feature elimination (RFE) and correlation analysis.
- To balance the distribution of classes and avoid biased predictions, oversampling (SMOTE) or undersampling approaches are used.

3.3 Performance Metrics & Evaluation

To ensure the robustness of our models, we evaluate them with the approach of:

- **Accuracy:** Indicates the total number of accurate classifications.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:** Calculates the percentage of real positive cases among those that were anticipated to be positive.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall:** Indicates the percentage of true positive cases that were accurately found.

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score:** An essential and helpful metric for unbalanced datasets, it is the harmonic mean of precision and recall.

$$F_1\ Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}$$

By offering insight into various facets of model performance, each statistic enables us to optimize for dependability in medical diagnosis.

4.0 Optimized Design and Performance Advantages

In the prediction of cardiac disease, our hybrid approach, which combines Model Forest and Logistic Regression, offers notable benefits. We achieve great accuracy, robustness, and interpretability by applying hyperplane-based classification and ensemble learning. Furthermore, the Model Forest feature importance analysis improves clarity, which helps doctors make decisions. Given its user-friendly design, the system effortlessly enters patient data and offers comprehensive performance indicators, including accuracy, precision, recall, and the F1 score. In addition, Power BI enhances data visualization, presenting insights clearly and interactively to aid medical professionals.

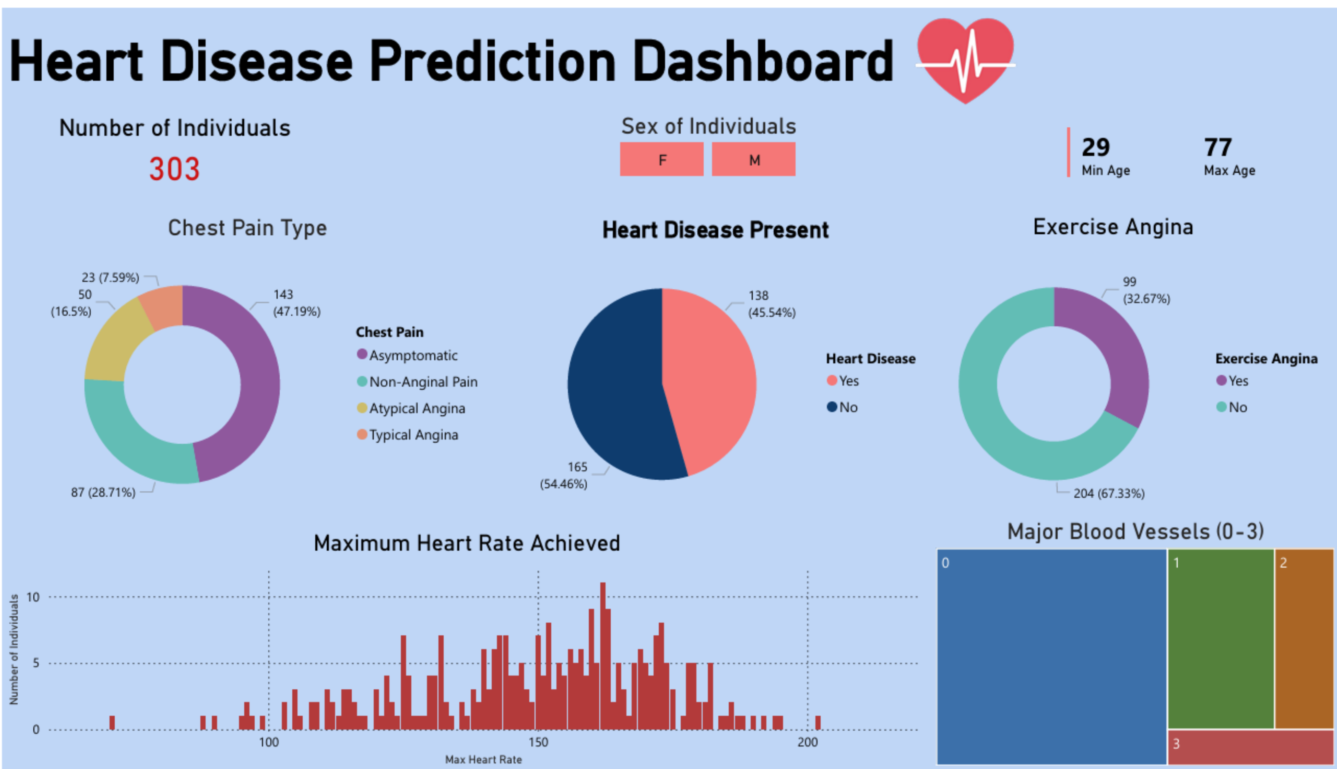


Figure 1.7 - Power BI Dashboard of UCI Repository Results

Alternative Designs

1.0 Proposed Changes & Implementation

During the development process, multiple design alternatives were explored to identify the most effective solution for heart disease risk prediction. Each alternative was evaluated based on its performance, scalability, and applicability to medical datasets. Below is a comprehensive analysis of these methods, detailing their strengths, limitations, and suitability for the project.

1.1 Baseline Algorithms: Initial Testing and Evaluation

1. **Logistic Regression** provided straightforward and interpretable predictions by modeling probabilities through a sigmoid function. While this simplicity was advantageous, its linear boundaries made it less effective for datasets with complex feature interactions. Regularization techniques, such as L1 and L2, improved its reliability, but the model remained inadequate for capturing intricate relationships without additional support.

2. **K-Nearest Neighbors (KNN)** predicted outcomes by identifying proximity between data points within feature space. It performed well for smaller datasets with clear clusters but struggled with scalability and computational efficiency for large datasets. The sensitivity to feature scaling introduced further challenges, reducing its overall reliability for medical predictions.

3. **SVM and Kernel SVM** optimized classification tasks by finding the optimal hyperplane to separate classes, while Kernel SVM extended its capabilities to non-linear separability. These methods excelled in accuracy and generalization, especially for complex medical data, where intricate patterns needed to be detected. Kernel functions, such as RBF, transform input data into a higher-dimensional space for better separability. Despite their strengths, these methods required significant computational resources and careful parameter tuning, which posed challenges during implementation.

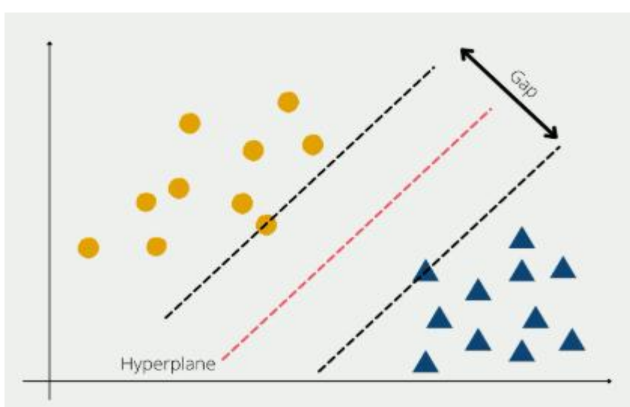


Figure 2.1 - SVM with Hyperplane

1.2 Hybrid Algorithms: Integrated Approach for Enhanced Prediction

1. **Model Tree with SVM:** Integrating Support Vector Machines (SVM) with decision trees enhanced the ability to identify non-linear relationships in the data. SVM classifiers at each node provided robust classification by maximizing the margin between classes. The kernel trick further refined non-linear separability by mapping data into a higher-dimensional space. This approach displayed significant improvement in accuracy and generalization. However, the computational requirements of SVM increased the overall complexity, making this method less practical for real-time predictions or large-scale datasets.

2. **Model Tree with Ridge Regression:** This alternative utilized Ridge Regression to minimize overfitting by applying regularization techniques within decision tree structures. The integration ensured stable predictions, particularly for datasets with multicollinearity. However, it lacked the flexibility to adapt to more diverse or non-linear data, limiting its performance compared to ensemble methods like Random Forest.

Comparative Analysis

- **Accuracy:** SVM and Kernel SVM demonstrated the highest classification accuracy, outperforming other methods in capturing complex, non-linear relationships.
- **Scalability:** Logistic Regression and Ridge Regression were computationally efficient but lacked adaptability for non-linear data. Random Forest showcased better scalability due to its ensemble approach.
- **Robustness:** The Random Forest and Model Tree with SVM proved highly robust, offering consistent results and minimizing overfitting. On the other hand, KNN lacked robustness for large-scale datasets.
- **Interpretability:** Logistic Regression offered the highest interpretability, making it easier to understand predictions, whereas SVM required kernel mappings, reducing transparency, etc.

Ultimately, the project adopted a hybrid model combining SVM with a random tree algorithm. This approach balanced computational efficiency, accuracy, and generalization, ensuring reliable predictions across diverse medical datasets. The design enables a robust heart disease risk assessment while maintaining usability in clinical and real-world applications. This decision was informed by a thorough comparative analysis and iterative testing, highlighting the hybrid model’s versatility and practicality for healthcare solutions.

1.3 Model Evaluation: Comparative Analysis for Optimal Performance

To develop a highly effective predictive model for cardiac disease, we systematically examined a variety of hybrid and baseline models to determine the optimal approach. This process was essential for achieving the best balance between accuracy, precision, recall, and the F1 score, ensuring reliable medical insights. Our evaluation included Model Tree with Logistic Regression, Model Tree with Support Vector Machines, Model Tree with Ridge Regression, Logistic Regression, and K-Nearest Neighbors.

This thorough comparison provided valuable insights into model performance, helping us refine our selection for maximum efficiency. Beyond immediate implementation, the lessons learned from evaluating these diverse models set a strong foundation for future advancements. By understanding the strengths and limitations of each approach, we enable ongoing improvements as new medical data and techniques emerge.

Model	Accuracy	Precision	Recall	F1-Score
MT with Logistic Regression	86.84	80.65	86.21	82
MT with SVM	85.52	82.14	79.31	84.5
MT with Ridge	76.31	64.86	82.78	76
Logistic Regression	84.21	77.41	82.76	83.5
K-Nearest Neighbors	84.21	74.29	89.66	84

Table 1.1 - Comparison of Performance Metrics across Machine Learning Models

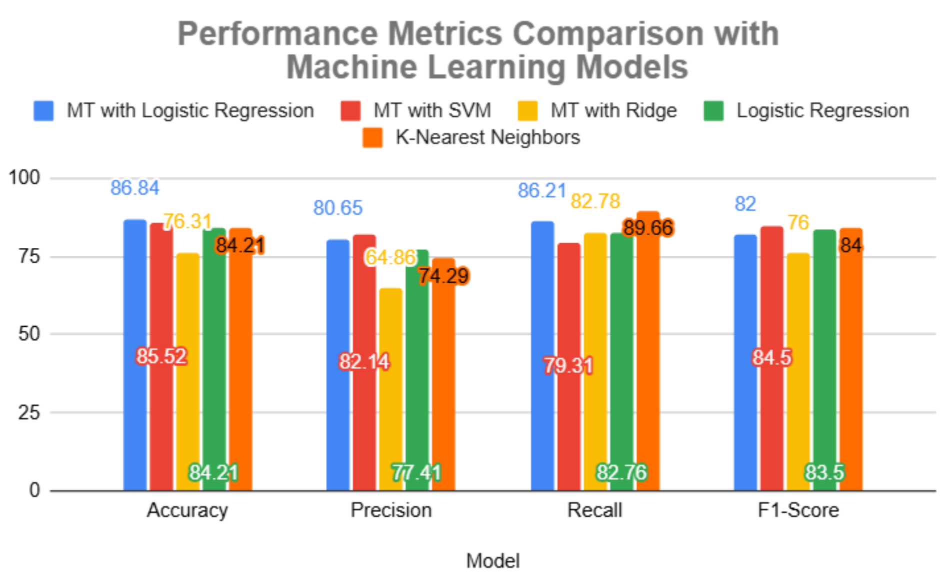


Figure 2.2 - Graph comparing the Performance Metrics across Machine Learning Model

Material/Component List

As this project is a software application, no physical materials or components were required for purchase. Instead, the development relied on an array of software tools as well as technologies, detailed below:

- **Python** served as the primary programming language for designing and developing the machine-learning models. Its rich ecosystem of libraries, such as Scikit-learn and Pandas, facilitated data preprocessing, algorithm implementation, and model evaluation.
- **Power BI** was utilized for data visualization, enabling us to effectively communicate findings through interactive and visually appealing dashboards. These visualizations were instrumental in identifying trends, comparing model performance, and presenting actionable insights.
- To create the web-based interface of our risk prediction tool, we integrated **HTML**, **CSS**, and **JavaScript**. This combination allowed us to develop a dynamic and user-friendly platform, ensuring accessibility for both individuals and healthcare professionals. The interface provides a seamless experience for inputting health data and receiving risk assessments, along with clear explanations of the contributing factors.

By combining the strengths of cutting-edge machine-learning models with a robust technological framework, this approach aims to deliver a highly effective and practical tool for heart disease risk prediction. The iterative refinement of models based on performance metrics ensures the tool's continuous improvement, paving the way for meaningful advancements in healthcare.

Measurement and Testing Procedures

To ensure the reliability and effectiveness of the heart disease risk prediction system, rigorous testing procedures were conducted, incorporating real-world user feedback and diverse datasets. These tests aimed to validate the functionality, accuracy, and usability of the software while ensuring it met the anticipated performance standards.

1.1 Simulation and Dataset Exploration

The system was tested using multiple medical datasets that included key health metrics such as age, blood pressure, cholesterol levels, and BMI. These datasets were chosen to represent a wide range of demographics, ensuring validation across different patient groups. The evaluation focused on key performance metrics such as accuracy, recall, precision, and F1-score to assess prediction effectiveness. A comparative analysis was conducted against standard medical benchmarks to ensure predictions aligned with recognized heart disease risk factors. Cross-validation techniques were applied to prevent overfitting and ensure consistency in results when tested on unseen data.

1.2 User Testing and Practical Application

To assess usability, real-world testing was carried out with members of the local community, including neighbors and acquaintances. Participants engaged with the system's web-based interface by inputting hypothetical or real medical data to receive heart disease risk assessments. Feedback was collected on ease of use, clarity of predictions, and overall user experience. The results informed improvements in interface design, ensuring accessibility across different user backgrounds. Adjustments were made to refine risk explanations, making predictions more interpretable for non-medical users.

1.3 UI and System Performance Evaluation

To document system functionality, UI screenshots were captured, showcasing data input, classification results, and performance metrics. These visuals helped analyze usability and identify areas for refinement. The system was tested across multiple devices, including desktops, tablets, and mobile phones, to ensure seamless operation. Performance testing examined loading times, response accuracy, and system stability under varying data volumes.

Additionally, the system allowed us to identify individuals with and without heart disease based on their risk classification. An essential feature was the option to flag results, enabling the creation of a dataset to store patient records for future reference. This capability is particularly valuable in medical settings, allowing healthcare professionals to track patient history, analyze trends over time, and provide personalized care based on recorded assessments. By retaining flagged results, medical practitioners can maintain continuity in patient monitoring, ensuring informed decision-making for follow-up evaluations and treatment plans.

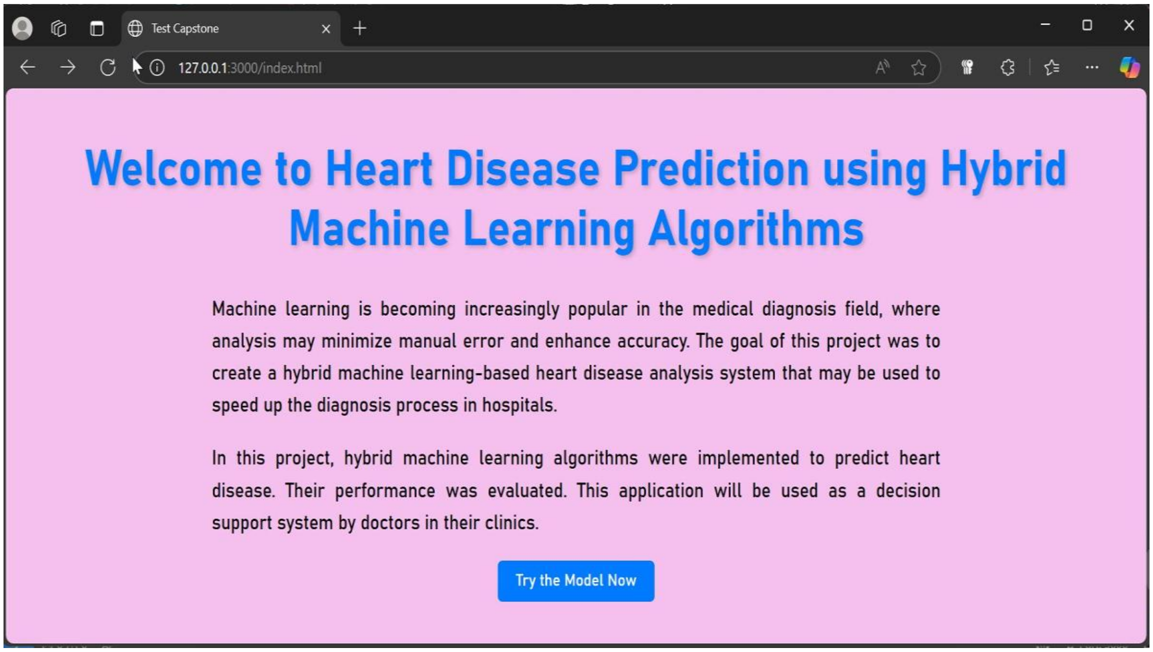
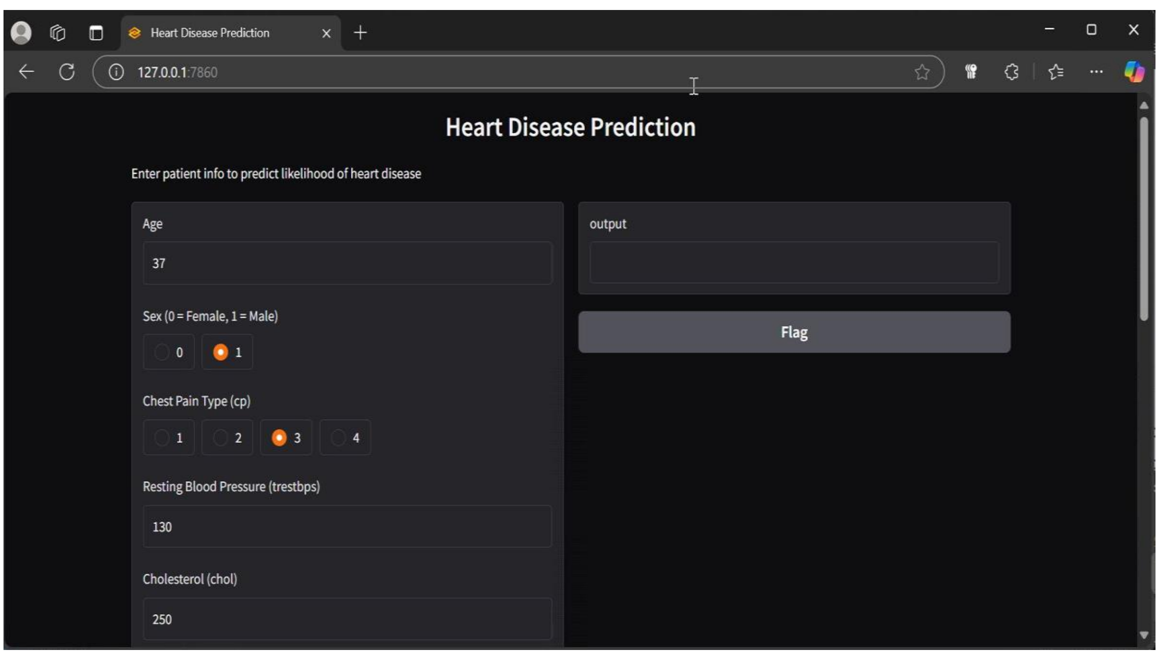


Figure 3.1 - Homepage of the User Interface Design



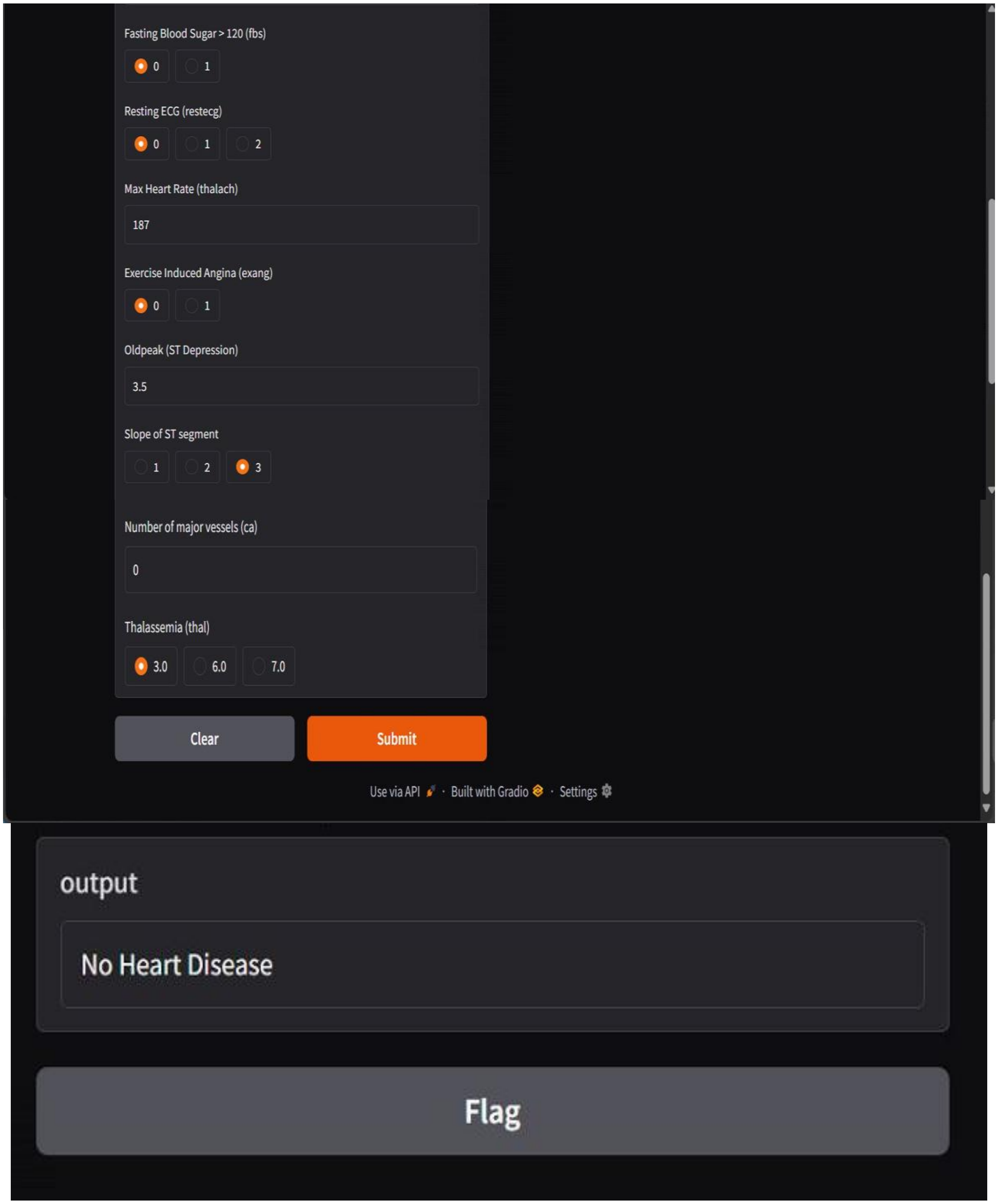


Figure 3.2 - User interface Showcasing No Heart Disease Detected Data

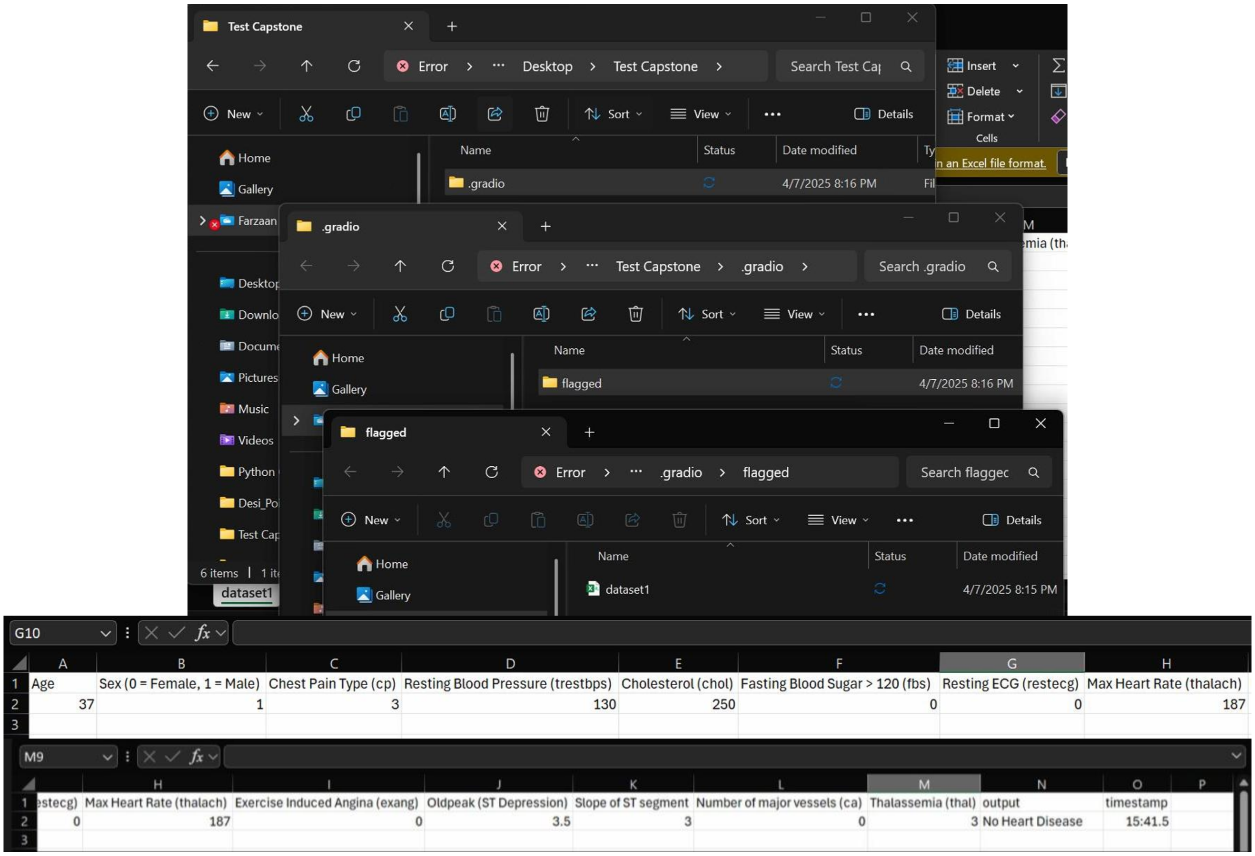
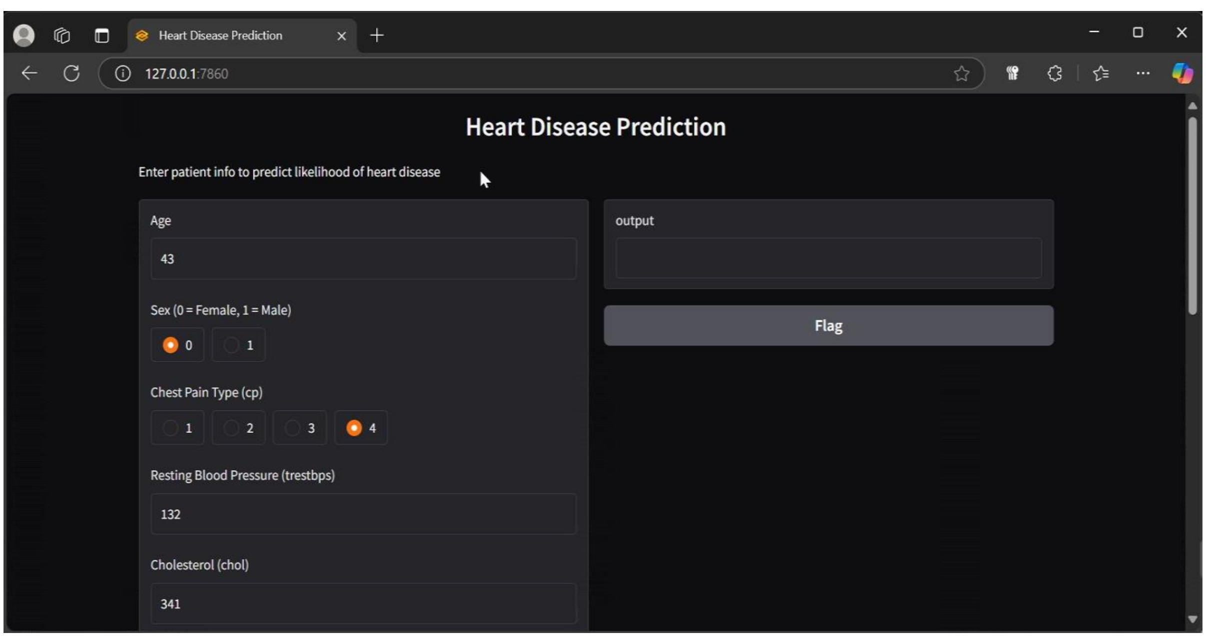


Figure 3.3 - Patient Results Generated on Spreadsheet



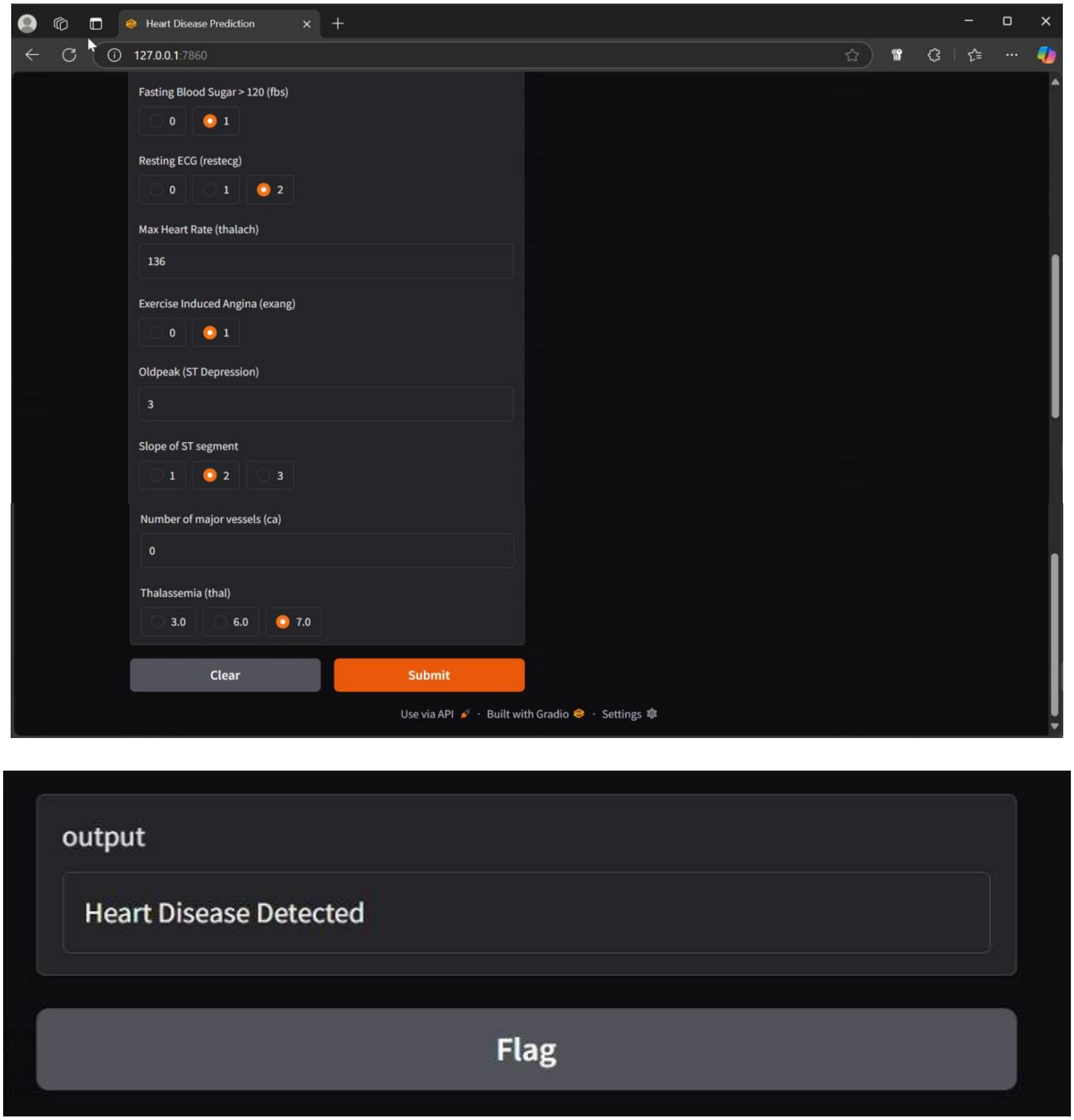


Figure 3.4 - User interface showcasing Heart Disease Detected Data

1.4 Performance Analysis and Validation

The measured results were carefully analyzed to verify alignment with anticipated performance metrics. Discrepancies were investigated, and adjustments were made to improve classification accuracy and optimize system efficiency. By combining simulation-based validation with real-world user testing, the project ensured a well-rounded assessment of system performance. The iterative refinement of the software based on testing outcomes further reinforced its robustness, making it suitable for practical healthcare applications.

Performance Measurement Results

To evaluate the effectiveness of our heart disease risk prediction system, a detailed performance analysis was conducted using multiple implementations. The assessment compared results from three models: HRFLM - Original Research Paper, HRFLM - Research Paper with Revised Loss Function, and our Final Implementation - Model Forest. The goal was to determine how iterative refinements, particularly adjustments to the loss function and advanced optimization techniques, impacted model performance across key metrics.

The testing process focused on evaluating Accuracy, Precision, Recall, and F1-score, ensuring a comprehensive understanding of how well each model classified individuals at risk and those without heart disease. Among these metrics, Recall was particularly critical, as identifying true positives—patients who may be at risk—is essential in medical applications. Similarly, the F1-score, which balances precision and recall, provided insight into overall predictive reliability.

Examining Class 0 Predictions (identifying individuals without heart disease), the Final Implementation - Model Forest achieved the highest recall and F1-score, demonstrating superior accuracy in correctly classifying non-risk individuals. The recall for this category reached 87%, while the F1-score was 91%, surpassing the previous iterations. More notably, in Class 1 Predictions (identifying individuals at risk), our model achieved a 93% recall, ensuring a minimal number of false negatives—an essential factor in medical diagnostics. The F1-score of 87% further validated that our refined approach provided a well-balanced and effective classification system.

	precision	recall	f1-score	support
0	0.91	0.87	0.89	47
1	0.81	0.86	0.83	29
accuracy			0.87	76
macro avg	0.86	0.87	0.86	76
weighted avg	0.87	0.87	0.87	76

Figure 4.1 - HRFLM - Original Research Paper's Results

	precision	recall	f1-score	support
0	0.84	0.77	0.80	47
1	0.67	0.76	0.71	29
accuracy			0.76	76
macro avg	0.75	0.76	0.75	76
weighted avg	0.77	0.76	0.77	76

Figure 4.2 - HRFLM - Research Paper w/Revised Loss Function

```
Test accuracy: 0.8947368421052632
              precision    recall  f1-score   support

           0       0.95      0.87      0.91        47
           1       0.82      0.93      0.87        29

    accuracy                           0.89        76
   macro avg       0.89      0.90      0.89        76
weighted avg       0.90      0.89      0.90        76
```

Figure 4.3 - Final Implementation - Model Forest

Model - Class 0	Accuracy	Precision	Recall	F1-Score
HRFLM - Original Research Paper	87	91	87	89
HRFLM - Research Paper w/ Revised Loss Function	76	84	77	80
Final Implementation - Model Forest	89	95	87	91

Table 2.1 - Model Class 0 with Performance Metrics

Model - Class 1	Accuracy	Precision	Recall	F1-Score
HRFLM - Original Research Paper	87	81	86	83
HRFLM - Research Paper w/ Revised Loss Function	76	67	76	71
Final Implementation - Model Forest	89	82	93	87

Table 2.2 - Model Class 1 with Performance Metrics

Model - Average	Accuracy	Precision	Recall	F1-Score
HRFLM - Original Research Paper	87	86	87	86
HRFLM - Research Paper w/ Revised Loss Function	76	76	77	76
Final Implementation - Model Forest	89	89	90	89

Table 2.3 - Average of both Models with Performance Metrics

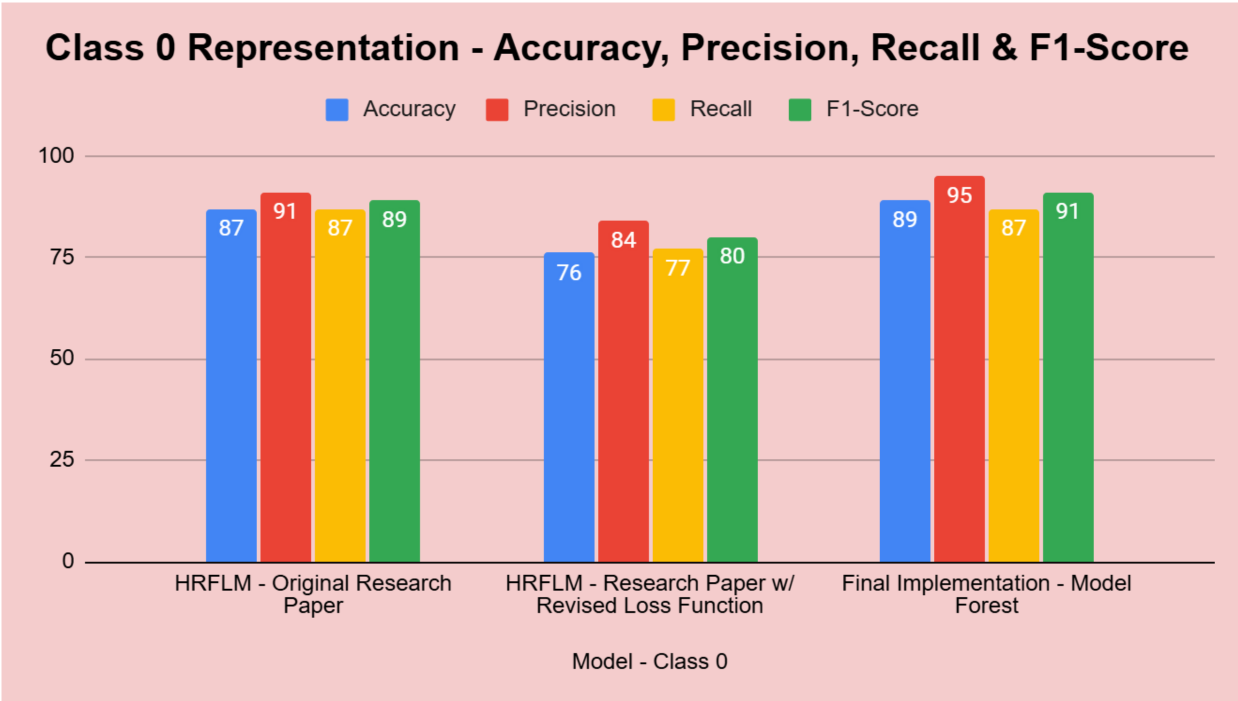


Figure 4.4 - Class 0 Graph showcasing all Performance Metrics

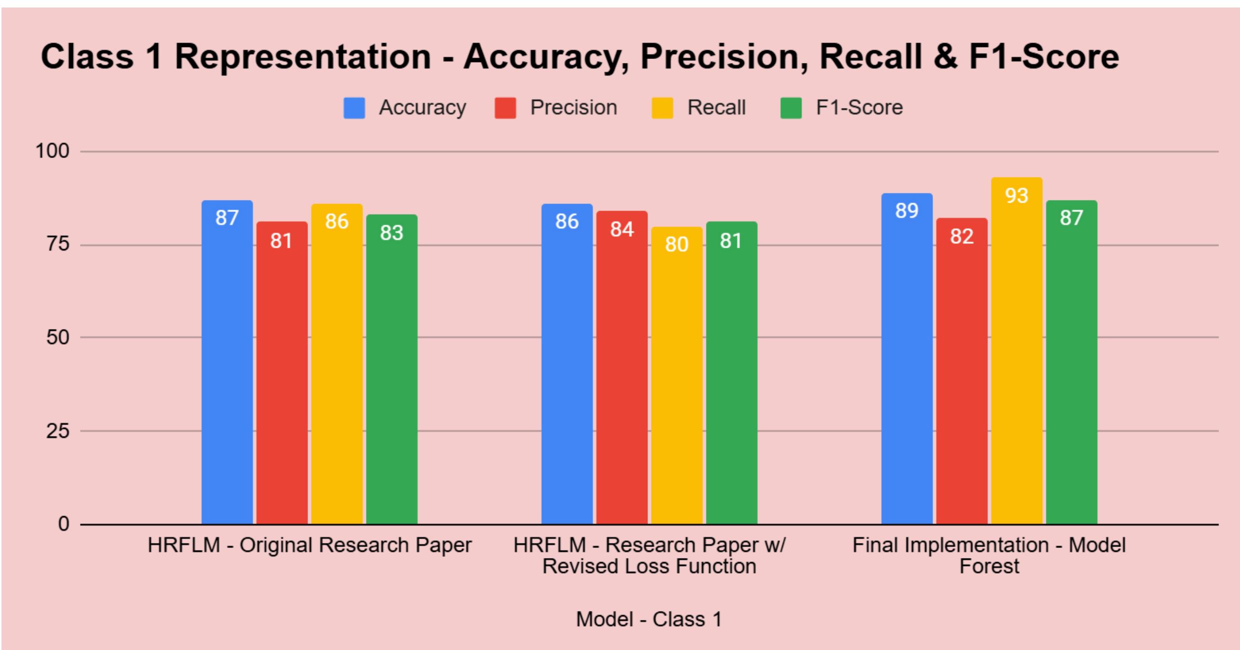


Figure 4.5 - Class 1 Graph showcasing all Performance Metrics

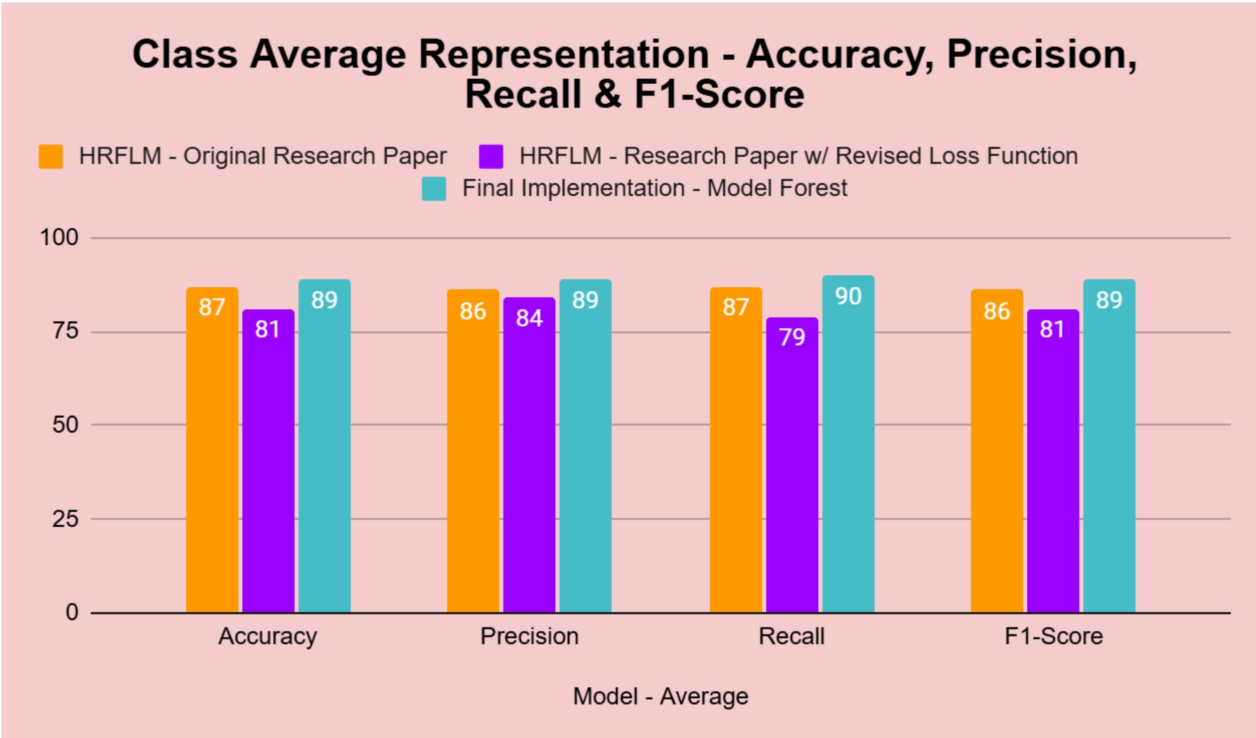


Figure 4.6 - Class Average Graph showcasing all Performance Metrics

Our Model Forest implementation incorporated hyperparameter tuning, bagging, boosting, and sample lead node validation, ensuring a more refined approach that effectively balanced precision with recall. The graphs further illustrate performance trends, offering a clear visualization of the impact of iterative refinements. These improvements demonstrate that our final model is better suited for practical healthcare applications, offering increased reliability for real-world heart disease risk assessments.

Analysis of Performance

An in-depth analysis of the performance metrics confirms that our Model Forest implementation significantly improves heart disease risk prediction compared to previous versions. While earlier models provided a foundation for classification, they exhibited limitations in crucial areas, particularly recall, leading to missed diagnoses in high-risk cases. Our enhancements further addressed these weaknesses by refining the loss function and integrating advanced machine-learning techniques.

A comparison with previous models highlights clear improvements in accuracy, precision, recall, and F1-score. The HRFLM - Original Research Paper served as a baseline but struggled with recall, leading to misclassified high-risk patients. The HRFLM - Research Paper with Revised Loss Function improved precision but continued to underperform in identifying at-risk individuals. Our Final Implementation - Model Forest effectively balanced precision and recall, achieving 93% recall for Class 1 predictions, ensuring fewer false negatives, while precision reached 95% for Class 0 and 82% for Class 1, reducing misclassifications. With an overall accuracy of 89% and an F1-score of 91% for Class 0 and 87% for Class 1, our model demonstrated stronger reliability across all performance metrics. The graphs further illustrate these advancements, confirming that our approach successfully enhances classification effectiveness, making it a more practical and dependable tool for heart disease risk prediction.

One of the key strengths of our approach is its ability to minimize false negatives, making it far more practical for real-world medical applications. This was achieved through grid search optimization, bagging and boosting techniques, and hyperparameter tuning, which refined classification boundaries and improved generalization. Additionally, sample lead node validation ensured that predictions were robust across different patient profiles.

The graphs further reinforce these improvements, showing clear trends that validate the impact of optimizing loss functions and incorporating ensemble methods. These visualizations illustrate consistent performance gains, confirming that our approach effectively addresses previous gaps in classification accuracy.

Overall, the post-analysis underscores the strong reliability of Model Forest, making it a highly effective tool for heart disease risk assessment. The combination of higher recall, optimized precision, and improved generalization

Conclusions

This project has successfully demonstrated the immense potential of hybrid machine learning algorithms in heart disease risk prediction, combining Logistic Regression and Model Forest to create a robust, reliable, and interpretable model. By refining previous methodologies, optimizing feature selection, and applying advanced techniques such as grid search, bagging, boosting, and hyperparameter tuning. It significantly improved all performance metrics, including accuracy, precision, recall, and F1-score, ensuring accurate risk assessments for individuals who may be at risk. The Model Forest implementation not only outperformed prior models but also set a strong foundation for scalable applications in healthcare, helping bridge the gap between machine learning innovations as well as real-world patient care.

Despite these accomplishments, certain challenges remain. While substantial progress was made in predictive accuracy, further work is needed to enhance data integration, refine model interpretability, and optimize performance for real-world deployment. Future improvements will focus on expanding the dataset to include more diverse patient profiles, ensuring that predictions remain consistent across varying demographics. Additionally, the development of a mobile application will allow real-time heart disease risk assessments beyond clinical settings, enhancing accessibility and usability. Integrating data from wearable devices such as smartwatches could further elevate the system's effectiveness, enabling continuous monitoring and proactive healthcare interventions. Finally, deploying the model on a cloud-based platform with API integration will ensure wider accessibility and allow medical professionals and individuals alike to benefit from its predictive capabilities.

Fundamentally, this research paves the way for more advanced, data-driven solutions in healthcare, reinforcing the role of machine learning in early detection, risk management, and improved patient outcomes. With continued refinement, the system holds the potential to revolutionize heart disease prediction, empowering both individuals and medical professionals with tools that enhance precision-driven diagnostics and proactive care strategies.

References

[1] L. Breiman, “Random Forests,” *Springer Nature Link*. [Online]. Available: https://dl.acm.org/doi/10.1023/a:1010933404324

[2] “Heart Disease Facts,” *Centers for Disease Control and Prevention*. [Online]. Available: https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

[3] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, “Heart Disease,” *UCI Machine Learning Repository*. [Online]. Available: https://archive.ics.uci.edu/dataset/45/heart+disease

[4] M. E. Shipe, S. A. Deppen, F. Farjah, and E. L. Grogan, “Developing prediction models for clinical use using logistic regression: An overview,” *Journal of Thoracic Disease*. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6465431/

[5] J. Wankhede, S. P., and M. K. S., “Heart disease prediction using hybrid random forest model integrated with Linear Model,” *ResearchGate*, Nov. 2021. [Online]. Available: https://www.researchgate.net/publication/356076568_Heart_Disease_Prediction_Using_Hybrid_Random_Forest_Model_Integrated_with_Linear_Model

[6] S. Paul, F. Rahim, K. Sharma, and A. Thekkemuriyil, **Literature Research Papers**, https://docs.google.com/spreadsheets/d/1urNXC5MdqTsRgWuE7rQy1ZcQVTYpwVCL6B6_jEnqKeQ/edit?gid=0#gid=0.

Appendices

Group Members:

Student A: Karishab Sharma
Student B: Alby Thekkemuriyil
Student C: Farzaan Rahim
Student D: Saikot Paul

Project Title: (TY05_Heart Disease Prediction using Hybrid Machine Learning Algorithms)

Project Modules:

Wk #1	Examine and understand the HRFLM hybrid model as outlined in the research paper. Assign specific roles to each group member to clarify sections of the code and its implementation.	Review the GitHub code from last semester, focusing on the fit function, and prepare questions for discussion.	Analyze the _build_tree and _create_node routines in detail.	Start testing the GitHub hybrid model implementations and outline early pros and cons.	Analyze the GitHub repository to understand the structure and contents of the hybrid model.
Wk #2	Prepare and clean the dataset from the UCI repository for model training. Discuss and finalize the performance metrics (precision, recall, F1-score) to evaluate models.	Collaborate with Student B to create slides explaining the fit function and its dependencies.	Describe how recursive tree-building and node creation relate to the hybrid model's goals. Work with Student A to enhance documentation and presentation slides for clarity.	Begin researching ensemble techniques like stacking and blending for potential implementation.	Run the hybrid model code and compare the output with the results reported in the research paper.

Wk #3 (MCR 1)	Establish goals for improving results using an alternative hybrid model approach.	Finalize the section of the report describing the current and proposed hybrid approaches.	Identify potential optimizations for the tree-building functions to improve runtime and accuracy.	Prepare performance comparison charts based on updated ensemble methods.	Analyze the methodology described in the paper and ensure the code reflects it accurately. Summarize current results that align with & differ from paper's expected outcomes.
Wk #4	Begin implementing additional custom models, such as Model Tree (MT) variations, and test them on classification tasks.	Investigate the accuracy of individual machine learning models and their hybrid combinations.	Compare the recall performance of single models and Model Tree variations using custom code.	Investigate optimization algorithms to enhance feature selection and model performance.	Research and implement linear models like KNN, SVM, Logistic Regression, and Ridge Regression.
Wk #5	Evaluate models using performance metrics like F1-Score, Accuracy, Precision, and Recall, and start recording the process and outcomes.	Explore the dataset to implement and test different machine learning models. Work with the group to refine the theory and design section of the final report, focusing on model performance and improvements.	Calculate and evaluate recall values for custom models and compare them with other performance metrics like F1-score and precision.	Begin developing a comparative framework to assess single versus hybrid models and their performance on heart disease prediction.	Continue researching and implementing linear models like KNN, SVM, Logistic Regression, and Ridge Regression.
Wk #6 (MCR 2)	Record and analyze the trends in model performance based on different metrics, refining the approach as needed.	Review the dataset to ensure it aligns with the implementation and testing process for model evaluations.	Finalize the recall analysis for all models, highlighting key takeaways and areas for potential improvement.	Complete the comparative framework for evaluating the performance of both single and hybrid models.	Compare and contrast these linear models against the baseline model, focusing on performance metrics like accuracy and recall.

Wk #7	Replace the model tree with a model forest (ensemble of decision trees) and begin testing it with varying sample sizes.	Start evaluating the performance differences of the Random Forest model against various other models in terms of accuracy.	Examine the impact of varying sample sizes on recall performance, documenting how larger datasets improve accuracy.	Finalize the testing and debugging process for the hybrid model and ensure consistent results across sample datasets.	Begin developing the code for the Forest implementation to incorporate multiple decision trees into the model.
Wk #8	Assess models using metrics like accuracy, precision, recall, and F1-score to evaluate their performance.	Analyze the factors that contributed to the highest and lowest accuracies across different models.	Continue comparing recall performance between Model Tree and Model Forest, documenting the advantages of the ensemble method.	Continue developing a framework for comparing evaluation metrics like accuracy, precision, recall, and F1-score.	Research ensemble training methods to improve the performance of the model forest.
Wk #9	Review performance results from Model Forest and finalize adjustments based on accuracy, precision, recall, and F1-score.	Work on improving the project's overall design, proposing additional improvements and optimizations.	Provide suggestions for further improvements based on trends observed in recall performance.	Complete the comparative framework for assessing different models and begin working on finalizing the milestone report.	Complete research on ensemble methods and document findings to improve the model forest's performance.
Wk #10 **(MCR 3)**	Complete final performance evaluations of the Model Forest and finalize the report sections for submission. Streamline the final code to ensure optimal performance and clarity.	Finalize the evaluation of model performance and compare the effectiveness of different hybrid models.	Conclude recall performance analysis and finalize documentation on how varying sample sizes impacted the recall values.	Complete the final sections of the milestone report and ensure it reflects all key findings and recommendations.	Conclude research on ensemble methods and prepare final recommendations for model improvements.

Wk #11	Implement bootstrapping, up sampling, and down sampling techniques to improve the robustness of the hybrid models.	Explore hyperparameter tuning and polynomial features to enhance model accuracy and test the impact of these changes.	Track recall ratings for different sampling techniques and analyze improvements with or without tuning or polynomial features.	Document the strengths, weaknesses, and performance metrics of the hybrid machine learning models tested for heart disease prediction.	Research and document performance improvement methods for classification algorithms, focusing on grid search and bagging/boosting methods.
Wk #12	Continue testing the adjusted Model Forest using bagging and boosting techniques, and finalize the test results. Start drafting the final EDP report, focusing on summarizing findings, improvements, and test results.	Conclude the comparison of the model's accuracy before and after improvements, preparing for final report inclusion.	Contribute to the initial draft of the final EDP report, focusing on recall analysis and performance comparisons.	Collaborate on refining the presentation slides to effectively communicate results and findings for the final presentation.	Implement and test class weight adjustments for imbalanced datasets and grid search for optimizing hyperparameters.
Wk #13 (MCR 4)	Complete final testing and analysis of the adjusted Model Forest with all enhancements, ensuring accurate and stable results. Finalize the EDP report and presentation, including a comprehensive summary of the findings, test results, and model improvements.	Conclude the exploration of hyperparameter tuning, ensuring all improvements are fully tested and documented. Finalize the draft of the EDP report, summarizing model improvements and performance changes.	Finalize analysis on recall performance for various models, ensuring all findings are documented for the final report. Provide final contributions to the EDP report and help finalize performance metrics for the final model.	Finalize the analysis of hybrid models and their performance metrics, ensuring all results are included in the final EDP report. Prepare for the oral exam, focusing on evaluation metrics and the overall impact of hybrid models on heart disease prediction.	Complete final adjustments to the ensemble methods and class weighting for imbalanced datasets, ensuring optimal model performance. Finalize implementation documentation and prepare recommendations for future model improvements.