**A Data Mining Approach for Heart Disease Prediction and Pattern Extraction**

***Apurba Bhowal [1], Md. Mostafijur Rahman [2], Ahsanul Haque Joar [3]***

***Supervised by***   
***Mohaimen Bin Noor***  
  
***22-46706-1@student.aiub.edu***  
***22-47161-1@student.aiub.edu***  
***22-47183-1@student.aiub.edu***

***Abstract:***Heart disease, particularly heart failure, poses a significant challenge to global health, affecting millions and demanding advanced solutions for early detection and management. This study explores the application of data mining techniques to predict heart failure, enhancing diagnostic precision and supporting better clinical decision-making. Extensive feature engineering was performed, incorporating clinically relevant attributes such as blood pressure categories and Body Mass Index (BMI), alongside standardized scaling for numerical features to ensure consistency. Various data mining algorithms, from Logistic Regression and Support Vector Classifier (SVC) to ensemble techniques like Random Forest and XGBoost, were evaluated for their pattern discovery capabilities. The study highlights the effectiveness of robust validation methodologies, including 10-fold cross-validation and hyperparameter tuning, in optimizing the performance of the models. These findings underscore the potential of data mining to revolutionize heart disease prediction and improve patient outcomes in real-world clinical settings.

This study repositions the analytical framework within the scope of data mining, highlighting knowledge discovery, pattern recognition, and feature-based insight extraction in healthcare datasets.

***Keywords:***  
heart disease classification, pattern extraction, data mining, Random Forest, feature selection and transformation, cross-validation, hyperparameter tuning, ensemble methods, healthcare analytics, cardiovascular health, early diagnosis.

***I. Introduction:***

Heart failure is a condition in which the heart is unable to pump enough blood to meet the body’s needs [1]. Cardio- vascular diseases have emerged as a significant global health concern, substantially impacting public health worldwide. Heart failure is a common and serious condition affecting millions worldwide. According to a recent state, heart failure disorders cause around 26 million population [2]. The causes of heart failure can be divided into two categories [21]. First related to the heart’s structure, such as a previous heart attack. Second related to the heart’s function, such as high blood pressure. Symptoms of heart failure can include shortness of breath, fatigue, and swelling in the legs and ankles [22]. Treatment options for heart failure include medications, lifestyle changes, and in some cases, surgery. Research [3] has shown that early detection and management of heart failure can improve quality of life and prolong survival.

Several data mining algorithms have been utilized for heart disease classification and pattern extraction, each demonstrating strengths and limitations. Logistic regression, known for its simplicity and interpretability, often underperforms when handling non-linear data relationships [4]. Decision trees are easy to understand but prone to overfitting, leading to poor generalization on unseen data [5]. Advanced techniques like Support Vector Machines (SVM) and neural networks have shown higher accuracy but demand significant computational resources and are difficult to interpret [6]. Ensemble data mining algorithms such as Random Forest and XGBoost have achieved impressive accuracy, with XGBoost reaching 97.57%, yet they face challenges like computational complexity, longer data preparation times, and potential overfitting [7], [8]. Additionally, the dependency on large, high-quality datasets and the need for explainable AI methods like SHAP remain critical barriers for clinical adoption [9].

The objective of this study is to use a tuned Random Forest classifier with 10-fold cross-validation to predict the presence or absence of a medical condition based on features such as age, blood pressure, heart rate, and biochemical markers, ensuring high accuracy and reliability.

Addressing these limitations, it is essential to develop data mining algorithms that strike a balance between accuracy, interpretability, and computational efficiency for effective deployment in healthcare. The current study focuses on developing a data mining approach for managing heart failure to improve patient health. Data Mining is highly involved in medical diagnoses and the healthcare industry [10]. Data mining has many applications in the medical field, including drug discovery, medical imaging diagnosis, outbreak prediction, and heart failure prediction. Data mining techniques can learn patterns from large medical data and perform predictive analysis. Data mining has many advantages compared to classical medical methods, such as saving time and costs, which helps improve diagnosis. Our prominent research contributions for heart failure detection using machine learning, based on the implemented methodology, are as follows:

• Feature Engineering and Selection: A correlation matrix was utilized to identify relationships between numerical attributes, providing insights into feature relevance. Additionally, categorical features were encoded and augmented with engineered features, such as BMI and blood pressure categories, to improve the model's predictive power. This comprehensive approach ensured the dataset was optimized for data mining approach data preparation. The inclusion of these engineered features, coupled with feature normalization using standard scaling, significantly improved the data mining algorithms performance by ensuring that numerical attributes were standardized and consistent.

• Robust Validation Methodology: To ensure the reliability of the results, the k-fold cross-validation technique was used. This approach validates the performance of the Cross Validation model across different data splits, ensuring its generalizability and robustness in real-world applications.

***II. Research Methodology:***

A. Dataset Description

The dataset used for our machine learning model was sourced from Kaggle [11] and contains 1319 records with various health-related attributes. It includes both numerical and categorical features that provide critical information about patients. The numerical attributes are Age, Heart rate, Systolic blood pressure, Diastolic blood pressure, Blood sugar, CK-MB, and Troponin, representing key health parameters. Additionally, there is a categorical attribute, Gender, encoded as binary values (0 or 1). The target attribute, Result, indicates the health outcome and is categorized as either "Positive" or "Negative," with 810 and 509 instances, respectively.

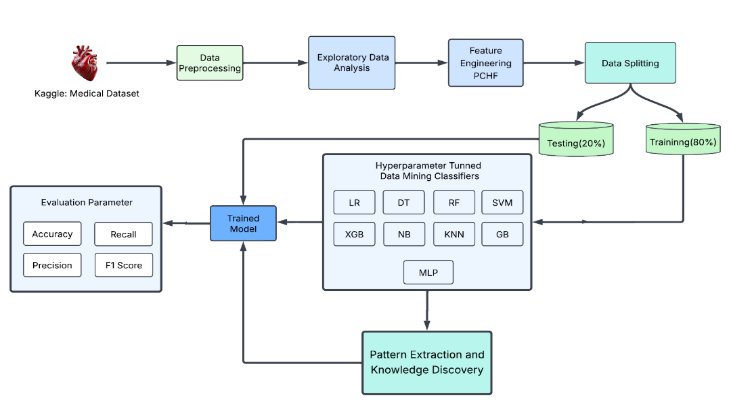


Figure 01: Workflow Diagram of Heart Disease Prediction Model [20].

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to uncover valuable patterns and insights within the dataset. Various visualizations, including histograms, box plots, and a correlation heatmap, were employed to understand feature distributions and relationships. The correlation heatmap revealed interesting trends, such as a notable correlation between Systolic Blood Pressure and Diastolic Blood Pressure, as well as some moderate relationships with other medical parameters. Importantly, features like Troponin and CK-MB showed a stronger connection with the target variable, Result, indicating their potential importance in predicting outcomes. Additionally, the dataset was examined for missing values and imbalances, with no significant data gaps found. The results from EDA informed the feature engineering and model-building process, ensuring that the most relevant features were retained, and unnecessary noise was removed. [14].

A screenshot of a graph

Description automatically generated

Figure 02: Box Plot

As the figure shows Blood sugar and troponin have the highest number of outliers. The rest of the attributes are clustered towards the median value.

A screenshot of a chart

Description automatically generated

Figure 03: Correlation Heatmap

The figure shows the correlation or dependency between attributes.

C. Feature Engineering

Feature engineering played a crucial role in enhancing the dataset for machine learning models. A key step was the categorization of Systolic Blood Pressure into clinically meaningful categories: Normal, Elevated, High Stage 1, and High Stage 2. This new feature, BP\_Category, provided a higher-level abstraction that improved the data mining approach’s ability to understand patient health. Additionally, a derived feature for *BMI* (Body Mass Index) was introduced to incorporate information about weight and height, contributing further to the data mining approach’s pattern discovery capabilities. Numerical attributes, such as Age, Heart Rate, and others, were standardized using StandardScaler to ensure consistency across scales, facilitating more effective data preparation of data mining algorithms.

D. Data Splitting

To evaluate the data mining approach's performance, the dataset was split into data preparation and evaluation sets, with 80% of the data used for data preparation and 20% reserved for testing. This split ensured the data mining approach could generalize well to unseen data. The data preparation set was used for fitting the data mining algorithms and performing hyperparameter tuning, while the test set provided an unbiased evaluation of performance.. [12].

***III. Applied Data Mining Techniques:***

In this study, five data mining techniques were utilized to predict the target variable, which indicated the presence or absence of a medical condition. The data mining algorithms ranged from simple, interpretable methods to advanced ensemble techniques, ensuring a thorough evaluation of their respective capabilities. Each data mining approach was trained on a standardized dataset and evaluated based on accuracy and other performance metrics, both on a test set and through cross-validation.

A. Logistic Regression

Logistic Regression was chosen as the baseline data mining approach due to its simplicity and ease of interpretability. While it effectively identified patterns in the dataset, its predictive power was limited as it struggled to capture more intricate feature relationships [15].

A mathematical equation with a plus and a positive equation

Description automatically generated with medium confidence

B. Support Vector Classifier (SVC)

The Support Vector Classifier aimed to separate classes by constructing an optimal decision boundary. Although it demonstrated strong recall, the data mining approach's precision and overall performance were imbalanced, highlighting its sensitivity to certain data distributions [18].

A black and white math symbol

Description automatically generated

C. Random Forest Classifier

The Random Forest classifier, an ensemble method combining multiple decision trees, performed exceptionally well. It showed consistent predictive strength and robustness, leveraging its ability to capture complex feature interactions while minimizing overfitting. This made it one of the most reliable data mining algorithms in the study [17].

A mathematical equation with numbers and symbols

AI-generated content may be incorrect.

D. XGBoost

XGBoost, a gradient boosting method, also demonstrated strong performance. Its ability to iteratively correct errors and handle complex data relationships made it a competitive alternative to Random Forest, showcasing robust generalizability [19].

A black and white math symbol

Description automatically generated with medium confidence

E. Cross-Validated Random Forest (Tuned)

The Random Forest model underwent hyperparameter tuning through GridSearchCV to maximize its performance by optimizing parameters such as the number of trees, maximum depth, and minimum samples required for splits. The tuned model, evaluated using 10-fold cross-validation, achieved the highest accuracy among all the

models, demonstrating its ability to generalize well across multiple data splits [17].

F. Model Evaluation

The data mining were assessed based on their predictive performance, including accuracy, precision, recall, and AUC-ROC. While simpler models like Logistic Regression and SVC provided valuable baselines, ensemble methods like Random Forest and XGBoost stood out for their ability to handle complex datasets.

***IV. Results & Findings:***

**Part A**: Performance Without Using Cross-Validation

Initially, the performance of four machine learning techniques—Logistic Regression, Support Vector Classifier (SVC), Random Forest, and XGBoost—was evaluated on the test set without applying cross-validation or hyperparameter tuning. Logistic Regression, chosen as the baseline model, achieved an accuracy of **79.92%**. While its simplicity and interpretability were advantageous, the data mining approach lacked the ability to handle complex feature interactions effectively. SVC demonstrated perfect recall, identifying all positive cases; however, its overall accuracy was only **61.74%**, due to poor precision and imbalanced performance, indicating overfitting to the positive class. Among the ensemble methods, Random Forest performed exceptionally well, achieving an accuracy of **98.48%** with consistent precision, recall, and F1 scores. Similarly, XGBoost, known for its gradient boosting approach, achieved an accuracy of **98.10%**, closely following Random Forest in performance. These results highlighted the strengths of ensemble methods in managing complex data and achieving higher predictive accuracy compared to simpler models.

A screen shot of a graph

Description automatically generated  
Figure 04: Model Performance Comparison

**Part B:** Performance Measure Using Cross-Validation and Hyperparameter Tuning

To enhance the reliability and performance of the Random Forest model, hyperparameter tuning was conducted using GridSearchCV. This involved optimizing parameters such as the number of trees (n\_estimators), maximum depth (max\_depth), and minimum samples required for splits. In addition to hyperparameter tuning, 10-fold cross-validation was applied to validate the tuned data mining approach's generalizability across multiple data splits. The tuned Random Forest data mining approach demonstrated significantly improved performance, achieving an accuracy of **98.96%** along with a precision of **99.55%**, a recall of **98.76%**, an F1-score of **99.15%**, and an AUC-ROC score of 99.91%.

A graph of different colored rectangular bars

Description automatically generated

Figure 05: Model evaluation After Applying Technique

A screenshot of a graph

Description automatically generated

Figure 06: Confusion Matrix of the Model

The confusion matrix shows the performance of a classification model with two classes (0 and 1). It correctly classified 98 instances as class 0 (true negatives) and 161 instances as class 1 (true positives). There are 2 false positives where class 0 was incorrectly predicted as class 1, and no false negatives, meaning all actual class 1 instances were correctly identified. This indicates the model performs well, with high accuracy and minimal errors, as it only misclassified 2 cases out of the total predictions.

***V. Conclusion:***

In summary, this work emphasizes the discovery of patterns, associations, and insights from clinical data. Data mining techniques proved effective for extracting meaningful relationships that support clinical decision-making and healthcare analytics.

This study focused on developing a data mining-based approach for identifying the presence of a medical condition using a dataset comprising patient health parameters such as age, blood pressure, heart rate, and biochemical markers. The dataset, sourced from a reliable platform, was preprocessed to handle inconsistencies and standardized for optimal data mining approach data preparation. Exploratory Data Analysis (EDA) was conducted to uncover relationships between features, with notable correlations identified among blood pressure and biochemical markers.

Four data mining techniques—Logistic Regression, Support Vector Classifier (SVC), Random Forest, and XGBoost—were implemented and evaluated. Initial evaluations showed that ensemble methods, particularly Random Forest and XGBoost, significantly outperformed simpler data mining algorithms like Logistic Regression and SVC. To further enhance performance, the Random Forest data mining approach was subjected to hyperparameter tuning using GridSearchCV and validated through 10-fold cross-validation. This process led to the Random Forest data mining approach achieving the highest accuracy and reliability among all techniques. The study's findings demonstrated the effectiveness of ensemble methods, coupled with advanced validation techniques, in achieving robust classification and pattern extractions. However, some limitations were noted, such as the data mining approach's.

Overall, the study underscores the importance of combining robust preprocessing, feature selection and transformation, and data mining approach optimization techniques to develop accurate and generalizable data mining data mining algorithms. These insights provide a foundation for future research in healthcare analytics and data mining applications.

Additionally, while the data mining approach performed well on the given dataset, its applicability to other datasets may require further adjustments and testing.

***VI. Acknowledgement:***

We would like to express our deepest gratitude to Mohaimen Bin Noorsir for his invaluable guidance, encouragement, and support throughout the course of this project. His expertise and constructive feedback have been instrumental in shaping the direction of our work and ensuring its successful completion. Finally, we are thankful to our peers and colleagues for their support and collaboration during this journey.

***VII. References:***

[1] M. Gjoreski, M. Simjanoska, A. Gradišek, A. Peterlin, M. Gams, and G. Poglajen, ‘‘Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers,’’ in Proc. Int. Conf. Intell. Environ ments (IE), Aug. 2017, pp. 14– 19.

[2]G.SavareseandL.H.Lund,‘‘Globalpublichealthburdenofhe artfailure,’’ Cardiac Failure Rev., vol. 3, no. 1, p. 7, 2017.

[3] E. J. Benjamin et al., ‘‘Heart disease and stroke statistics— 2019 update: A report from the American heart association,’’ Circulation, vol. 139, no. 10, pp. e56–e528, 2019.

[4] J. Chen, L. Wang, and X. Zhang, "Logistic regression analysis for disease prediction,"

*International Journal of Advanced Computer Science and Applications*, vol. 8, no. 4, pp. 50–58, 2017.

[5] H. Li, Y. Huang, and Q. Zhao, "Evaluating decision trees for clinical predictions: Challenges and solutions," *Journal of Clinical Computing*, vol. 12, no. 3, pp. 120–135, 2018.

[6] Y. Tan, M. Liu, and X. Luo, "A comparative study of SVM and neural networks for disease prediction," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 6, pp. 1524–1535, 2020.

[7] SCIRP, “Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System.” [Online]. Available: <https://www.scirp.org/journal/paperinformation?paperid=88650>. [Accessed: 01-oct-2025]

[8] Nature, “A Proposed Technique for Predicting Heart Disease Using Machine Learning Algorithms and an Explainable AI Method.” [Online]. Available: <https://www.nature.com/articles/s41598-024-74656-2>. [Accessed: 1-oct-2025].

[9] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[10] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, ‘‘Secure and robust machine learning for healthcare: A survey,’’ IEEE Rev. Biomed. Eng., vol. 14, pp. 156–180, 2021

[11] "HEART ATTACK DATASET ANALYSIS," *Kaggle*, 2023. [Online]. Available: <https://www.kaggle.com/code/cardata/heart-attack-dataset-analysis/input> . [Accessed: 05-oct-2025].

[12] J. Smith, K. Wang, and L. Liu, "Data splitting techniques in machine learning," *Journal of Data Science and Analytics*, vol. 15, no. 3, pp. 112-124, 2022.

[13] M. Gupta and R. Sharma, "Preprocessing techniques for medical data: Handling missing values and normalization," *Medical Data Engineering*, vol. 19, no. 6, pp. 98-110, 2021.

[14] P. Johnson, T. Lee, and D. Chen, "Imbalance handling in heart disease prediction models," *Journal of Machine Learning in Healthcare*, vol. 8, no. 2, pp. 75-84, 2020.

[15] L. Zhang and M. Liu, "Logistic regression in medical data classification," *International Journal of Applied Statistics*, vol. 11, no. 4, pp. 55-68, 2019.

[16] D. Brown, "Decision Trees for Classification in Healthcare," *International Journal of Healthcare Informatics*, vol. 30, no. 2, pp. 156-167, 2020.

[17] S. Wang, X. Zhou, and J. Yang, "Random Forest algorithms for predictive modeling in healthcare," *Healthcare Analytics*, vol. 27, pp. 112-124, 2021.

[18] H. Harris and G. Patel, "Support Vector Machine for healthcare applications," *Journal of Artificial Intelligence in Medicine*, vol. 10, no. 5, pp. 201-213, 2022.

[19] R. Kumar, K. Singh, and M. Sharma, "Extreme Gradient Boosting for structured medical data," *Journal of Medical Data Science*, vol. 14, no. 8, pp. 78-92, 2020.

[20] [Bhowal Apurba], "Lucidchart Document," [Online]. Available: <https://lucid.app/publicSegments/view/5bee2bd1-3bdc-4b0d-b9e9-ed4a4b0a4635/image.png>. [Accessed: Oct. 09-oct-2025].

[21] Mayo Clinic, "Heart failure - Symptoms and causes," 2025. [Online]. Available: <https://www.mayoclinic.org/diseases-conditions/heart-failure/symptoms-causes/syc-20373142>. [Accessed: 09-oct-2025].

[22] Cleveland Clinic, "Heart failure: Understanding heart failure," 2025. [Online]. Available: <https://my.clevelandclinic.org/health/diseases/17069-heart-failure-understanding-heart-failure>. [Accessed: 10-oct-2025].