HEART FAILURE RISK PREDICTION & KEY HEALTH INDICATOR ANALYSIS

Final Project Report

Lanxin LI
DSBA M2, CentraleSupelec Paris, France
lanxin.li@student-cs.fr

Meng XIA
DSBA M2, CentraleSupelec Paris, France
meng.xia@student-cs.fr

Hanqi YANG
DSBA M2, CentraleSupelec Paris, France
hanqi.yang@student-cs.fr

Abstract

Heart disease remains a leading global health challenge, requiring timely and accurate diagnosis to enable early intervention. Traditional medical approaches often struggle with the complexity and volume of data involved. This study leverages machine learning (ML) techniques to develop a predictive model for heart disease diagnosis using the Kaggle Heart Failure Prediction Dataset, which contains 13 clinical features. The research addresses two core questions: how to build an efficient predictive model based on health metrics, and which variables are most influential in predicting heart disease risk.

Random Forest, SVM, and XGBoost models were evaluated through rigorous experimentation. The dataset was preprocessed with feature scaling and one-hot encoding for categorical variables, followed by 5-fold cross-validation and grid search for hyperparameter optimization. Evaluation metrics, including Recall, F1-score, and ROC-AUC, were prioritized to balance false negatives and overall model performance. Among the models, Random Forest demonstrated superior robustness and balanced performance, achieving a Recall of 89.72% and an ROC-AUC of 94.74%. Key predictors identified included ST segment slope, cholesterol levels, and exercise-induced angina, aligning with established medical knowledge.

This study highlights the Random Forest model's ability to deliver high predictive accuracy while maintaining interpretability, making it an ideal tool for clinical decision-making. Future research will focus on integrating more diverse datasets, dynamic features, and advanced explainability techniques to further enhance model utility and reliability in real-world medical applications.

1 Introduction

Detecting warning signs of heart disease is critical for timely intervention. However, traditional medical approaches are often limited by the complexity and volume of data, making it challenging to comprehensively assess a patient's health status and risk factors. In recent years, machine learning (ML), as an artificial intelligence technology, has gained significant attention for its ability to identify patterns and relationships in large datasets. By analyzing and integrating various types of information, including clinical symptoms, laboratory test results, and imaging data, ML can achieve higher accuracy in diagnosing heart failure and predicting disease progression and patient outcomes.

In this study, we aim to design and develop a machine learning-based predictive model using the Heart Failure Prediction Dataset provided by Kaggle, which includes 13 features. Our goal is to identify individuals at high risk of heart failure and address the following core questions: How can an efficient predictive model be built based on health indicators? Which variables are the most influential in predicting heart failure?

2 Problem Definition

Our problem can be formally defined as a binary classification task, with the objective of training a machine learning model $f(X;\theta)$ to predict the risk of heart failure $\hat{y}$ based on the health indicators X of patients. The symbols and model definitions are as follows:

- **Input feature matrix** X: Represents the key health indicators of each patient. The dataset includes 12 feature variables, such as Age and Resting Blood Pressure (RestingBP).
- **Target variable** y: Indicates whether a patient is diagnosed with heart failure (Heart Disease). This is a binary variable where:
  - $y = 1$: The patient is diagnosed with heart failure.
  - $y = 0$: The patient is not diagnosed with heart failure.
- **Predictive model** $f(X;\theta)$: Uses parameters θ to predict the output $\hat{y}$ based on the input features X. Here, f can be any supervised learning algorithm.
- **Loss function** $L(y;\hat{y})$: Measures the error between the predicted value $\hat{y}$ and the true label y. The loss function used is cross-entropy loss:

$$L(y;\hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right],$$

  where n is the number of samples, $\hat{y}_i$ is the predicted probability for sample i, and y_i is the true label for sample i (either 0 or 1).

The constraints of this study include the following: The dataset has a fixed dimensionality of 13 features, and there may be missing or anomalous values that require appropriate data cleaning and preprocessing. The dataset might exhibit class imbalance, where positive samples are underrepresented. To address this issue, techniques such as class weight adjustment or sampling strategies (e.g., oversampling or undersampling) need to be applied.

In this project, based on existing literature, we selected Recall, F1-score, and ROC-AUC as the primary evaluation metrics for the model. The optimization goal is to maximize these three metrics. By combining these metrics, we aim to comprehensively evaluate the model's performance in terms of missed diagnosis control, false positive trade-offs, and overall effectiveness, thereby ensuring the practical utility and reliability of the predictive model in medical scenarios. Specifically:

- **Recall**: This is regarded as the most critical metric, as the core objective in cardiovascular disease prediction is to avoid missed diagnoses (False Negatives). A missed diagnosis could result in patients missing timely treatment opportunities.
- **F1-score**: Used to balance Recall and Precision. While a high Recall reduces missed diagnoses, it may introduce more false positives. The F1-score ensures that the model maintains high Recall without sacrificing too much Precision. This makes it an essential metric for balancing the trade-off between missed and false diagnoses.
- **ROC-AUC**: Evaluates the model's overall performance and is suitable for comparing the classification capabilities of different models. While not directly optimizing Recall or Precision, it reflects the model's performance across different decision thresholds.

3 Related Work

The prediction of heart disease has been extensively explored through various machine learning approaches, as demonstrated in six key studies. These works provide a solid foundation for understanding data processing, feature selection, model choice, and evaluation metrics. Below is a summary of the key methodologies and findings from these studies:

- **Uddin et al. (Epistemological Advancements in Cardiological Forecasting)**: Used the UCI Heart Failure dataset with 299 records and 13 clinical features. The study employed Random Forest, which achieved high recall and ROC-AUC, but without feature reduction, the risk of redundancy and overfitting was noted.
- **Samar L. Elhawy (Harnessing Statistical Analysis and Machine Learning Optimization for Heart Attack Prediction)**: Applied Random Forest for feature selection, identifying three key features (chest pain type, maximum heart rate, and number of major vessels). Logistic Regression was chosen as the optimal model, achieving high recall through hyperparameter tuning.
- **Fang Zhou Qu et al. (Construction of Clinical Predictive Models for Heart Failure Detection)**: Used statistical tests (t-tests) to select four key features on a small dataset. Logistic Regression was again selected due to its simplicity and high recall, though overfitting risks were significant due to the limited data size.
- **Wu et al. (Interpretable Machine Learning for Personalized Medical Recommendations)**: Focused on integrating model interpretability using LIME, demonstrating its utility alongside Random Forest, SVM, and GBDT models. Random Forest emerged as a robust and interpretable choice for medical applications.
- **Arooj et al. (A Deep Convolutional Neural Network for the Early Detection of Heart Disease)**: Introduced a CNN model that excelled in accuracy but lacked the interpretability critical for clinical applications. While deep learning proved effective for capturing complex relationships, its resource-intensive nature posed challenges.
- **Hajiarbabi et al. (Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm)**: Highlighted the Jellyfish Optimization Algorithm for feature selection and compared Random Forest, AdaBoost, and SVM. Random Forest achieved the best balance of recall and ROC-AUC, particularly after feature dimensionality reduction.

3.1 Relation to Our Project

Our project synthesizes these findings and addresses their limitations to develop a robust heart disease prediction model.

- **Feature Selection:** The reviewed studies emphasize the need for targeted feature selection to balance interpretability and predictive power. Inspired by the Random Forest-based selection used by Elhawy and Hajiarbabi, and the statistical tests employed by Fang Zhou Qu, we combined feature importance analysis and interaction assessments to optimize our input features. Unlike Uddin et al., we avoided including all features, reducing the risk of redundancy.
- **Model Selection:** Random Forest consistently emerged as a robust performer in Uddin, Hajiarbabi, and Wu's studies, particularly for its balance between interpretability and recall. Logistic Regression, favored by Elhawy and Fang Zhou Qu for small datasets, demonstrated simplicity but lacked scalability. Deep learning, as explored by Arooj et al., showed promise for complex patterns but was resource-intensive. Based on these insights, we selected Random Forest, as it provides a strong trade-off between performance and interpretability, crucial for clinical applications.
- **Evaluation Metrics:** Recall was universally prioritized for its critical role in minimizing false negatives, as emphasized by Uddin, Elhawy, and Qu. F1-score and ROC-AUC complemented recall to balance precision and overall model performance. These metrics guided our evaluation framework, ensuring alignment with clinical priorities.

3.2 How Our Project Differs or Complements Previous Work

- **Comprehensive Feature Analysis:** While most studies focused on individual feature importance, our project explores feature interactions to uncover latent relationships, addressing gaps in the reviewed works.
- **Advanced Evaluation Framework:** Unlike earlier studies that focused predominantly on a single metric, we integrated recall, F1-score, and ROC-AUC into a weighted evaluation framework, ensuring a holistic assessment.
- **Scalable and Interpretable Approach:** Deep learning approaches like those by Arooj et al. were resource-intensive and lacked interpretability. Our Random Forest-based model balances performance and clinical applicability, leveraging feature importance for decision-making insights.

4 Methodology

4.1 Data and Feature Engineering

The data used in this study is sourced from the publicly available Heart Failure Prediction Dataset on the Kaggle platform, which contains 918 observations. A detailed description of the dataset features can be found in the EDA code file.

Feature Engineering

We began by checking for missing values and duplicate entries. Due to the high quality of the dataset, no duplicates or missing values were found.

Based on the meaning and type of variables, the features were manually categorized into: Numerical features (Age, RestingBP, Cholesterol, MaxHR, Oldpeak) and Categorical features (Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina, ST_Slope).

Categorical features were one-hot encoded, splitting each category into binary variables to improve the model's ability to interpret categorical information. Numerical features were standardized using Z-score normalization, which helps mitigate the impact of scale differences between features and improves the learning efficiency of the model.

4.2 Exploratory Data Analysis

The target variable HeartDisease is relatively balanced, with 55% labeled as 1 and 45% as 0. To address potential data imbalance, relevant parameter adjustments were incorporated during optimization.

Univariate analysis highlighted key patterns: Age is concentrated in middle-aged and older individuals, showing a near-normal distribution. RestingBP typically falls between 120-140 mmHg, reflecting hypertension traits. Cholesterol is right-skewed, indicating elevated levels in some patients. MaxHR is symmetrically distributed, mostly between 140-160 bpm, while Oldpeak also right-skewed, concentrates between 0-2, suggesting mild myocardial ischemia.

Multivariate analysis revealed heart disease trends: its prevalence rises with Age, confirming it as a major risk factor. Cholesterol levels are higher and more variable in heart disease patients, underscoring its diagnostic value. MaxHR is notably lower, reflecting impaired heart rate regulation. Oldpeak values are elevated, emphasizing the diagnostic importance of ECG abnormalities. Males have a higher prevalence, while asymptomatic chest pain (ASY) and elevated fasting blood sugar (FastingBS ¿ 120 mg/dl) strongly associate with heart disease. Flat or downsloping ST_Slope is also predominant in patients.

Combined analysis showed male patients with ASY exhibit higher resting blood pressure, linking it to hypertension. Heart disease patients have higher cholesterol, especially in males, and MaxHR is significantly lower, with sharper declines in males.

In summary, Oldpeak, MaxHR, and Age are critical features strongly correlated with HeartDisease, providing essential insights for feature selection and diagnosis.

4.3 Model

The experiment involved four steps: dataset splitting, cross-validation, hyperparameter tuning, and model evaluation. The dataset was split into 80% training and 20% testing data for robust evaluation. 5-fold cross-validation optimized the model while reducing randomness from data splits. Recall was the optimization target during hyperparameter tuning. Finally, model performance was assessed on the test set to ensure comprehensive evaluation.

4.3.1 Random Forest

Hyperparameter Tuning Method

Grid Search was employed with recall as the optimization target, focusing on improving the detection rate of positive cases (disease samples). Through hyperparameter tuning, the optimal combination was identified as: {'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 100}

Overfitting Prevention Measures

To prevent overfitting, the model's complexity was controlled by setting the maximum tree depth (max_depth) to 10, limiting the learning of noise in training data. Minimum samples for node splitting (min_samples_split=5) and leaf nodes (min_samples_leaf=1) further constrained tree growth, avoiding overly complex decision boundaries.

A random feature sampling strategy (max_features='sqrt') introduced randomness at each split, reducing reliance on individual features and enhancing model robustness. To address class imbalance, a balanced class weight parameter (class_weight='balanced') was applied, improving the model's detection of underrepresented classes. Finally, Random Forest utilized an ensemble approach (Bagging) to combine predictions from multiple trees, enhancing generalization and minimizing overfitting risks. These strategies collectively ensured high recall, accuracy, and robust generalization, making Random Forest a reliable tool for heart disease diagnosis.

Training Results and Model Performance

The model achieved a recall score of 0.8972 on the test set, demonstrating strong capability in identifying heart disease cases and reducing missed diagnoses in clinical applications. A high ROC-AUC value (0.9474) further confirmed consistent performance in distinguishing positive and negative samples across thresholds. Additionally, the Random Forest model excelled in F1-Score and Precision, showcasing balanced and robust overall performance.

4.3.2 SVM

4.3.2.1 Grid Optimization with Recall as the Objective Hyperparameter Tuning

Grid Search was employed with Recall as the optimization target. The hyperparameter tuning process identified the optimal combination as follows:

```
{'svc__C': 0.01, 'svc__class_weight': None, 'svc__coef0': 0.0,
 'svc__degree': 2, 'svc__gamma': 0.001, 'svc__kernel': 'rbf'}
```

Overfitting Prevention Measures

To prevent overfitting, we limited the model's complexity by setting the regularization parameter C during the Grid Search. Additionally, 5-fold cross-validation was applied to ensure the model demonstrated consistent performance across different subsets of the data, thereby improving its generalization capability.

Training Results and Model Performance

The final performance of the model on the test set is summarized in Table 1. In this training run, we observed a perfect Recall of 1.0; however, the F1-Score was only 0.7133. Both Accuracy and Precision hovered around 0.55, indicating low positive class prediction accuracy and frequent misclassification of negative samples. This result suggests that the model overfitted to the positive class.

To address this overfitting issue in the SVM model, we abandoned the strategy of solely optimizing for Recall during Grid Search. Instead, we adopted a custom composite scoring metric to retrain and fine-tune the model.

4.3.2.2 Grid Optimization with a Composite Scoring Metric

Optimal Parameters and Overfitting Mitigation

To address the overfitting issue caused by optimizing only for Recall, we defined a custom scoring function, `custom_score`, which combines Recall (weight 0.4), F1-Score (weight 0.3), and ROC-AUC (weight 0.3). This approach mitigates the problem of model bias that arises from optimizing a single metric.

The optimal hyperparameter combination identified was:

```
{'svc__C': 0.1, 'svc__class_weight': None, 'svc__coef0': 0.1,
 'svc__degree': 3, 'svc__gamma': 'scale', 'svc__kernel': 'poly'}
```

To further prevent overfitting, we selected a polynomial kernel. The polynomial kernel (degree=3) is capable of capturing complex feature interactions. With an appropriately designed coef0, its complexity aligns well with the data distribution without leading to overfitting.

Training Results and Model Performance

The final performance of the model on the test set is presented in Table 1. Using the composite scoring approach, the F1-Score showed a significant improvement, reaching 0.9135. Precision increased to 0.8962, while Recall decreased slightly to 0.9314 but still demonstrated strong detection capability for positive cases (heart disease). This approach also achieved a better balance in detecting negative cases (non-disease).

4.3.3 XGBoost

Hyperparameter Tuning

Grid Search CV was employed to optimize the model, and the best parameter combination identified was:

```
{'colsample\_bytree': 0.8, 'gamma': 0, 'learning\_rate': 0.01, 'max\_depth': 3,
 'min\_child\_weight': 1, 'n\_estimators': 50, 'reg\_alpha': 0, 'reg\_lambda': 10,
 'scale\_pos\_weight': 1, 'subsample': 1.0}
```

Overfitting Prevention Measures

To prevent overfitting, tree depth was limited to 3 (`max_depth`), and a high L2 regularization parameter (`reg_lambda=10`) was applied to reduce complexity and enhance stability. Feature and sample randomization (`colsample_bytree=0.8`, `subsample=1.0`) minimized reliance on specific features, mitigating overfitting risks. Additionally, a low learning rate (`learning_rate=0.01`) ensured gradual optimization, improving model robustness.

Training Results and Model Performance

XGBoost achieved excellent Recall (0.9159), effectively identifying positive cases. A high ROC-AUC (0.9272) validated its ability to distinguish between classes across thresholds. However, lower Precision and F1-Score indicated reduced effectiveness in handling negative samples and a potential risk of false positives.

5 Evaluation

5.1 Model Evaluation

5.1.1 Quantitative Comparison of Model Performance

Model	Recall	Precision	F1-Score	Accuracy	ROC-AUC
Random Forest	89.72%	91.43%	90.57%	89.13%	94.74%
SVM (Recall Optimized)	100.00%	55.43%	71.33%	55.43%	92.27%
SVM (Composite Optimized)	93.14%	89.62%	91.35%	90.22%	93.64%
XGBoost	91.59%	82.35%	86.73%	83.70%	92.72%

Table 1: Performance comparison of different models.

Random Forest achieved the best performance across Recall, Precision, F1-Score, and ROC-AUC, demonstrating the most balanced overall performance. XGBoost performed well in Recall and ROC-AUC but was slightly inferior to Random Forest in Precision and F1-Score. SVM (optimized for a composite metric) showed improvements in Precision and Accuracy, offering decent overall performance but with a tendency toward overfitting.

5.1.2 RandomForest vs SVM

When SVM was optimized solely for Recall, it encountered overfitting issues. To mitigate this, a custom composite scoring metric was introduced. This metric was also used to re-optimize the Random Forest model via grid search, but the optimal parameters for Random Forest remained unchanged, highlighting its robustness. The reasons for this distinction are explored below:

Random Forest and SVM differ fundamentally in their decision-making mechanisms. Random Forest employs a voting system across multiple trees, balancing Recall and Precision among trees to reduce bias and error. In contrast, SVM relies on a single decision boundary, which, when overly focused on Recall, often sacrifices Precision and Accuracy.

For class imbalance, Random Forest excels by using Gini impurity or information gain, naturally accounting for class distribution. Its trees independently optimize Recall, enhancing positive case detection while maintaining diversity to avoid excessive bias. SVM, on the other hand, defaults to equal class weighting. Without explicitly adjusting the `class_weight` parameter, optimizing Recall can lead to significant bias toward the positive class.

Random Forest also shows greater robustness. Its Bagging and feature randomization strategies make it less prone to over-optimization, enabling reliable handling of complex data distributions. In contrast, SVM's sensitivity to parameter tuning and optimization goals makes it more susceptible to overfitting when Recall is heavily prioritized.

5.1.3 Experimental Findings and Model Selection

The Random Forest model outperformed SVM and XGBoost in this project, demonstrating several key advantages.

Firstly, Random Forest exhibited a balanced performance in critical metrics such as Recall and Precision. This balance effectively reduces both false negatives and false positives, making it particularly suitable for heart disease diagnosis, where high Recall and reliability are essential.

Additionally, Random Forest's robustness and resistance to overfitting further enhanced its performance. Its voting mechanism and feature randomization strategy ensured stable performance even when dealing with complex data distributions. In contrast, SVM and XGBoost were more sensitive to parameter tuning and data distribution.

In conclusion, Random Forest was identified as the optimal model for heart disease diagnosis in this study.

5.2 Feature Importance Evaluation

5.2.1 Feature Importance Analysis of the Random Forest Model

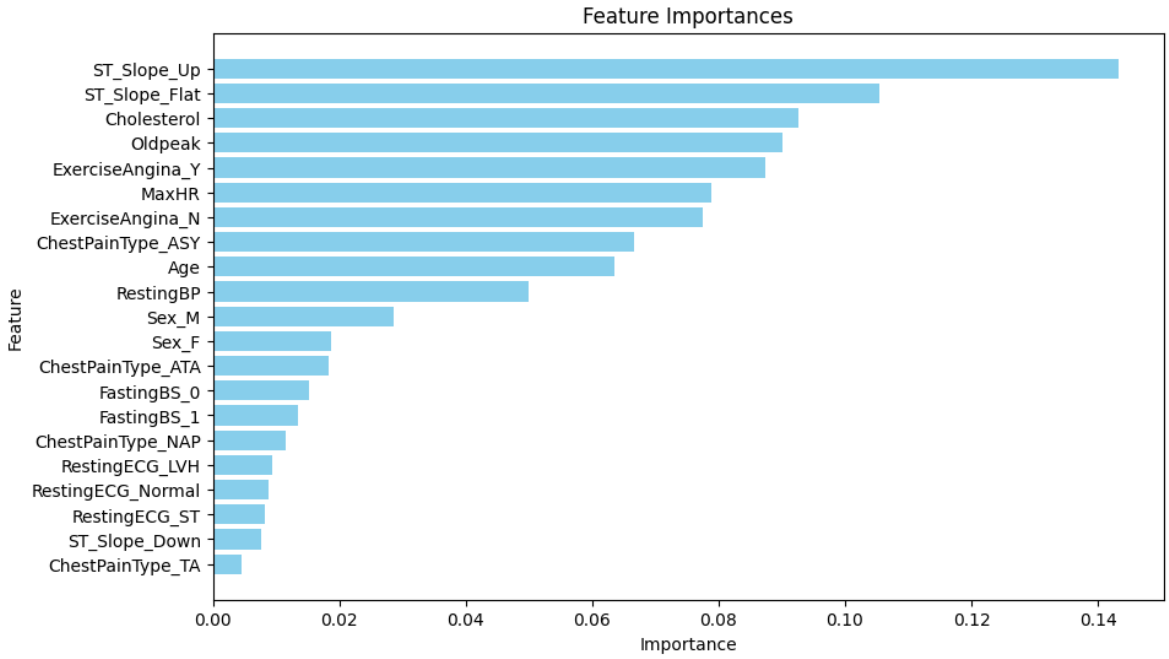

Figure 1: Feature Importance of Random Forest

`ST_Slope_Up` (0.143) and `ST_Slope_Flat` (0.106) are critical indicators for predicting heart disease. The ST segment slope directly reflects cardiac health, where `ST_Slope_Up` is typically associated with normal or low-risk conditions, while `ST_Slope_Flat` indicates potential cardiac abnormalities. This finding aligns closely with medical knowledge, further validating the importance of ST segment shape as a diagnostic marker. Additionally, Cholesterol (0.093) and Oldpeak (0.090) also demonstrated high importance. Elevated cholesterol levels are a key risk factor for cardiovascular disease, while Oldpeak reflects changes in the ST segment during exercise, suggesting potential myocardial ischemia or exercise intolerance.

Among secondary features, `ExerciseAngina_Y` (0.087) and `MaxHR` (0.079) showed strong predictive capabilities, reflecting the patient's response to exercise stress and dynamic cardiac health, respectively. Age (0.064), a moderately important feature, highlights the significant increase in heart disease risk with advancing age. Furthermore, sex-related features (`Male: 0.029`, `Female: 0.019`) had weaker predictive power but supported the conclusion that males have a slightly higher risk of heart disease compared to females.

Less significant features, such as `RestingECG_Normal` and `ChestPainType_TA`, contributed minimally, likely due to weak signals or overlap with more predictive features, limiting their role in distinguishing heart disease risk.

5.2.2 Comparison with SVM and XGBoost

All three models consistently identified ST slope-related features (e.g., `ST_Slope_Up` and `ST_Slope_Flat`) as critical indicators for heart disease prediction, reinforcing their core importance. Cholesterol and

Oldpeak were also highlighted across Random Forest, SVM, and XGBoost, emphasizing their key role in heart disease forecasting. Additionally, `ExerciseAngina_Y` demonstrated strong discriminative power in all models, underlining its efficiency in distinguishing positive and negative cases.

6 Conclusions

This study identified Random Forest as the optimal model for heart disease diagnosis, offering high recall, robust performance, and strong interpretability. Key features like ST slope, cholesterol levels, and `ExerciseAngina_Y` were accurately identified, aligning with medical insights and providing valuable clinical support. By prioritizing Recall to minimize false negatives, the model achieved a balance between accuracy and interpretability. Future efforts will focus on diversifying datasets, incorporating dynamic features, and enhancing interpretability tools to improve generalizability and clinical applicability.

References

- Ahmad, A.A. and Polat, H. (2023) 'Prediction of heart disease based on machine learning using jellyfish optimization algorithm', *Diagnostics*, 13(14), p. 2392. doi:10.3390/diagnostics13142392.
- Arooj, S. et al. (2022) 'A deep convolutional neural network for the early detection of heart disease', *Biomedicines*, 10(11), p. 2796. doi:10.3390/biomedicines10112796.
- Elhawy, S.L. (2025) 'Harnessing Statistical Analysis and machine learning optimization for heart attack prediction', *Multicriteria Algorithms with Applications*, 6, pp. 57–65. doi:10.61356/j.mawa.2025.6456.
- Hajiarbabi, M. (2024) 'Heart disease detection using machine learning methods: A comprehensive narrative review', *Journal of Medical Artificial Intelligence*, 7, pp. 21–21. doi:10.21037/jmai-23-152.
- Qu, F.Z. et al. (2024) 'Construction of clinical predictive models for heart failure detection using six different machine learning algorithms: Identification of key clinical prognostic features', *International Journal of General Medicine*, 17, pp. 6523–6534. doi:10.2147/ijgm.s493789.
- Wu, Y. et al. (2023) 'Interpretable machine learning for personalized medical recommendations: A lime-based approach', *Diagnostics*, 13(16), p. 2681. doi:10.3390/diagnostics13162681.
- Bushra, U., Mohammad Shihab, U., Sharmin, S., MD Sanowar Hossain, S., Fariha Ferdous, N., & MD Salah, U. (2025) 'Epistemological advancements in cardiological forecasting: Machine learning as a paradigm for prognostic precision', *Preprint*. doi:10.13140/RG.2.2.18990.01602.