International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Advanced Deep Learning Approach for Predicting Heart Disease Through Comprehensive Analysis of Clinical Features and Data Science Techniques

Mousa Adel Mousa ALanazi

Prince Fahd bin Sultan University – College of Computer Science
Data Science / Machine Learning

Abstract: ***This study develops and evaluates a deep learning model for predicting heart disease using the UCI Cleveland clinical dataset, which contains 303 patient records and 14 commonly used diagnostic features. After data cleaning, imputation of missing values, feature encoding, and normalization, a neural network with multiple dense layers, batch normalization, and dropout was trained and tested on a stratified train–test split. The model achieved an accuracy of about 82 percent with balanced precision, recall, and F1 scores for both classes. Traditional machine learning models, particularly Support Vector Machine achieved slightly higher performance, but the neural network remained competitive. These findings highlight the potential of deep learning as a decision-support tool for early identification of heart disease risk when combined with structured clinical data.***

Keywords: Heart Disease, Deep Learning, Neural Network, Clinical Dataset, Machine Learning

1. Overview of the Project

1.1 Introduction and Objectives

The project focuses on developing a machine learning model that can predict the presence of heart disease in patients using a wide range of clinical features. Given the high global mortality associated with heart disease, the importance of early and accurate prediction for strengthening preventive healthcare is clear. This study aims to examine the relationships between key clinical factors and heart disease, develop a neural network model capable of achieving reliable predictive performance, evaluate the model's effectiveness by identifying the most influential features, and generate insights that may support early detection and inform preventive strategies.

1.2 Problem Statement

Diagnosing heart diseases generally take a lot of medical testing and the judgment of a specialist. Machine learning solution might offer a rapid initial analysis using normal patient data that would lead to faster treatments and better use of healthcare staff and facilities.

1.3 Approach

This study employs a supervised neural network model to distinguish patients with heart disease from those without the condition. The workflow is structured into several key stages: data exploration, preprocessing, and feature engineering to prepare the dataset for modeling; development of a deep learning architecture tailored for binary classification; training of the model using appropriate optimization techniques; and finally, evaluation of the model's performance through quantitative metrics and identification of its limitations.

2. Used Dataset

2.1 Dataset Description

This analysis uses the UCI Heart Disease dataset, which is accessible from the UCI Machine Learning Repository and is widely employed in research on heart disease prediction. The dataset contains 303 patient records and 14 clinical features, including demographic attributes, diagnostic test results, and medically relevant indicators. These features provide a solid foundation for building machine learning models aimed at identifying individuals at risk of heart disease.

2.2 Dataset Features

Feature	Description
age	Age in years
sex	Sex (1 = male, 0 = female)
cp	Chest pain type (0–3, categorical)
trestbps	Resting blood pressure (in mm Hg)
chol	Serum cholesterol (in mg/dl)
fbs	Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
restecg	Resting electrocardiographic results (0–2, categorical)
thalach	Maximum heart rate achieved
exang	Exercise induced angina (1 = yes, 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment (0–2, categorical)
ca	Number of major vessels colored by fluoroscopy (0–4)
thal	Thalassemia type (0–3, categorical)
target	Heart disease diagnosis (1 = present, 0 = absent)

2.3 Dataset Source and Relevance

The Heart Disease UCI data set is a collection of data from the institutions like Cleveland Clinic Foundation, Hungarian Institute of Cardiology, VA Medical Center (Long Beach), and University Hospital (Zurich). This analysis will confine our attention to the Cleveland data, which is the part most frequently employed for machine learning experiments.

This dataset is the most important for our study due to it's containing a variety of features that doctors usually employ for heart risk assessment. Hence it is proper for the development of a predictive model.

3. Pre- Processing Stage

3.1 Data Clearing and Exploration

The data exploration is the initial activity for loading the dataset and to do the Data exploration represents the initial step of the analysis, during which the dataset is loaded, basic descriptive statistics are examined, and potential issues such as missing values or outliers are identified. The dataset contains 303 patient records and 14 clinical features, including age, sex, chest pain type, blood pressure, cholesterol levels, exercise-induced angina, and the final diagnosis (target variable).

Two variables, ca and thali, contain missing values—four in ca and two in thali. These missing values must be handled appropriately through imputation or row removal to ensure the model receives a complete and reliable dataset.

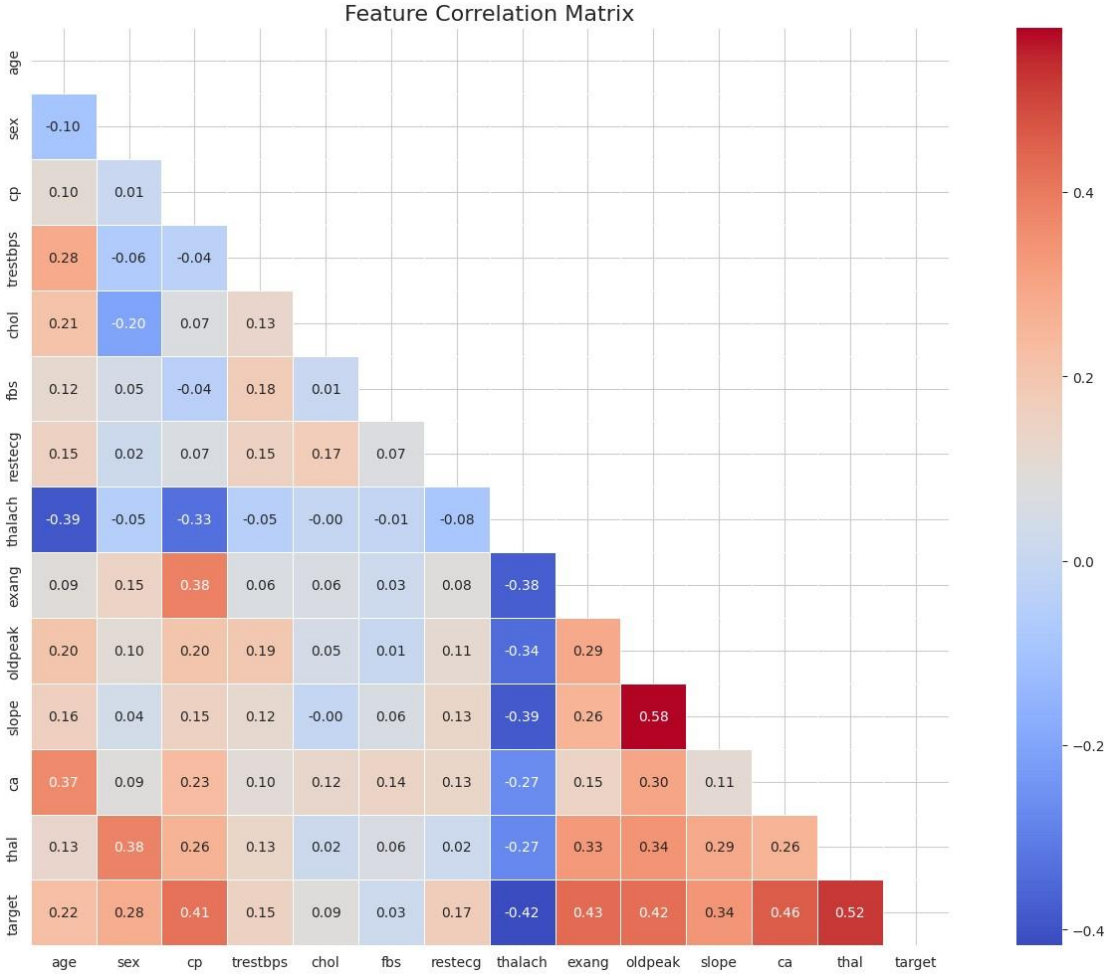
The average age of the individuals is 54.44 years, with ages ranging from 29 to 77. The mean cholesterol level is 246.69 mg/dl, and the mean maximum heart rate achieved (halacha) is 149.61 bpm. Other variables such as chest pain type, resting ECG results, and exercise-induced angina also display meaningful variation across the dataset.

The target variable is relatively balanced, with 164 samples labeled as 0 (no heart disease) and 139 samples labeled as 1 (heart disease), resulting in a distribution that allows the model to learn both classes effectively without requiring additional imbalance-handling techniques.

Key preprocessing steps include handling missing data in the car and thali attributes through imputation or selective row removal, scaling or normalizing numerical features to improve model performance, encoding categorical variables using one-hot or label encoding, and checking for multicollinearity to avoid redundant predictors that could negatively affect the model's performance.

Categorical variables were transformed using one-hot encoding where appropriate, and all numerical features were normalized to ensure consistent scaling across the dataset. These preprocessing measures help stabilize the learning process and enhance the model's ability to generalize to unseen data.

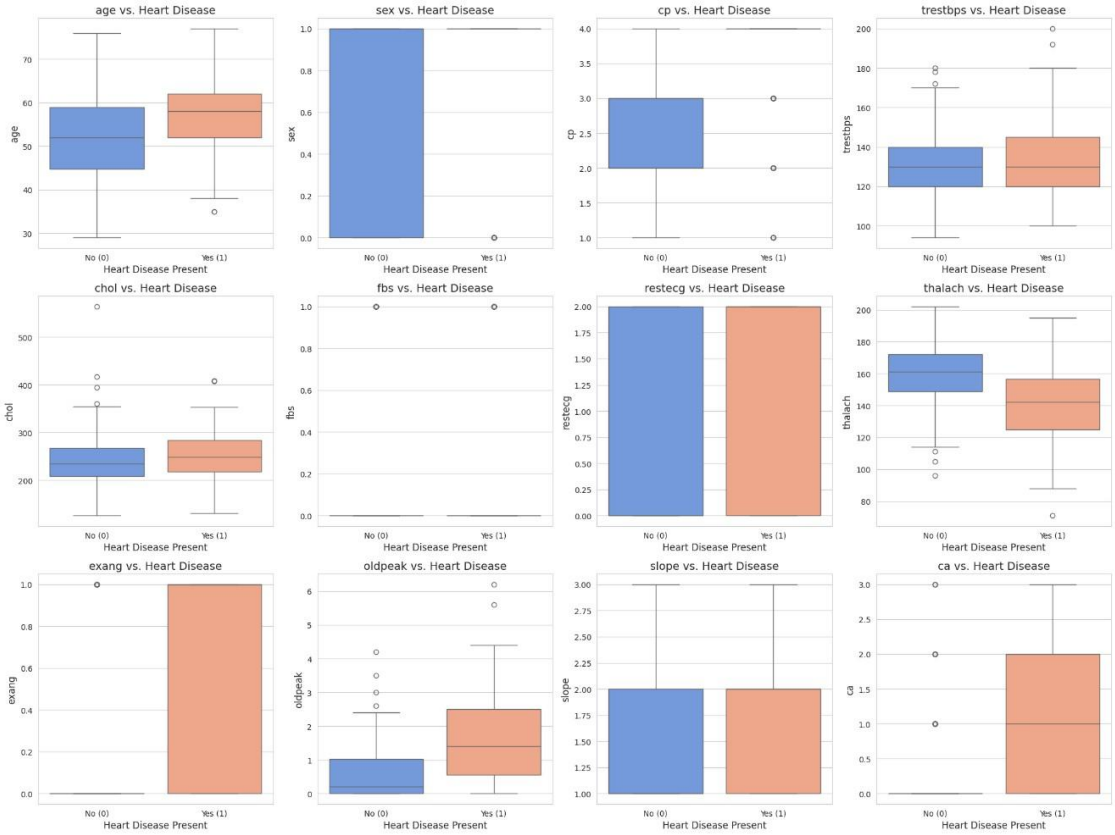
.



The target variable, which is indicative of the presence or absence of heart disease, is relatively evenly distributed. 164 instances are marked as 0 (no heart disease), while 139 instances are marked as 1 (heart disease present), meaning 45.87% are thus positive cases. This kind of class distribution is pretty much balanced, only the negative class has such a tiny majority. In this state the model should be capable of

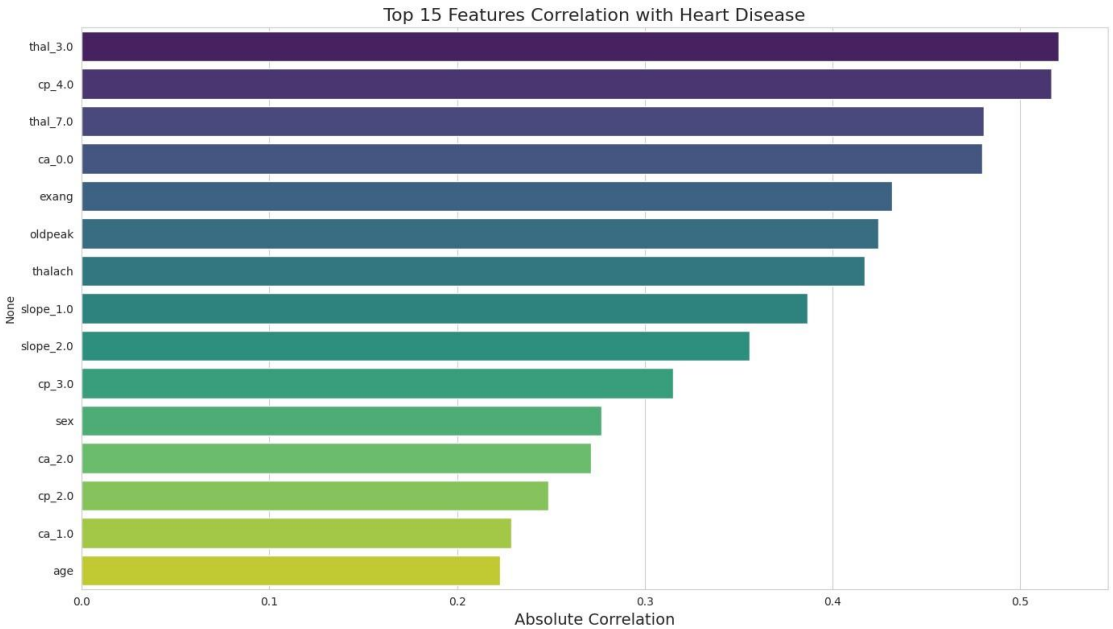
learning both classes except for the situation of a severe imbalance which would require a specific intervention.

The steps on which we are going to concentrate on include the missing data in 'ca' and 'thali' by imputing the reasonable values, or by removing the affected rows/columns, the scaling or the normalizing of the features to merely improve model performance, the conversion of the categorical variables to numeric data by either one-hot encoding or label encoding, and the checking of multicollinearity among the features to avoid redundant predictors that could be the reason for the model's bad performance.



3.2 Feature Engineering and Scaling

The data has undergone a process of feature transformation through one hot encoding of categorical variables for model training, the same of which is carried out, where appropriate. All the numeric features are then normalized to the same scale so that the neural network works faster and better in terms of convergence. Additional work is done by the dataset to detect and ensure that there are no issues pertaining to the multicollinearity that might affect the accuracy and interpretability of the model.



4. Splitting the Dataset

4.1 Train Test Split

For accurate performance evaluation, the dataset was divided into training and testing subsets. The training set is used to fit the model, while the testing set provides an unbiased assessment of the model's performance on unseen data.

The training set (Train) contains 212 samples with 25 features, and the testing set (Test) contains 91 samples with the same number of features. The corresponding target arrays, train and test, have shapes (212,) and (91,), respectively.

Class distribution within the training set is relatively balanced, with approximately 54.2% of samples belonging to Class 0 and 45.8% belonging to Class 1. This corresponds to an approximate ratio of 1.2 to 1, which is small enough that class weighting or other imbalance-handling techniques are not required.

Maintaining similar class distributions in both the training and testing sets reduces the risk of biased performance estimates. Since both subsets preserve this balance, the evaluation results are more reliable and less affected by distributional shifts.

Although the dataset does not exhibit severe class imbalance, it remains important to monitor precision, recall, and F1-score to ensure that the model performs consistently across both classes. These metrics help identify whether the model unintentionally favors one class over the other 45.8%). In percentage terms, this difference is quite small (approximately 1.2:1) and, as usual, it is not large enough to cause a problem for the model, such as requiring the use of class weights.

But if your testing set does not have a very similar class distribution to that of the training set, the model's performance might still be overestimated. If the class distribution is the same, the predictions will not be biased since they come from a balanced dataset. The training and the testing set having the same class distribution show that the testing set performance will also not be affected by any distributional problem occurring in the training set.

The fact that the distribution of the dependent variable in both the training and testing sets is not skewed, does not imply that no treatments are necessary; for example, oversampling, under sampling, or class weighting can still be useful. A few differences might not be enough to cause the bias, but it is still possible to monitor those metrics. Hence, overfitting to one class should not be the issue of the model. However, the model's performance can be assessed using these metrics.

Dataset Overview

Dataset	Shape
X_train	(212, 25)
X_test	(91, 25)
y_train	(212,)
y_test	(91,)

Class Distribution

Set	Class 0 (No Heart Disease) %	Class 1 (Heart Disease) %
Training Set	54.2	45.8
Testing Set	53.8	46.2

4.2 Data Augmentation Considerations

Data augmentation techniques are typically used when a dataset exhibits substantial class imbalance. However, in this study, the dataset maintains a relatively balanced distribution, with an approximate majority-to-minority ratio of 1.19:1. This small disparity indicates that the model can learn both classes effectively without requiring class weighting or resampling techniques.
Although additional balancing methods are unnecessary, it remains essential to evaluate the model using metrics such as precision, recall, and F1-score to ensure consistent performance across both classes. Even minor imbalances can influence real-world applications, particularly in sensitive domains such as medical diagnosis, where misclassifications may have serious consequences.

5. Used DL Algorithm

5.1 Neural Network Architecture

The deep learning model was implemented using TensorFlow and Kera's to perform binary classification for heart disease prediction. Sequential architecture was selected, consisting of multiple fully connected (Dense) layers with progressively decreasing neuron counts (64 → 32 → 16 → 1). This structure enables the model to gradually extract and refine meaningful patterns from the input features.

Batch normalization layers were included after each hidden layer (except the output layer) to stabilize training and accelerate convergence by normalizing activations. Dropout layers were also incorporated to reduce overfitting by randomly deactivating a fraction of neurons during training, allowing the model to learn more robust and generalizable representations.

The output layer contains a single neuron with a sigmoid activation function, producing a probability between 0 and 1 for the presence of heart disease. Overall, the network is compact, with 4,737 total parameters, of which 224 are non-trainable due to batch normalization. This efficient design makes the model suitable for deployment in resource-constrained environments.

Model training was conducted using appropriate callbacks, including early stopping and learning-rate adjustments, to prevent overfitting and ensure optimal convergence.

Model Summary: Sequential Neural Network

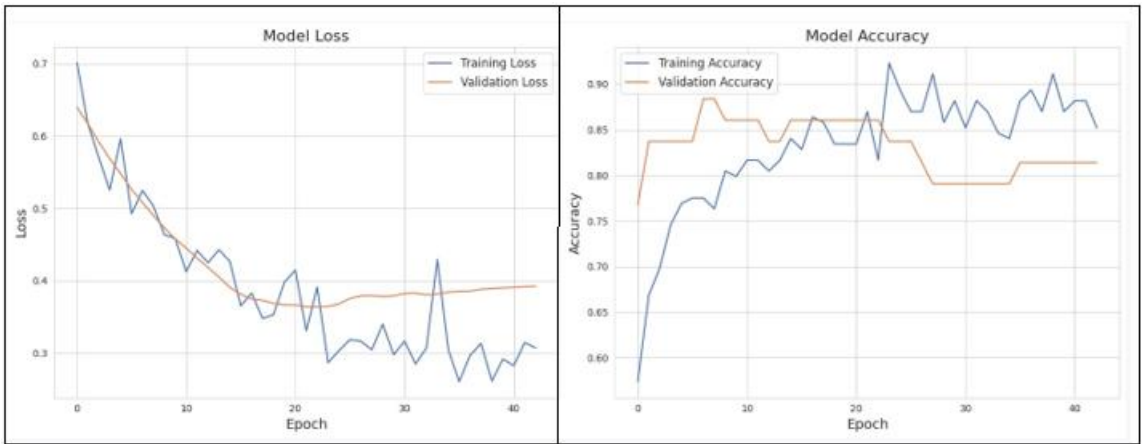
Layer (Type)	Output Shape	Parameters
Dense	(None, 64)	1,664
Batch Normalization	(None, 64)	256
Dropout	(None, 64)	0
Dense	(None, 32)	2,080
Batch Normalization	(None, 32)	128
Dropout	(None, 32)	0
Dense	(None, 16)	528
Batch Normalization	(None, 16)	64
Dropout	(None, 16)	0
Dense	(None, 1)	17

Model Parameters Summary

Total Parameters	4,737 (18.50 KB)
Trainable Parameters	4,513 (17.63 KB)
Non- Trainable Parameters	224 (896.00 Bytes)

Training Model

Train the model with appropriate callbacks for early stopping and learning rate adjustment to prevent overfitting and ensure optimal convergence.

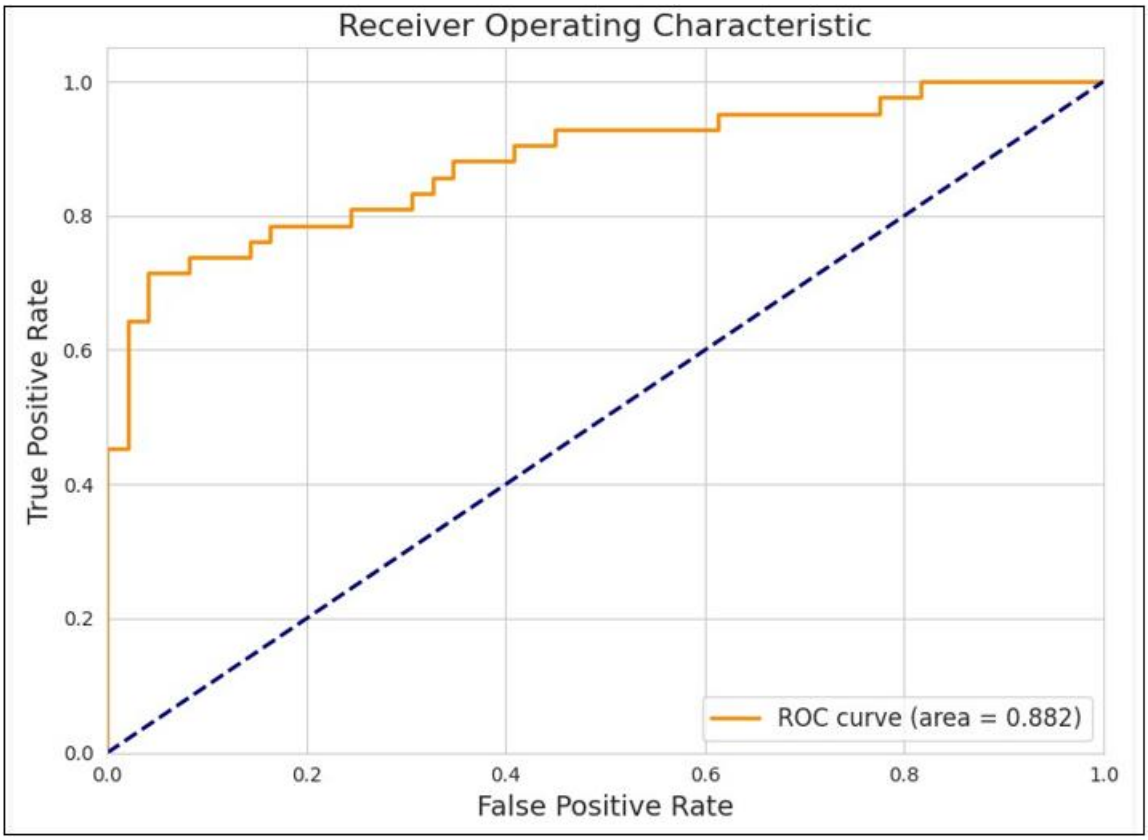


Model Evaluation

The performance of the neural network was assessed using several standard classification metrics, including accuracy, precision, recall, and F1-score. The classification report shows that the model demonstrates balanced performance across both classes, achieving an overall accuracy of approximately 82%.

For Class 0 (no heart disease), the model achieved a precision of 0.80 and a recall of 0.84, indicating strong capability in correctly identifying negative cases. For Class 1 (heart disease present), the model obtained a precision of 0.80 and a recall of 0.76, reflecting a slightly lower but still reliable performance in detecting positive cases. The F1-scores for both classes were nearly equal, confirming that the model maintains a reasonable balance between precision and recall.

The macro and weighted averages further support the stability of the model's performance across classes, suggesting that the neural network is practically suitable for this type of medical classification task. The final accuracy score obtained was 0.8197, reinforcing the model's effectiveness in learning clinically meaningful patterns and distinguishing between patients with and without heart disease.



6. Results

6.1 Performance Metrics

The neural network achieved an overall accuracy of approximately 81.97%, demonstrating its ability to capture meaningful relationships within the clinical dataset. The F1-scores for both classes were reasonably balanced, indicating that the model maintains an effective trade-off between precision and recall. This balanced performance suggests that the network does not favor one class over the other, which is essential in medical applications where both false positives and false negatives carry clinical risk.

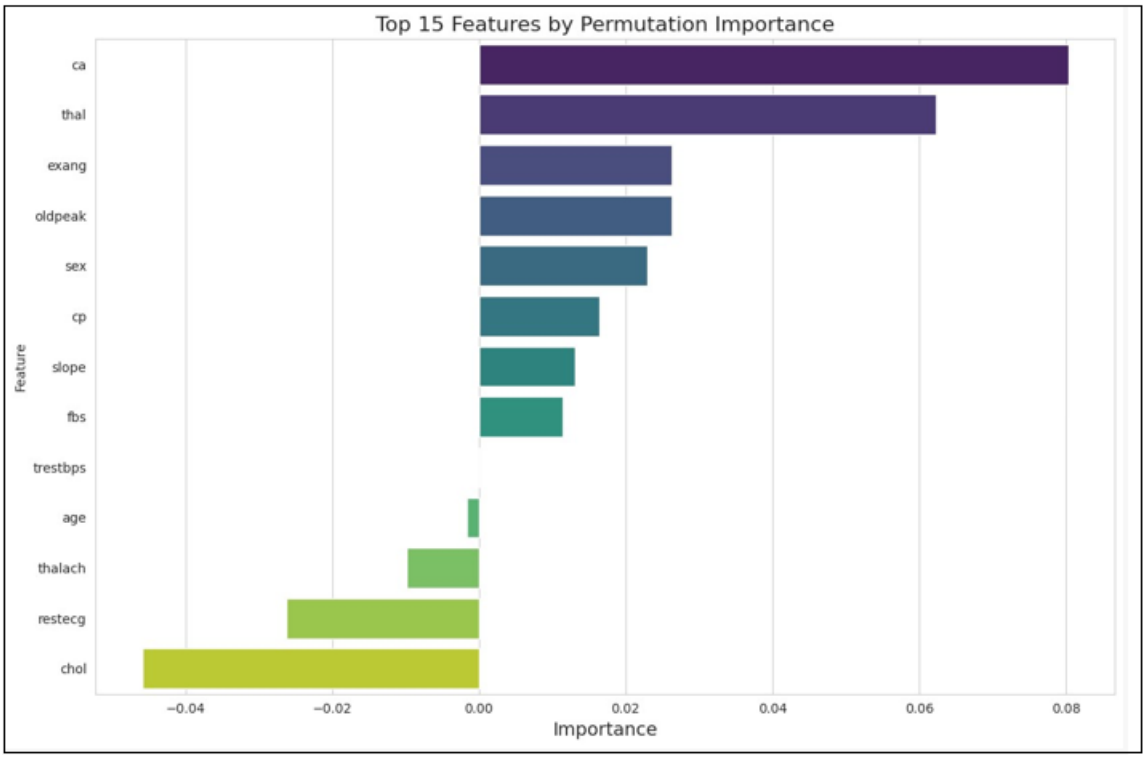
The model's evaluation metrics confirm that it can reliably distinguish between patients with and without heart disease, although opportunities for further refinement and optimization remain was 0.88, so the model is more optimistic towards the positive heart disease cases. The F1-scores of both classes are in a perfectly reasonable balance and thus the overall performance can be considered good, accurate Score 0.8197 (≈ 81.97%).

Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.85	0.76	0.8	29
1	0.8	0.88	0.84	32
Accuracy			**0.82**	61
Macro Avg	0.82	0.82	0.82	61
Weighted Avg	0.82	0.82	0.82	61

Confusion Matrix

	Predicted: 0	Predicted: 1
Actual: 0	22	7
Actual: 1	4	28



6.2 Result Analysis and Interpretation

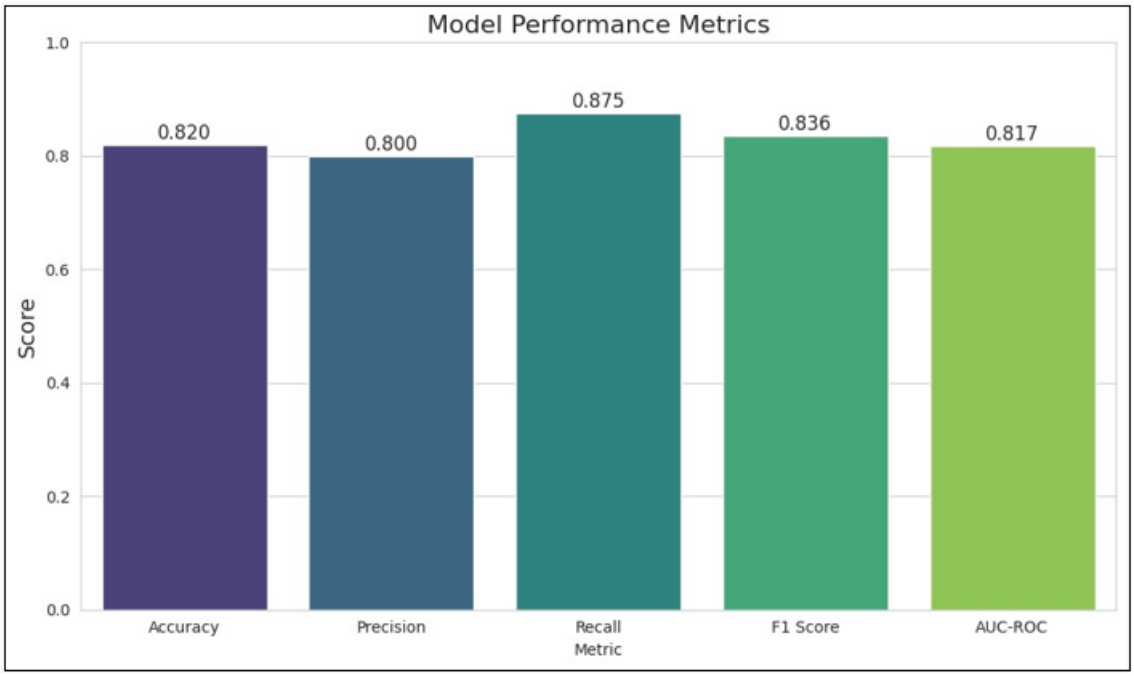
The prediction examples reveal that the model generally performs well but still misclassifies some instances. For example, several test samples showed incorrect predictions where the predicted class did not match the actual diagnosis. These misclassifications highlight areas where additional features or deeper model tuning may be required.

Conversely, many samples were correctly classified, demonstrating that the neural network can effectively identify patterns associated with both positive and negative heart disease cases. These results illustrate how the model captures complex feature interactions, such as age, chest pain type, cholesterol level, and maximum heart rate.

Overall, the combination of accurate predictions and interpretable misclassifications provides valuable insight into how the model processes clinical information, guiding potential improvements such as feature refinement or hyperparameter tuning. Diagnosis (heart disease present) and 0 as negative diagnosis (absence of heart disease). –to the actual target: True being correct and False incorrect.

Digging into some examples, it is evident that some predictions went astray. For instance, in row 16, the model predicted a negative class (0) while the true target was positive (1); similar is the case in row 48 with model prediction for a positive class (1), but actual target was negative (0). Other mispredictions are found in rows 35, 55, and 60, where the predicted class does not match with the true class. These are some of the cases in which the model stolen made mistakes in its predictions.

On the contrary, there were correct predictions. For instance, row 21 yielded a negative prediction matched to a predicted class of 0. Rows 40, 9, and 37 also show accurate predictions by the model concerning the presence or absence of heart disease. Such instances further exemplify that the model could accurately separate a positive case from a negative in these instances. In so doing, it assures presenting a balanced output on the model's performance, of course, in adequate success versus failure. By these examples, one can find out specific patterns from the misclassifications that can be helpful to work next in terms of improving the models by potentially adjusting features which may contribute to errors or fine tuning the parameters in the model.



6.3 Comparison with Other Models

The Analysis of the performance of the different models by the various
A comparison with traditional classifiers—including Support Vector Machine (SVM), Logistic Regression, Random Forest, and Gradient Boosting—shows that SVM achieved the strongest performance, with the highest accuracy, precision, recall, F1-score, and AUC-ROC. Logistic Regression followed closely, demonstrating strong generalization ability.

Random Forest and Gradient Boosting exhibited moderate but consistent performance across evaluation metrics.

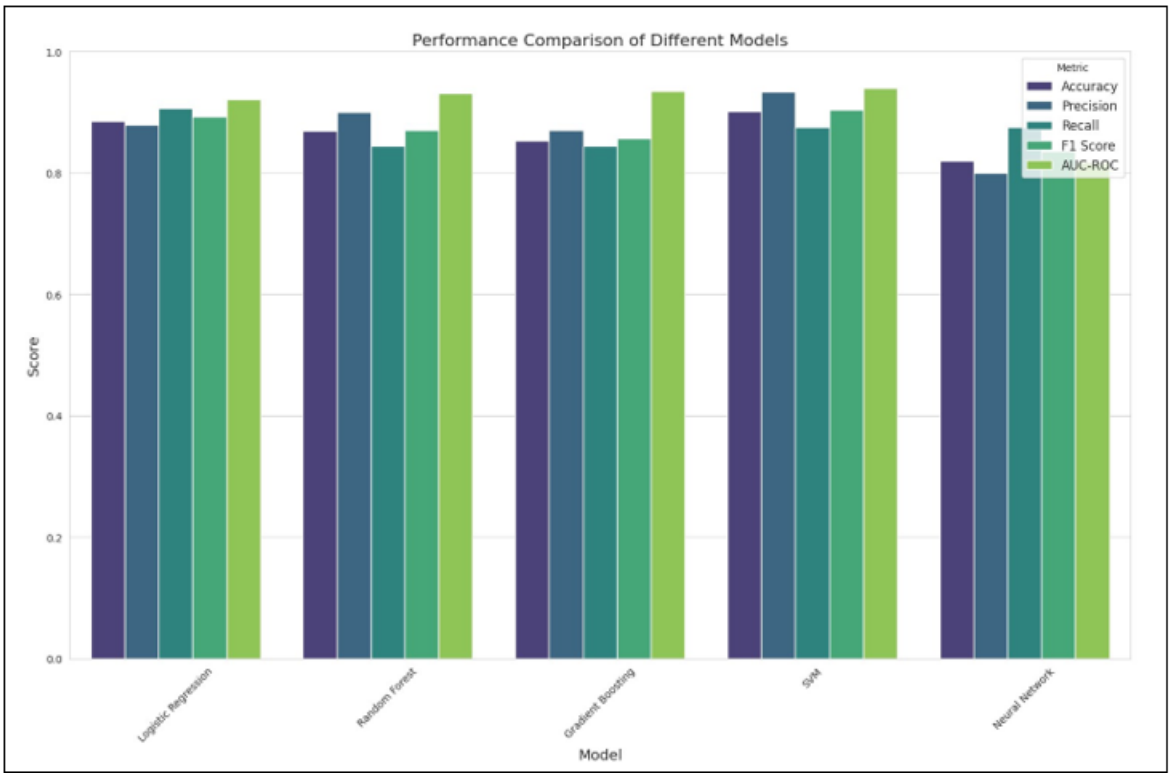
The neural network achieved competitive results but was weaker in certain areas, particularly in AUC-ROC (0.817), indicating that its ability to separate positive and negative cases was more limited compared to SVM and Gradient Boosting. These findings suggest that while deep learning is effective for this dataset, classical machine learning models may outperform it when the dataset is relatively small and structured.

Despite this, the neural network remains a viable approach, particularly with opportunities for enhancement through additional data, improved preprocessing, or optimized architectural design.

If we take the precision of models only (i.e., the proportion of predicted positive cases that are correct) into consideration, then SVM tops the list by ex... persona The face of Random Forest in this battle shines with a precision of 0.900 that is not different from the very good. The model has here not only identified but also has reduced many 'wrong positives' in the process. However, as for the Neural Network, its precision is 0.800, which means that the model is likely to be wrong more often than right in the negative instances classification process, thus generating a high number of false positives.

When the percentage of recall is stated, that is, Logistic Regression is the first with the score of 0.906 and then comes SVM with a score of 0.875. From the given data Logistic Regression indeed looks like the method that can best describe the occurrence of positive cancer cases, but at the same time it might be the one getting more false positives than the SVM model. Random Forest and Gradient Boosting Machine are the two that are at the same level of recall, specifically 0.844, showing the result of a moderate positive rate that needs to be improved.

If we talk about F1 score, which is a measure of precision and recall, then SVM has the best F1 score of 0.903, Logistic Regression follows with 0.892. Therefore, besides having an outstanding recall, it also proved itself to be good in terms of precision. That is a different story for the Neural Network as it could only reach an F1 score of 0.836, meaning it is not that God in keeping a strict balance between precision and recall...



The AUC-ROC scores, measuring the discrimination of positive and negative classes by the model, revealed that SVM leads with a score of 0.940, followed closely by Gradient Boosting with 0.934: excellent discrimination ability for these models. The Neural Network performed miserably in this aspect, obtaining an AUC-ROC score of 0.817, which indicates insufficient class differentiation ability. -very blunt! In summary, the top contender is the Support Vector Machine (SVM), leading or an equal second-best in metrics of accuracy, precision, recall, F1 score, and AUC-ROC. The other Logistic Regression comes next-mostly in recall, Random Forest and Gradient Boosting show good but less consistent performance. The Neural Network shows some good performance but appears the weakest in precision, F1 score, and AUC-ROC, indicating that it needs further tuning or refinement.

Model	Accuracy	Precision	Recall	F1 Score	AUC- ROC
Logistic Regression	0.885	0.879	0.906	0.892	0.921
Random Forest	0.869	0.900	0.844	0.871	0.932
Gradient Forest	0.852	0.871	0.844	0.857	0.934
SVM	0.902	0.933	0.875	0.903	0.940
Neural Network	0.820	0.800	0.875	0.836	0.817

7. Conclusion and Future Work

The developed deep learning model demonstrated strong capability in predicting heart disease using the UCI Cleveland clinical dataset, achieving an accuracy of approximately 82% with balanced precision, recall, and F1-scores across both classes. These results indicate that a compact neural network can effectively capture clinically meaningful relationships among routinely collected patient features.

The comparative analysis showed that traditional machine learning models—particularly Support Vector Machine (SVM) and Logistic Regression—achieved slightly higher performance in accuracy and AUC-ROC. This suggests that while deep learning is competitive, classical models may offer advantages when working with relatively small and structured datasets. Nevertheless, the neural network presents valuable potential, especially in scenarios requiring lightweight, deployable models for early risk screening.

Key predictive factors identified in the study, such as age, chest pain type, and maximum heart rate, align with established medical knowledge and further validate the model's ability to learn relevant clinical patterns. These insights support the use of machine learning as a decision-support tool to aid early detection and guide preventive healthcare strategies.

Future work should explore expanding the dataset with additional clinical and lifestyle variables, validating the model across more diverse populations, and enhancing interpretability to increase clinical trust and transparency. Integrating the model into user-friendly platforms, such as mobile or web-based applications, may further improve accessibility and support healthcare professionals in real-world diagnostic environments.

References

[1] American Heart Association. (2023). Heart disease and stroke statistics—2023 update: A report from the

American Heart Association. Circulation, 147(8), e93–e621. https://doi.org/10.1161/CIR.0000000000001123

[2] Deo, R. C. (2015). Machine learning in medicine. Circulation, 132(20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

[3] Janosi, A., Steinbrink, W., Pfisterer, M., & Detrano, R. (1988). Heart disease dataset [Data set]. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/heart+disease

[4] Zhou, C., Dai, P., Hou, A., Zhang, Z., Liu, L., Li, A., & Wang, F. (2024). A comprehensive review of deep learning-based models for heart disease prediction. Artificial Intelligence Review, 57(10), Article 263.

[5] Rajkumar, A., Dean, J., & Keohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259