Heart Failure Prediction

Authors: Rotem Geva and Diana Gurevich

Keywords: Heart Failure Prediction, Machine Learning, Clinical Data Analysis, Cardiovascular Health, Predictive Modeling, Deep Learning

Table of Contents

Abstract

Heart failure remains a significant global health concern, necessitating accurate and early prediction to improve patient outcomes. This project leverages the Heart Failure Prediction dataset from Kaggle, which comprises clinical and demographic data of patients, including attributes such as age, gender, cholesterol levels and HR data. Utilizing machine learning models, the study aims to predict the likelihood of heart failure with high accuracy. The results of this work could assist healthcare professionals in identifying high-risk individuals and implementing preventive strategies effectively.

Introduction

Heart failure is a chronic condition characterized by the heart's inability to pump blood effectively, leading to reduced oxygen supply to the body's tissues. It is a leading cause of morbidity and mortality worldwide, significantly impacting patients' quality of life and placing a substantial burden on healthcare systems. Early detection of heart failure risk is crucial for timely intervention, improved prognosis, and reduced healthcare costs.

Machine learning has emerged as a powerful tool for analyzing complex medical data and identifying patterns that may not be evident through traditional methods. The Heart Failure Prediction dataset from Kaggle offers an extensive collection of clinical and demographic features, such as age, gender, blood pressure, cholesterol levels, and smoking status. This dataset provides an excellent foundation for building predictive models that can estimate the likelihood of heart failure based on patient-specific characteristics.

In this project, we aim to develop and evaluate machine learning algorithms to predict the risk of heart failure. By leveraging the diverse features in the dataset, we seek to identify critical factors contributing to heart failure and construct a model that can aid healthcare professionals in making data-driven decisions for patient care. The insights from this study have the potential to enhance preventive care and improve clinical outcomes for individuals at risk of heart failure.

Literature Review

Framingham Heart Study by Lucy Soto

Initiated in 1948 in Framingham, Massachusetts, the Framingham Heart Study is a long-term, multigenerational study **designed to identify common factors contributing to cardiovascular disease** (CVD). Researchers enrolled over 5,000

participants who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke. Through extensive data collection and analysis over several decades, the study has significantly advanced our understanding of heart disease**. Key findings include the identification of major risk factors such as high blood pressure, high cholesterol levels**, smoking, obesity, and physical inactivity. The study also refined the understanding of "good" (HDL) and "bad" (LDL) cholesterol and established the Framingham Risk Score, a tool used to estimate an individual's 10-year risk of developing coronary heart disease.

Seven Countries Study by Ancel Keys

Launched in the late 1950s by Ancel Keys, the Seven Countries Study was the first major research endeavor to investigate the relationship between diet, lifestyle, and cardiovascular disease across different populations and cultures. The study examined cohorts in seven countries: the United States, Finland, the Netherlands, Italy, Yugoslavia, Greece, and Japan. Over an extended follow-up period, researchers found significant differences in heart disease rates among these populations. **A key finding was the direct correlation between the level of total serum cholesterol and the risk of heart attack and stroke**, both at the population and individual levels. The study also highlighted the impact of dietary patterns, particularly the benefits of the Mediterranean diet, on heart health.

Predicting Heart Disease Using Machine Learning Techniques: A Comprehensive Study

This study evaluates the effectiveness of various machine learning models in predicting heart disease using a dataset from Kaggle, which includes 1,190 instances with 11 clinical features. The models tested include Logistic Regression, Random Forest, XGBoost, Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), and a Voting Classifier ensemble model. Grid Search hyperparameter tuning was applied to optimize model performance. The Voting Classifier achieved the highest accuracy (95.4%), outperforming other models such as Random Forest (94.5%) and XGBoost (93%), demonstrating the power of ensemble learning. Soft Voting, feature scaling, and hyperparameter tuning were key factors in performance improvement. The study concludes that combining multiple models enhances prediction accuracy, making machine learning a promising tool for heart disease diagnosis and clinical decision support.

Methodology

The study began with **Exploratory Data Analysis (EDA)** to understand the data and uncover underlying patterns. We started by identifying variable types (numerical and categorical), checking for missing values, duplicate rows, and outliers.

Some columns contained categorical data, while others were numerical. To standardize the dataset, we applied label encoding to the categorical columns.

Then we performed a **univariate analysis** to explore the distribution of each variable and calculate key statistical measures. To detect outliers, we used boxplots, which revealed that the **Cholesterol** column contained numerous widely spread outliers and several zero values. Since a cholesterol level of zero is physiologically impossible, these zeros likely represent missing data.

Given that missing values accounted for approximately 20% of the dataset, our initial approach was to apply various imputation techniques, such as filling the missing values with the median or mean, generating random values based on the distribution, and using the KNN algorithm. However, **all these methods resulted in an inaccurate distribution** that negatively impacted the results. Consequently, **we decided to remove all rows containing missing data.**

Regarding the **RestingBP** column, higher systolic pressure is common in individuals with heart disease. Therefore, it is reasonable to assume that most people in the dataset have relatively high but normal blood pressure, while some exhibit even higher values**. Given its clinical relevance, we decided to retain these values** as they naturally occur.

Since age is likely to relate to heart disease as a broader age-group problem rather than a specific age, **we convert the continuous age variable into four bins**, using 4 quartiles.

Multivariate analysis was conducted next, revealing a strong negative correlation between ST_Slope and OldPeak. However, since these two features showed opposite correlations with the target variable, it was decided not to combine them. The remaining variables showed some correlation, but they were not significant enough to warrant any further action at the moment.

To address scale differences among numerical variables, **standardization** was applied.

Given the dataset's relatively small size, we split it into a **60-20-20 ratio** for training, validation, and testing.

We applied two models:

1. **Logistic Regression**, with performance evaluated using recall, precision, F1 score, and accuracy.
2. **A fully connected neural network** with one hidden layer consisting of 16 neurons, the default activation function and learning rate of 0.001.

Multiple experiments were conducted, each varying one of the following hyperparameters: the number of neurons, learning rate, and activation function.

We also tested the following data scenarios:

1. **Improved Data:** Outliers were removed, and the target variable classes were balanced. We re-ran the initial model architecture under these conditions.
2. **Adverse Data:** We artificially worsened data quality by deleting columns correlated with the target variable to simulate a suboptimal scenario and assess model performance.

We also experimented with **enhanced architecture for the fully connected neural network** by incorporating weight decay, adjusting the batch size and number of epochs, and changing the activation function to improve performance. To further boost the network's performance, we designed the network with 12 neurons in the hidden layer, using the sigmoid activation function for this layer.

The Matthews Correlation Coefficient (**MCC**) is a robust metric **used to evaluate the performance of classification models**, particularly in imbalanced datasets. Unlike accuracy, which can be misleading when class distributions are skewed, MCC considers true positives, true negatives, false positives, and false negatives, providing a more balanced assessment. It ranges from -1 to +1, where +1 indicates perfect classification, 0 represents random guessing, and -1 suggests complete misclassification.

To investigate the effects of data balancing on the basic neural network, we created three levels of data balancing: unbalanced data, moderate balanced data and fully balanced data.

Finally, we applied **Principal Component Analysis (PCA)** and retained five components to reduce dimensionality while preserving most of the variance.

Results

ML Model - Logistic Regression

We applied a **machine learning model of logistic regression** to the dataset. The performance of the model was evaluated using a confusion matrix, along with key metrics such as accuracy, precision, recall, and F1-score.

Logistic Regression Metrics Results table presents the **precision, recall, and F1-score** for a binary classification model. Both classes show relatively high precision, recall, and F1-scores. There's a slight trade-off between precision and recall for each class. Class 0 has higher recall but lower precision, while class 1 has higher precision but lower recall. The F1-scores are quite close, indicating a balanced performance across both classes.

Accuracy: 85%

Table 0-1 Logistic Regression Metrics Results

	Precision	Recall	F1 Score
0	82%	89%	85%
1	89%	82%	86%

Logistic Regression Confusion Matrix has a higher number of False Negatives (14) compared to False Positives (8).

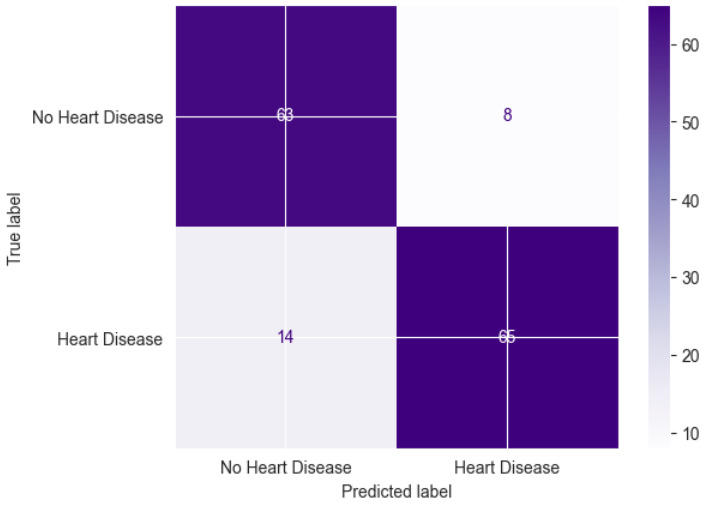

Figure 0-1 Logistic Regression Confusion Matrix

Basic Neural Network

We used a **basic neural network** with 16 neurons in the hidden layer, with a default activation function (None) and learning rate of 0.001.

Basic Neural Network Training and Validation shows that both the training and validation loss decreased with the number of epochs. The validation loss starts to plateau after 10 epochs indication the model is no longer improving on the validation set. The noticeable gap between the two graphs indicates a slight overfit.

Figure 0-2 Basic Neural Network Training and Validation

Basic Neural Network Confusion Matrix shows a model with relatively good performance in predicting heart disease, but with a higher number of false negatives (14) than false positives (8). Accuracy: 85%.

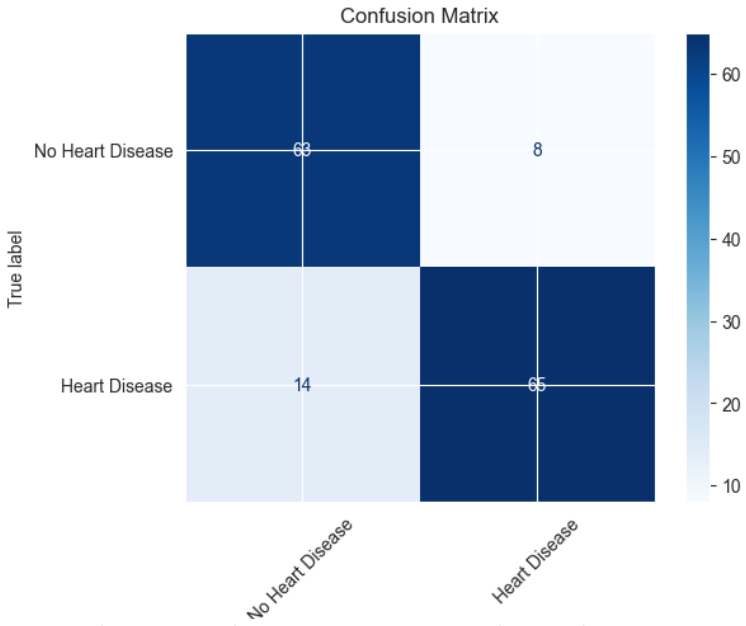

Figure 0-3 Basic Neural Network Confusion Matrix

Hyperparameters Fine-Tune

After the neural network assembly, we conducted 9 experiments on the network by changing one of the following hyperparameters:

- **Learning rate variations**:
  - 0.1
  - 0.01
  - 0.001
- **Number of neurons:**
  - 4
  - 16
  - 100
- **Activation functions:**
  - Sigmoid
  - ReLU
  - Tanh

Learning Rate:

Learning rate: 0.1:

Accuracy: 82%

Exp1 – Training and Validation Loss reveals significant fluctuations in the validation loss, indicating instability and potential overfitting, while the training loss shows a more stable, albeit less smooth, decrease. The large spikes in validation loss suggest the model's performance on unseen data is inconsistent and unreliable. Exp1 - Confusion Matrix shows a concerning imbalance with more false negatives (18) than false positives (8).

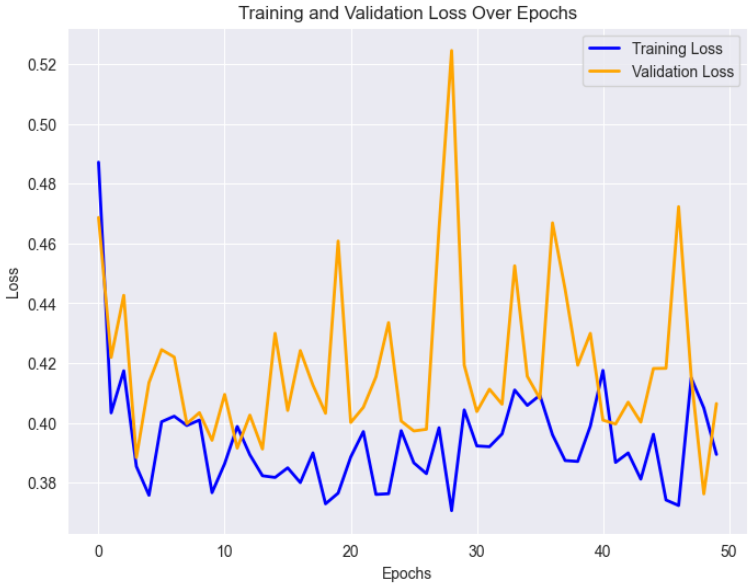

Figure 0-4 Exp1 – Training and Validation Loss

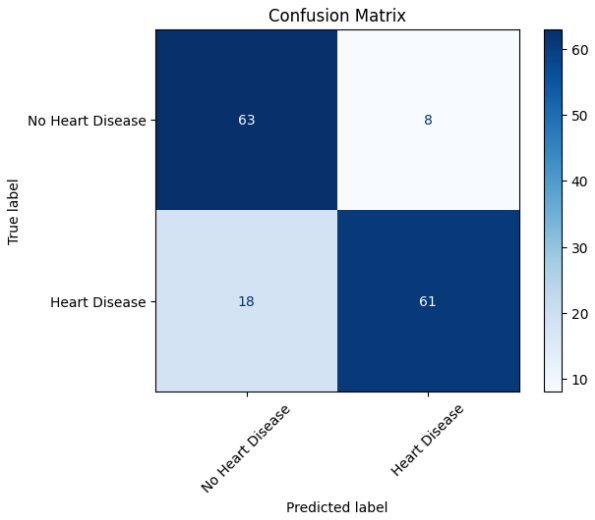

Figure 0-5 Exp1 - Confusion Matrix

Learning rate: 0.01:

Accuracy: 85%

Exp2 – Training and Validation Loss graph demonstrates a rapid initial decrease in training loss followed by a plateau, while the validation loss shows a slight initial dip and then stabilizes at a higher level, suggesting potential overfitting. The confusion matrix is identical to the one above.

Figure 0-6 Exp2 – Training and Validation Loss

Learning rate: 0.001:

Accuracy: 85%

Exp3 – Training and Validation Loss reveals a rapid decrease in both losses initially, followed by a plateau, with the training loss consistently lower than the validation loss, indicating potential overfitting. Exp3 Confusion Matrix shows a concerning imbalance with more false negatives (14) than false positives (8).

Figure 0-8 Exp3 – Training and Validation Loss

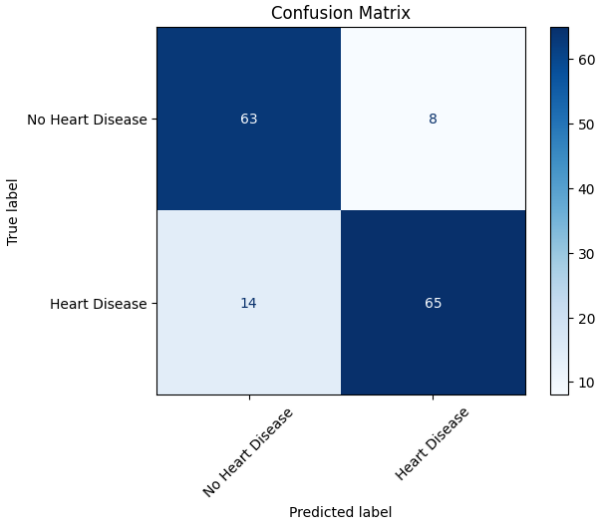

Figure 0-7 Exp3 Confusion Matrix

It can be seen that the higher the learning rate, the noisier the graphs.

Number of Neurons:

Number of neurons: 4

Accuracy: 84%

Exp4 - 4 Neurons graph shows a consistent and converging decrease in both losses, with the validation loss slightly above the training loss.

Figure 0-9 Exp4 - 4 Neurons

Number of neurons: 16

Accuracy: 85%

Exp5 - 16 Neurons graph shows a rapid initial decrease in both losses, followed by a plateau, with the validation loss consistently slightly higher than the training loss.

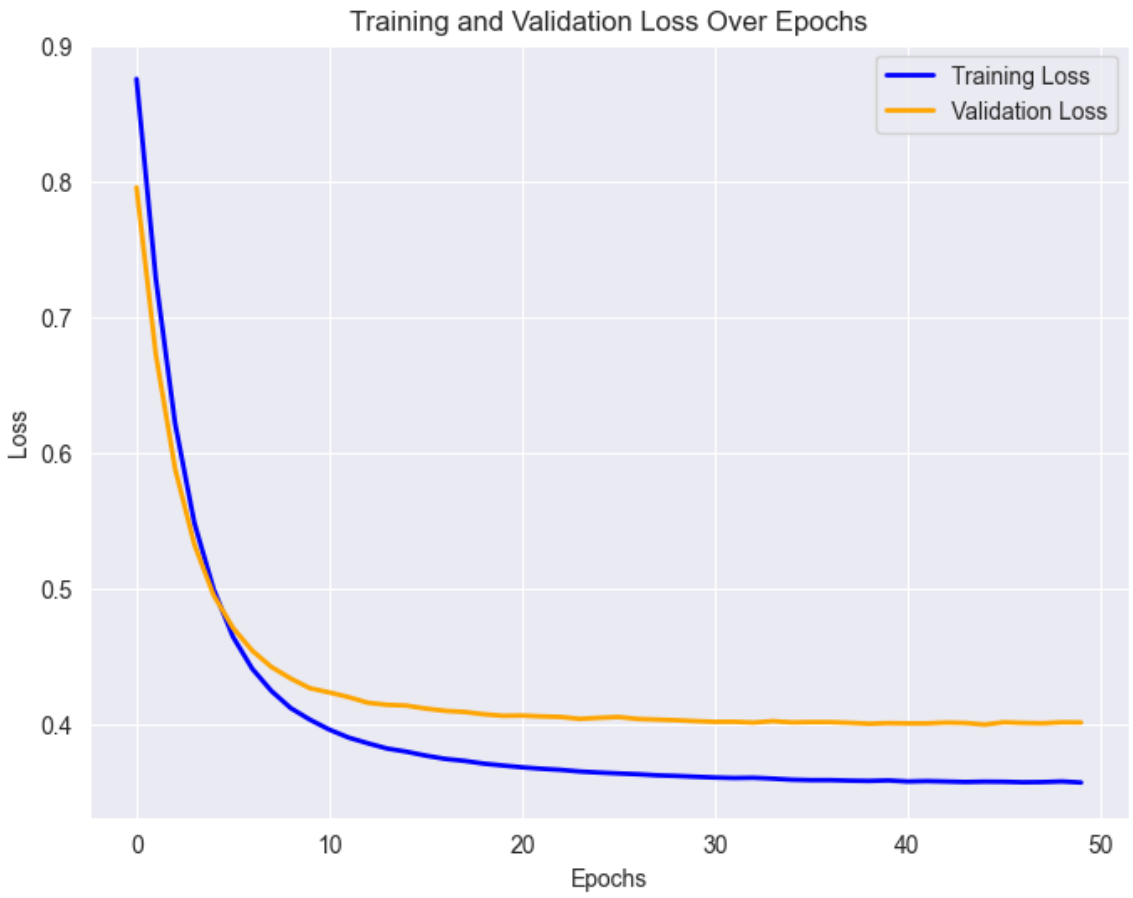

Figure 0-10 Exp5 - 16 Neurons

Number of neurons: 100

Accuracy: 85%

Exp6 - 100 Neurons graph shows a rapid initial decrease in training loss followed by a plateau, while the validation loss shows a slight initial dip and then stabilizes at a noticeably higher level, strongly suggesting overfitting.

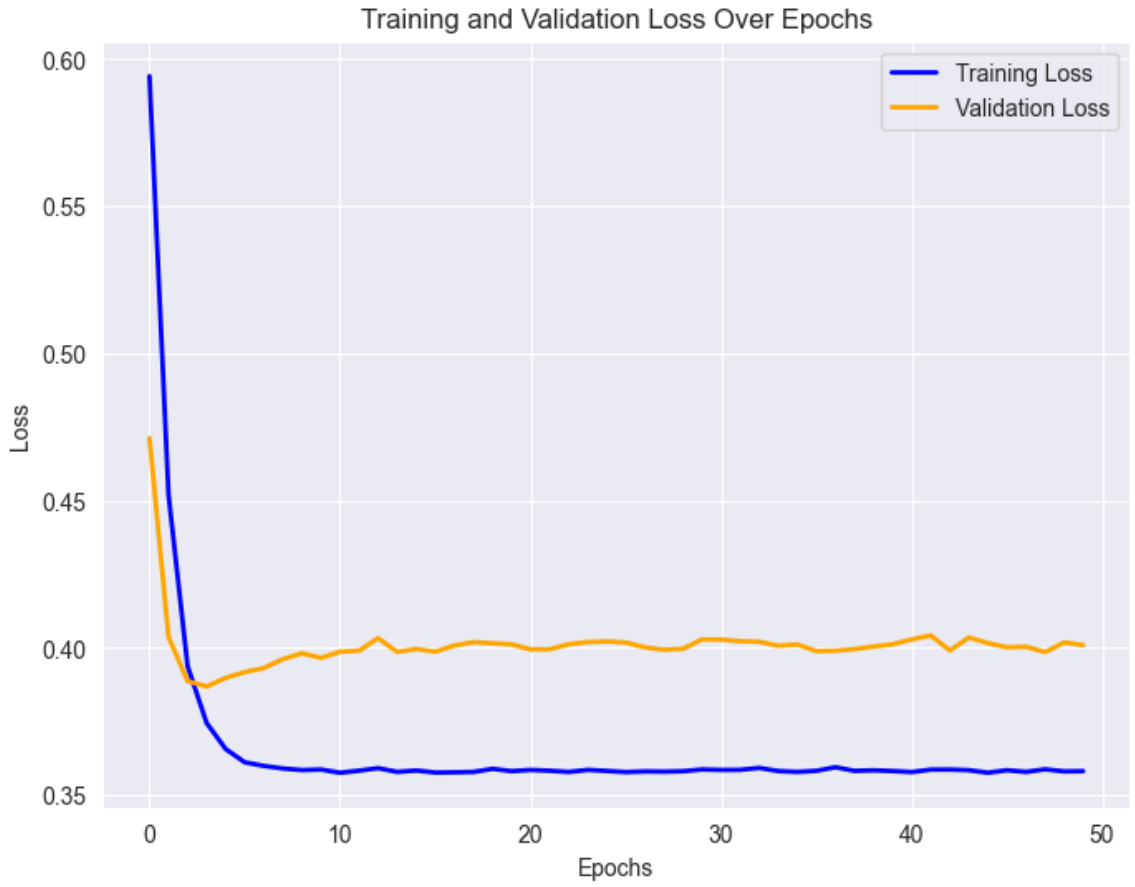

Figure 0-11 Exp6 - 100 Neurons

The confusion matrices are pretty similar between the three experiments.

It can be seen that the higher the number of neurons, the higher the validation loss from the training loss, which can be a sign of overfit. Also, the higher the number of neurons the lower the training graph and the lower the training starting loss.

Activation Function:

Activation Function: None

Accuracy: 85%

Exp7 - Default Activation Function graph shows both losses decreasing rapidly initially and then plateauing, but the validation loss consistently slightly higher than the training loss.

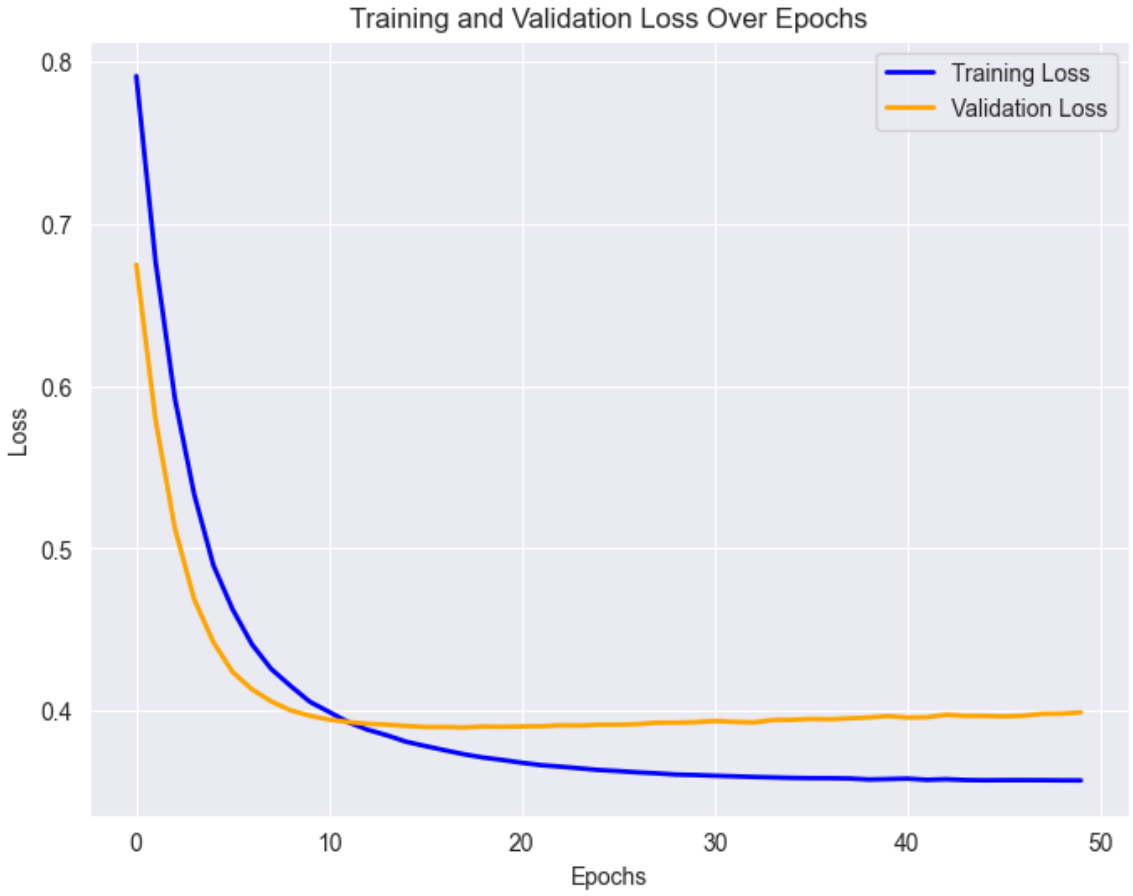

Figure 0-12 Exp7 - Default Activation Function

Activation Function: ReLU

Accuracy: 86%

Exp8 - ReLU Function graph demonstrates a consistent decrease in both losses with the validation loss slightly above the training loss throughout.

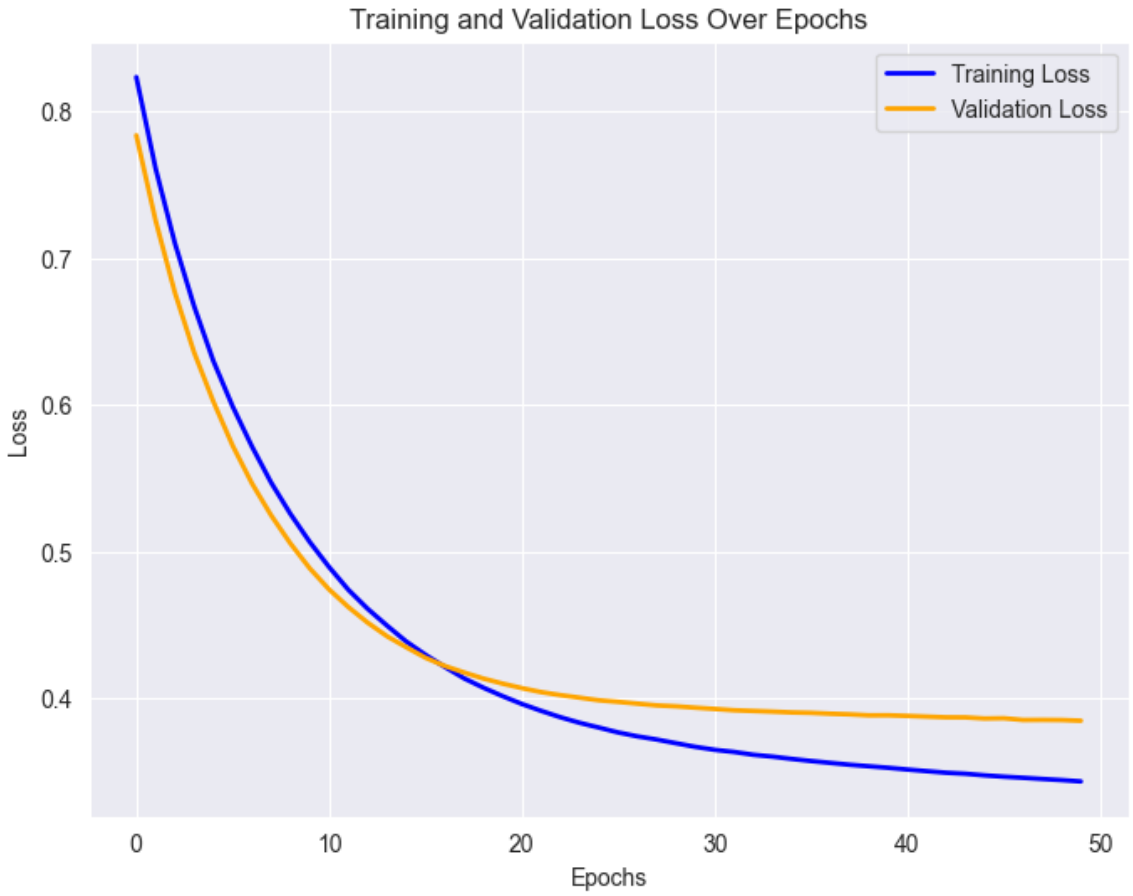

Figure 0-13Exp8 - ReLU Function

Activation Function: Tanh

Accuracy: 84%

Exp9 - Tanh Function graph shows a consistent decrease in both losses with the validation loss slightly above the training loss.

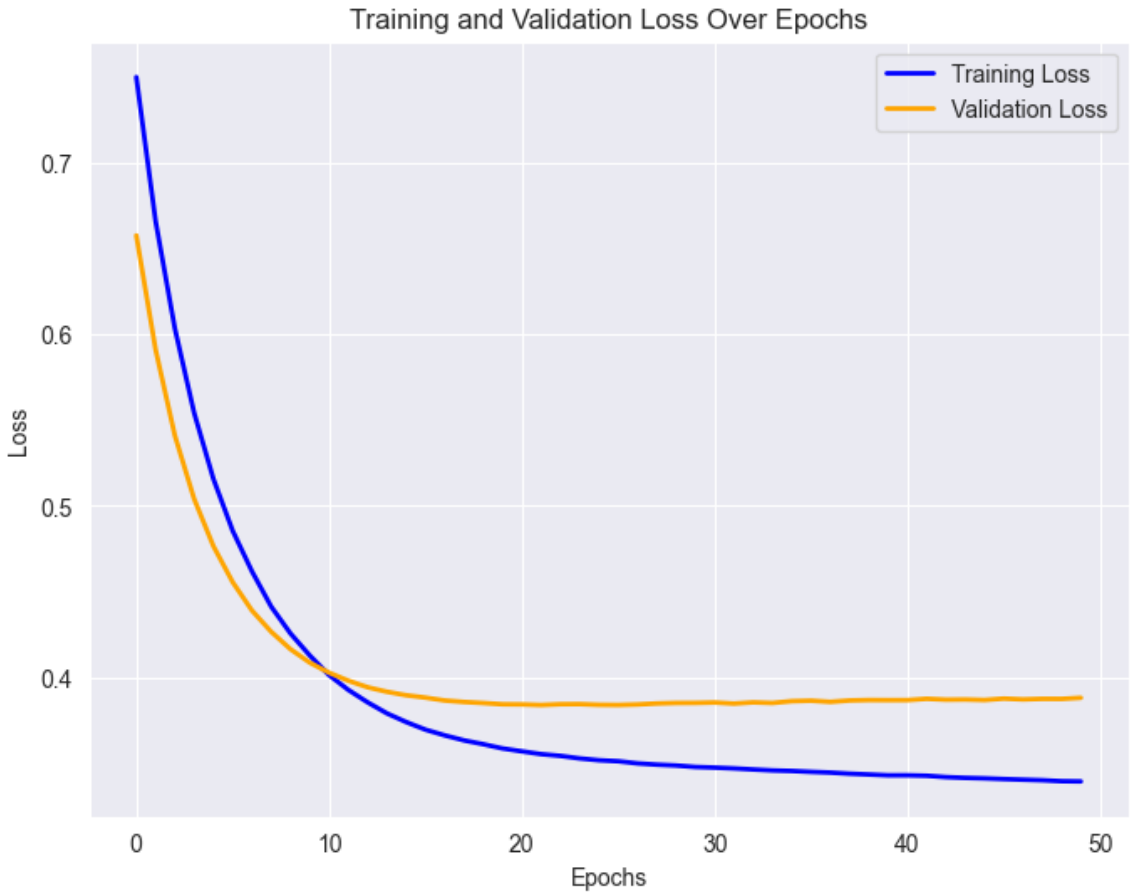

Figure 0-14 Exp9 - Tanh Function

No significant change between the experiments. Confusion matrices are very similar, a bit better for the ReLU function.

Modified Dataset

To explore the influence of the data on the network performance, we examined two data scenarios:

- **Improved Data** – we removed outliers and balanced the classes – accuracy: 91%.
- **Bad Data** – we removed key features – accuracy: 77%.

Improve Dataset Scenario Training and Validation Loss shows both training and validation losses decrease initially and then level off, the validation loss remains slightly higher, hinting at the possibility of mild overfitting.

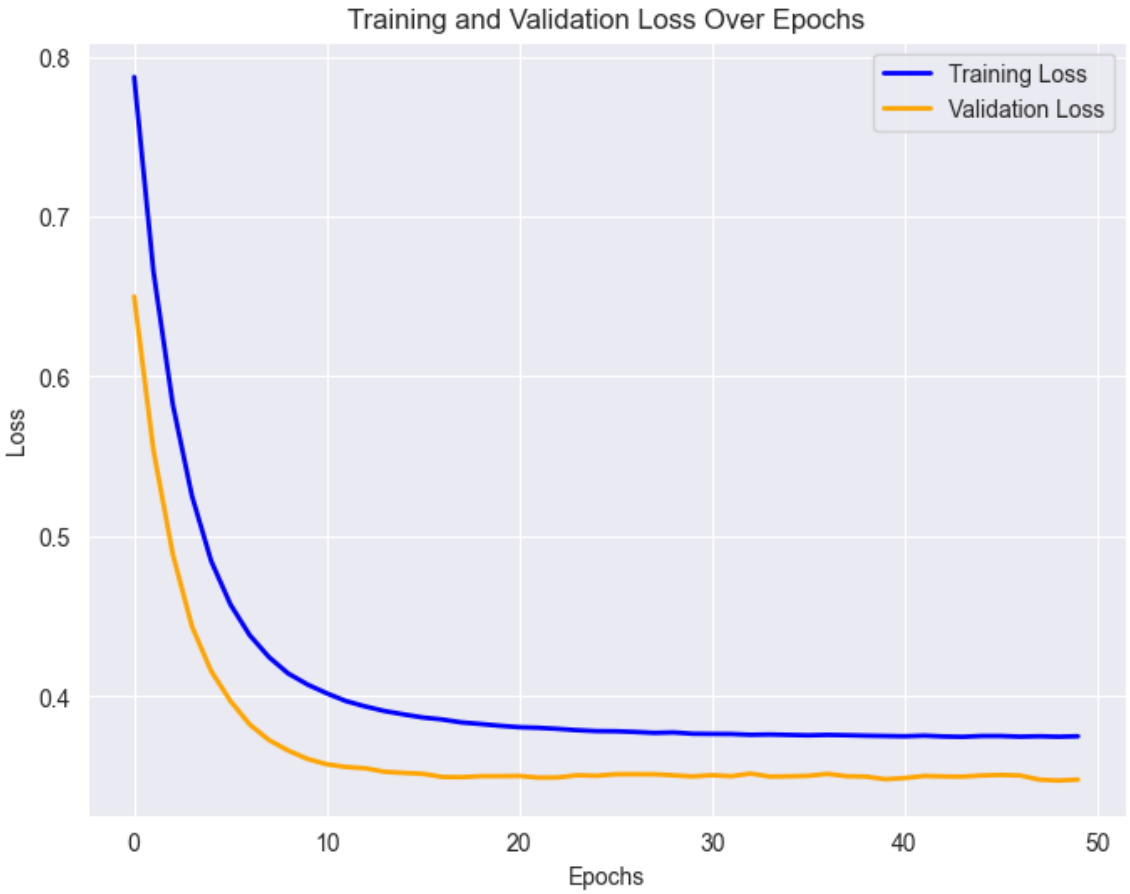

Figure 0-15 Improve Dataset Scenario Training and Validation Loss

Improved Dataset Scenario Confusion Matrix showing 63 true negatives and 65 true positives, but also 4 false positives and 8 false negatives. The model shows good overall accuracy but has more false negatives than false positives.

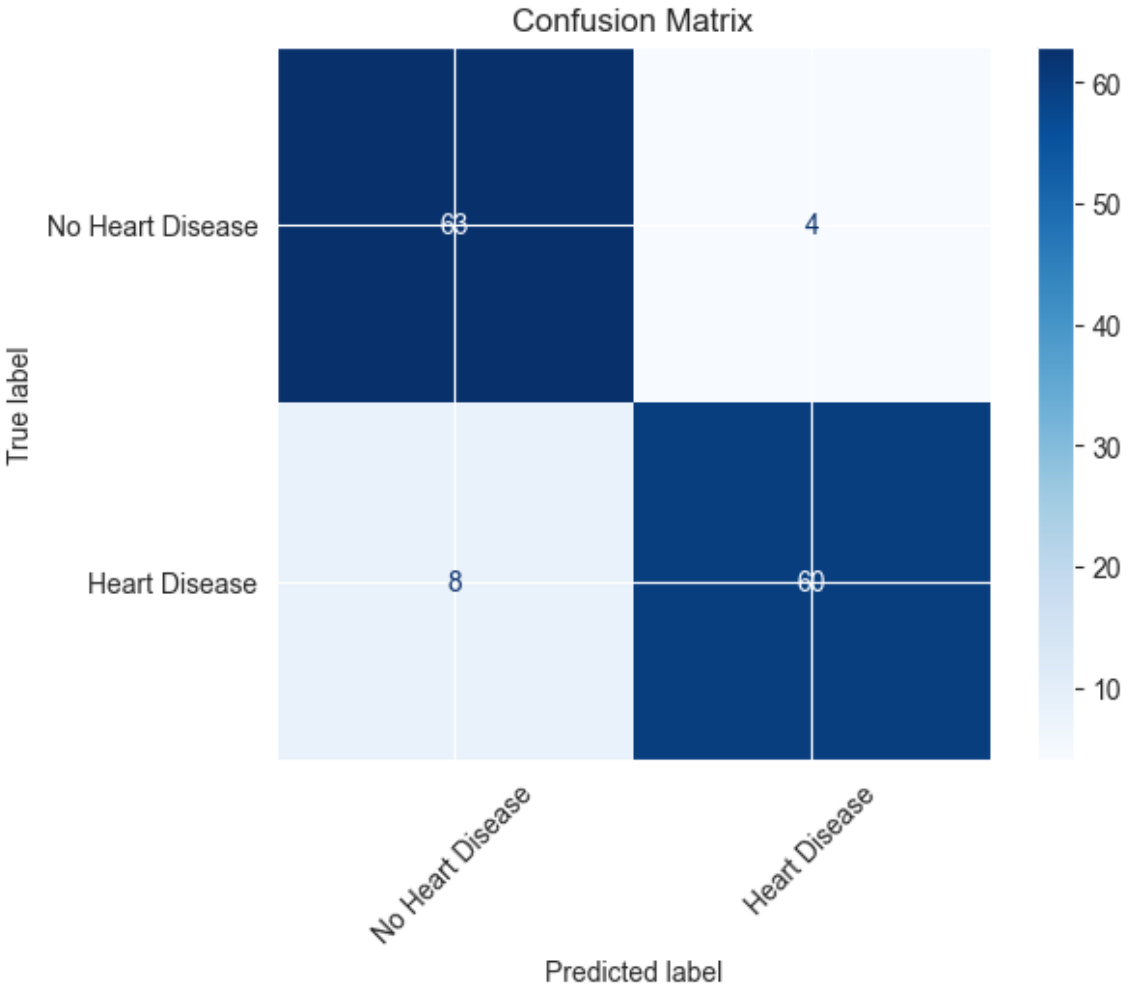

Figure 0-16 Improved Dataset Scenario Confusion Matrix

Bad Dataset Scenario - Training and Validation Loss shows that the model's training and validation loss decreases as the number of epochs increases. The convergence and close proximity of the loss curves indicate a successful training process.

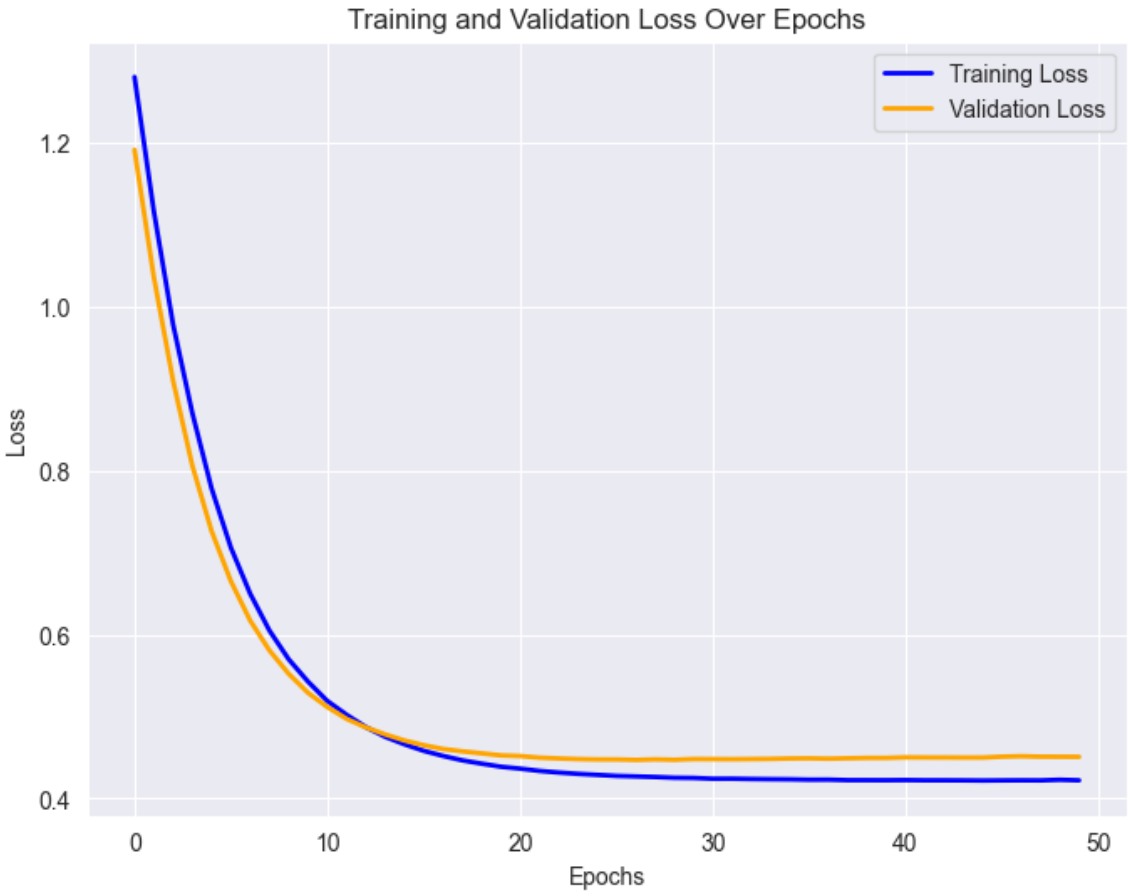

Figure 0-17 Bad Dataset Scenario - Training and Validation Loss

Bad Dataset Scenario - Confusion Matrix reveals 59 correctly identified individuals without heart disease and 57 correctly identified individuals with heart disease, but also 12 false positives and 22 false negatives, indicating a higher error rate for misclassifying those with heart disease as not having it.

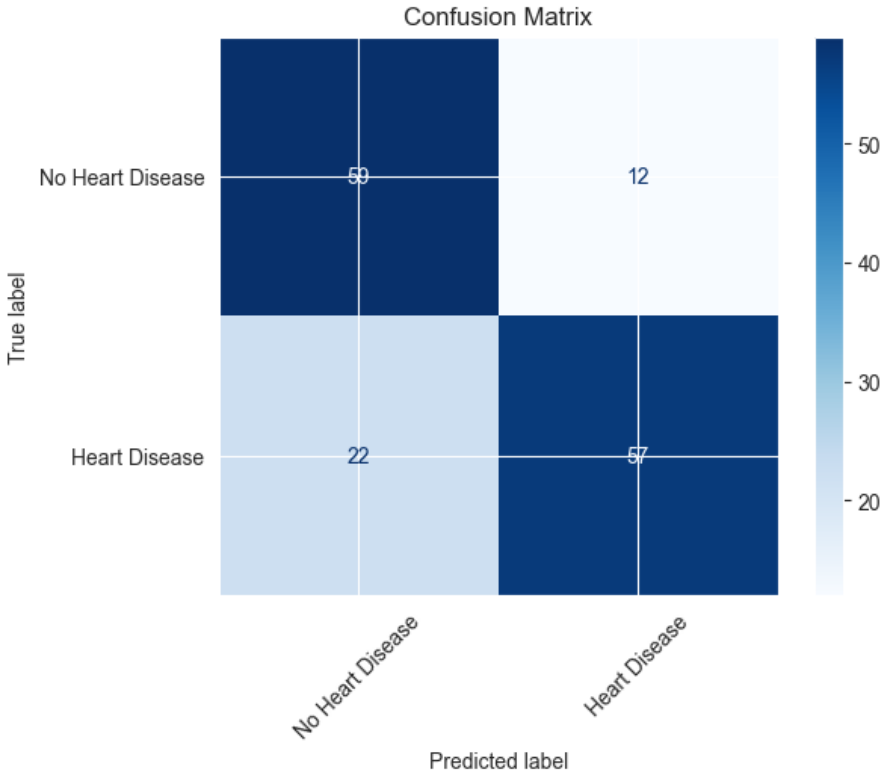

Figure 0-18 Bad Dataset Scenario - Confusion Matrix

Improved Model Architecture

Improved Model Scenario Training and Validation Loss graph displays the training and validation loss over 40 epochs, showing a **rapid** initial decrease in both losses followed by a gradual plateau, with the validation loss consistently above the training loss, suggesting good learning with a slight possibility of overfitting that could be addressed with further tuning.

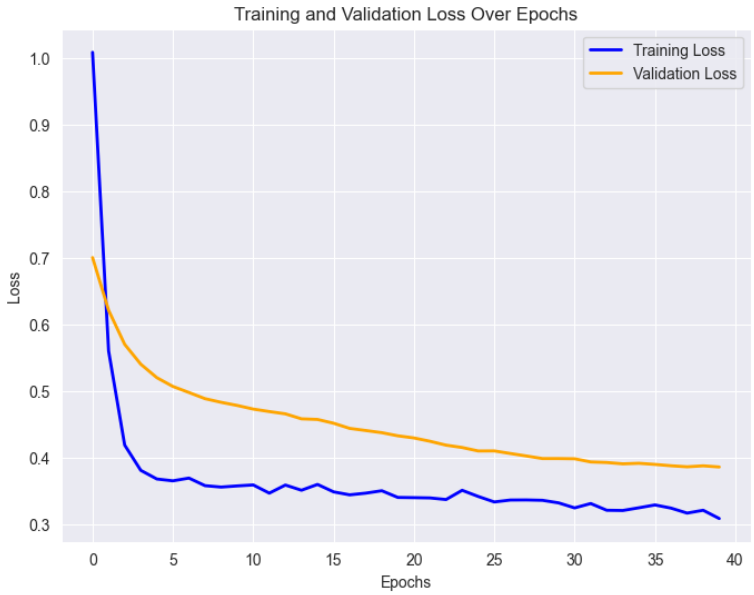

Figure 0-19 Improved Model Scenario Training and Validation Loss

Improved Model Scenario Confusion Matrix shows 62 true negatives and 70 true positives, but also 9 false positives and 9 false negatives.

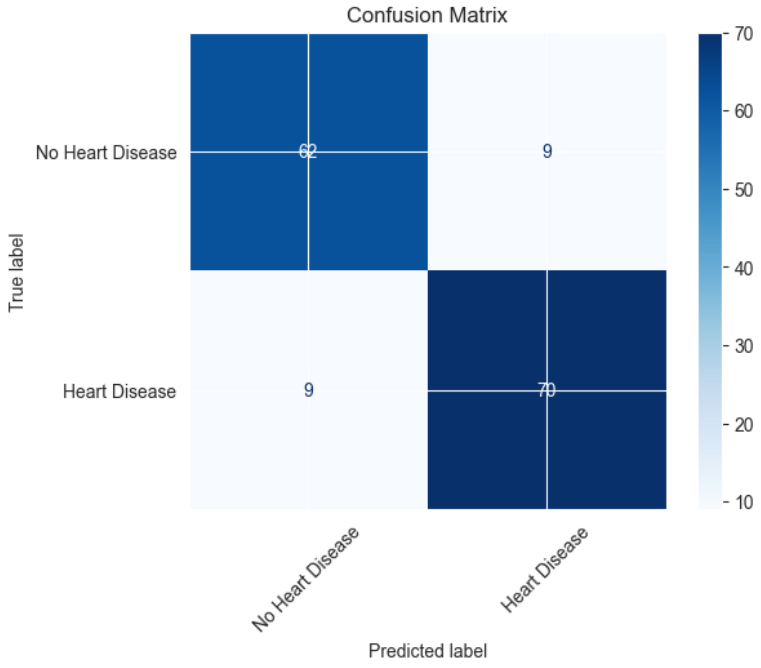

Figure 0-20 Improved Model Scenario Confusion Matrix

New Metrics

The MCC score starts near 0, indicating random classification at the beginning of training. It rises steeply within the first 10 epochs, showing rapid improvement in the model's ability to distinguish between classes. After approximately 15 epochs, the MCC score stabilizes around **0.75–0.8**, indicating a strong correlation between predictions and actual labels.

The relatively flat curve in later epochs suggests that the model has reached a convergence point, with minimal fluctuations in MCC, meaning overfitting is unlikely.

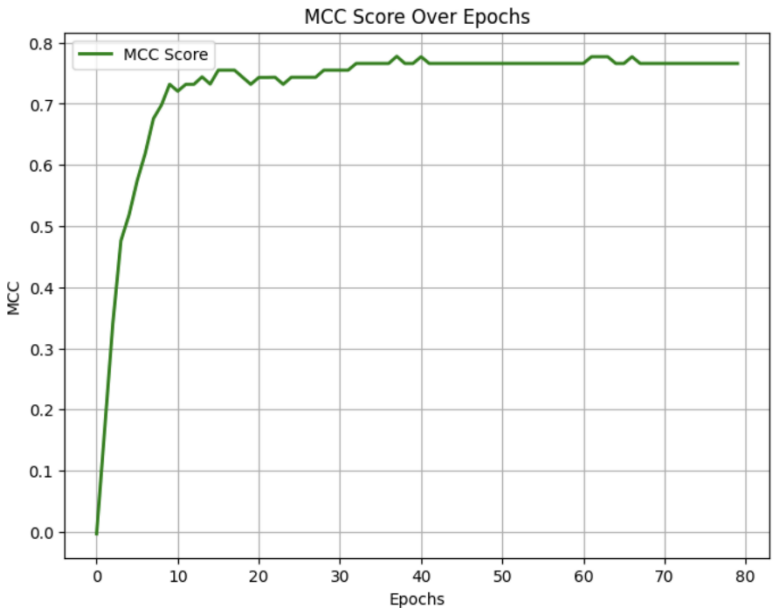

Figure 0-21 MCC Score Over Epochs

Imbalanced Data

Unbalanced Data Training and Validation Loss shows a rapid initial decrease in both losses followed by a gradual plateau, with the validation loss consistently slightly higher than the training loss. Unbalanced Data Confusion Matrix shows that the model has a balanced error rate, with an equal number of false positives and false negatives and shows reasonable overall accuracy.

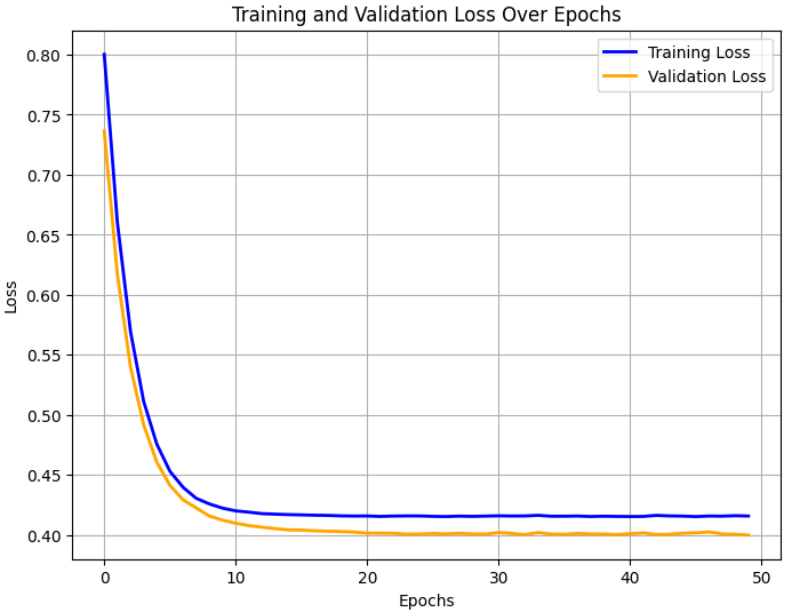

Figure 0-23 Unbalanced Data Training and Validation Loss

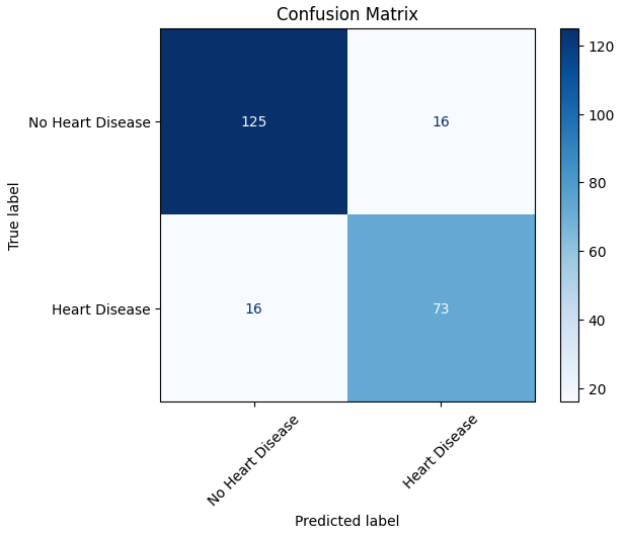

Figure 0-22 Unbalanced Data Confusion Matrix

Error! Reference source not found. depicts a consistent decrease in both losses, with the validation loss slightly lower than the training loss. The confusion matrix is as above.

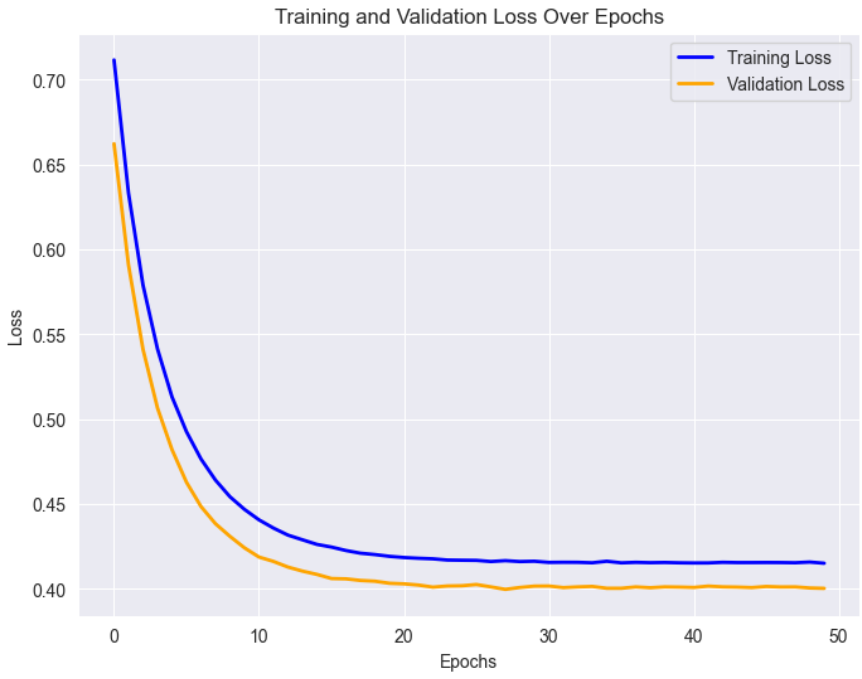

Figure 0-24 Slightly Unbalanced Training and Validation Loss

Error! Reference source not found. illustrates a consistent decrease in both losses, with the validation loss consistently lower than the training loss. This confusion matrix shows a heart disease prediction model with high accuracy in identifying those without heart disease (176 true negatives), but a significant number of false negatives (40), indicating a risk of missing actual heart disease cases.

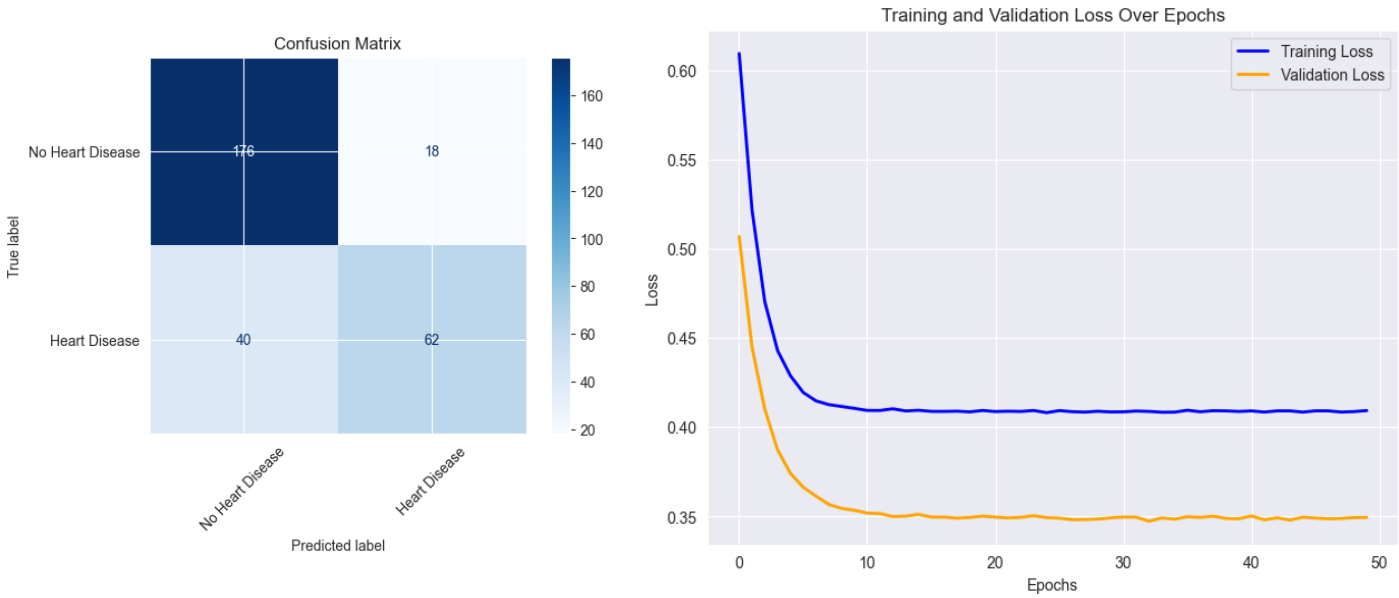

Figure 0-25 Balanced Data Training and Validation Loss and Confution Matrix

PCA

After performing PCA, only five components were retained. The training and validation loss, as well as the confusion matrix, remain consistent with the original results.

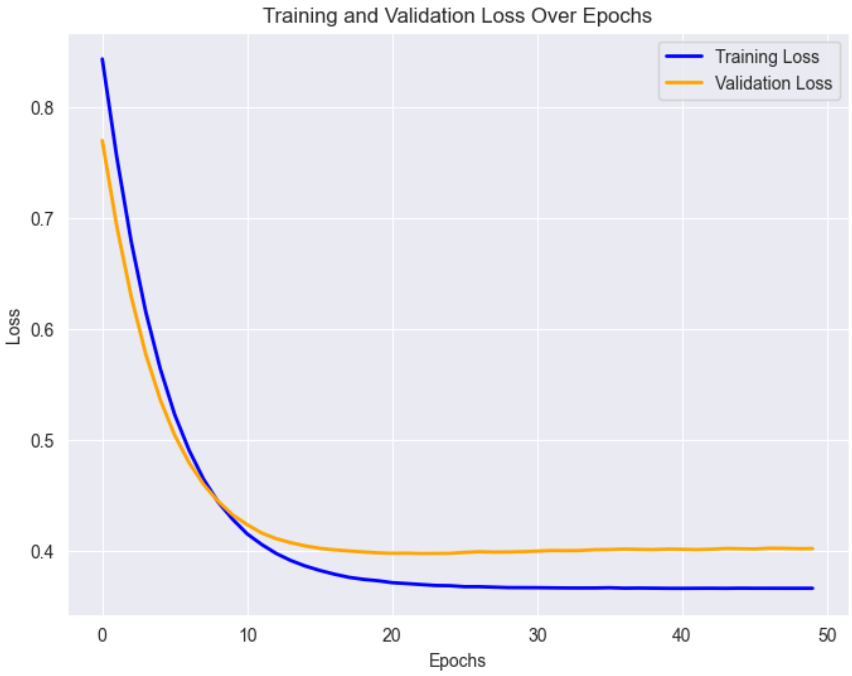

Figure 0-27 PCA Training and Validation Loss

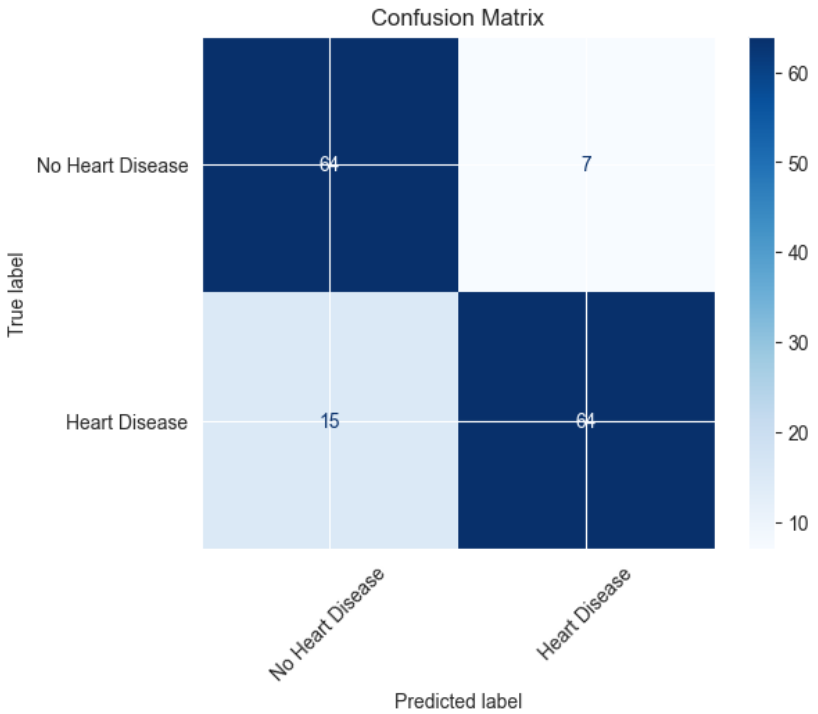

Figure 0-26 PCA Confusion Matrix

Discussion and Conclusions

ML Model Vs. Basic Neural Network

The results indicate that while logistic regression performed well with an accuracy of 85%, the basic neural network with 16 neurons and a default activation function demonstrated similar accuracy but exhibited signs of slight overfitting. The confusion matrices of both the ML model and the neural network turned out to be exactly the same therefore, F1 score, precision and recall are identical.

Hyperparameters Fine-Tune

Hyperparameter tuning revealed that increasing the learning rate resulted in greater instability, causing fluctuations in validation loss. Additionally, a higher number of neurons led to overfitting, as indicated by a noticeable gap between training and validation loss. While different activation functions (ReLU, Sigmoid, Tanh) showed only marginal performance differences, ReLU generally provided better stability throughout training.

Dataset Modifications

The quality of the dataset played a crucial role in model performance. In the improved dataset, where outliers were removed and the target classes were balanced, the network achieved an accuracy of 91%, demonstrating smoother training convergence with reduced fluctuations in weight updates. This dataset allowed the model to identify patterns more effectively, as indicated by a lower number of false negatives in the confusion matrix.

Conversely, in the worst dataset, where key features were removed, accuracy dropped to 77%. Training and validation loss curves showed instability, and the model failed to generalize, likely due to the loss of critical predictive information. The confusion matrix further highlighted that a significant number of heart disease cases were misclassified, which could have severe real-world implications for patient outcomes.

Improved Network Architecture

To enhance the model's architecture, several modifications were introduced, leading to a more robust and generalizable network. Batch Normalization was incorporated to stabilize learning and accelerate convergence, while the Adam Optimizer with Weight Decay (learning rate = 0.001, weight decay = 0.01) helped mitigate overfitting and improve generalization. Additionally, a revised network structure with 12 neurons and a sigmoid activation function provided better overall stability. As a result, the improved model achieved 85.33% accuracy. These enhancements made the network a reliable tool for heart disease prediction, demonstrating the importance of fine-tuning both architectural design and optimization strategies.

MCC Metric

The Matthews Correlation Coefficient (MCC) is a reliable metric that evaluates classification performance by considering true positives, true negatives, false positives,

and false negatives, making it particularly useful for imbalanced datasets.
The MCC score of ~0.75–0.8, confirming a strong correlation between predictions and actual classifications.

Imbalanced Dataset Scenarios

To assess the impact of class distribution on model performance, three levels of balancing were applied: unbalanced data achieved 84% accuracy, slightly balanced data improved accuracy to 86.09%, while fully balanced data resulted in a decrease to 80.74% accuracy. Although balancing helped reduce class bias, over-balancing led to a decline in performance, likely due to information loss from excessive resampling. These findings highlight the importance of maintaining a strategic balance in dataset preparation, rather than aiming for a strict 50-50 class distribution, to ensure optimal model generalization and predictive accuracy.

PCA

To assess the effect of dimensionality reduction on model performance, Principal Component Analysis (PCA) was applied, retaining **five principal components**. The explained variance ratio of these components was **[0.26, 0.12, 0.11, 0.09, 0.08]**, indicating that while the dataset's dimensionality was significantly reduced, a substantial portion of the information was preserved.

Interestingly, despite the lower number of features, the model's training and validation loss remained consistent, and the confusion matrix **showed no significant difference from the original model**. This suggests that **the key predictive features were effectively retained** even after dimensionality reduction.

Future Work

Several potential improvements can be explored to enhance model performance and practical applicability:

1. **Expanding the Dataset:** Increasing sample size or integrating additional clinical datasets may improve generalizability.
2. **Feature Engineering:** Incorporating domain-specific medical insights could refine predictions.
3. **Real-World Implementation:** Deploying the model in a clinical decision-support system and validating it with real patient data would assess its practical utility.
4. **Explainability Techniques:** Applying SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) can help increase model interpretability for healthcare professionals.
5. **Time-Series Analysis:** Considering longitudinal patient data rather than static clinical records could improve early disease detection.
6. **PCA and Other Feature Selection Comparisons:**

While PCA was used for dimensionality reduction, alternative techniques like LASSO Regression, Recursive Feature Elimination (RFE), and Autoencoders could be compared to assessing the best approach for maintaining predictive power while reducing computational complexity.

By addressing these areas, future work can improve heart failure prediction models, making them more reliable, interpretable, and effective in real-world clinical applications.

References

L. Soto, Framingham Heart Study, Framingham, MA: National Heart, Lung, and Blood Institute, 1948.

A. Keys, Seven Countries Study: Diet, Lifestyle, and Cardiovascular Disease, University of Minnesota, 1950s.

Zainab Alwaeli. (2025). *Predicting heart disease using machine learning techniques: A comprehensive* study. In [Smart Infrastructures in the IoT Era], pp. 1095–1103. https://link.springer.com/chapter/10.1007/978-3-031-72509-8_89#citeas