BUILDING NEURAL NETWORKS AND CNN'S

Team Mate 1: Samyak Shah - 50604267

Team Mate 2: Nischal Seemantula - 50605666

1 Dataset Overview

The dataset consists of 766 samples and 8 features. A brief summary of the features is as follows:

- Number of samples: 766
- Number of features: 8
- Feature types:
  - f1, f2, f4, f5, f6, f7: object
  - f3, target: integer
- No missing values in the dataset.

Key Statistics:

```
f1          float64
f2          float64
f3            int64
f4          float64
f5          float64
f6          float64
f7          float64
target        int64
```

2 Data Preprocessing

The following preprocessing steps were applied to clean and prepare the data:

- Replaced invalid characters in f1, f2, f4, f5, f6, f7 with None.
- Converted categorical features to numeric using pd.to_numeric.
- Handled missing values by filling them with the column mean.
- Detected and handled outliers using Interquartile Range (IQR). In particular, outliers in f3 and f4 were handled by replacing them with the mean.

3 Data Visualizations

3.1 Box Plot of f3

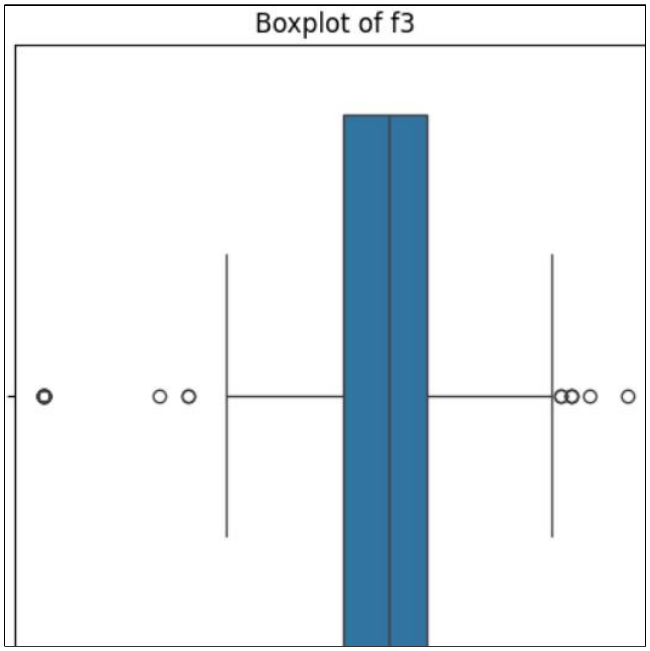


Figure 1: Box plot of feature f3.

This plot shows the distribution and presence of outliers in the f3 feature.

3.2 Correlation Heatmap

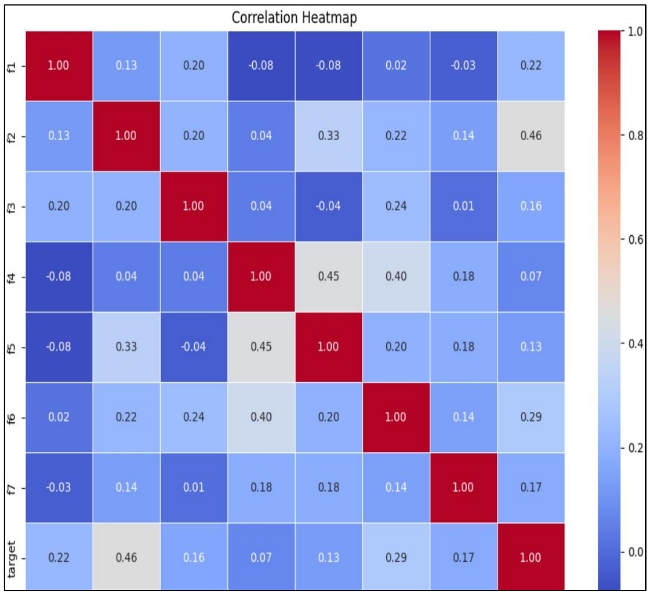


Figure 2: Correlation Heatmap between features.

The heatmap highlights the correlation between different features, with darker colors indicating stronger relationships.

3.3 Histogram of Target Variable

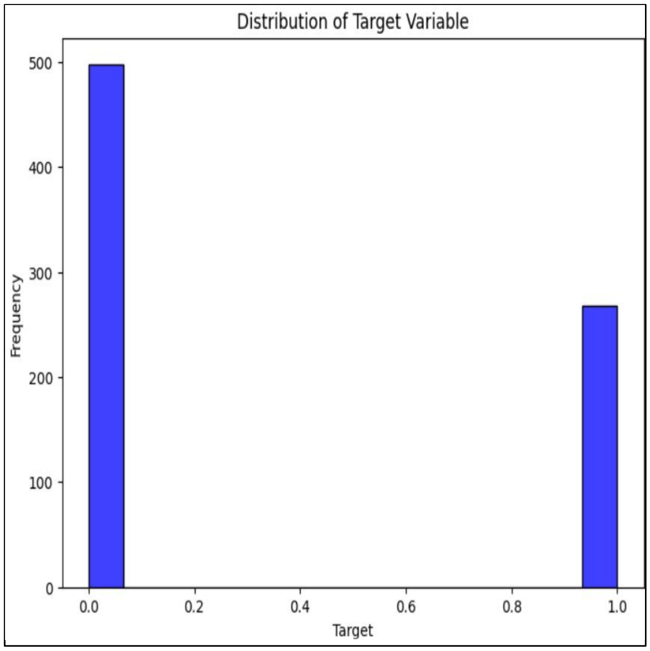


Figure 3: Distribution of the target variable (0/1).

The histogram shows the distribution of the target variable, with values of 0 and 1.

4 Neural Network Architecture

The neural network implemented is a simple feedforward network with the following architecture:

- Input layer: 6 features.
- Two hidden layers, each with 64 neurons, followed by batch normalization and dropout (50%).
- Output layer: Single neuron with sigmoid activation for binary classification.

5 Performance Metrics and Analysis

The model was trained for 200 epochs. The final performance metrics on the test set are as follows:

- Test Accuracy: 79.22%
- Test Loss: 0.4564
- Precision: 0.6786
- Recall: 0.7308
- F1 Score: 0.7037

5.1 Training, Validation, and Test Loss Over Epochs

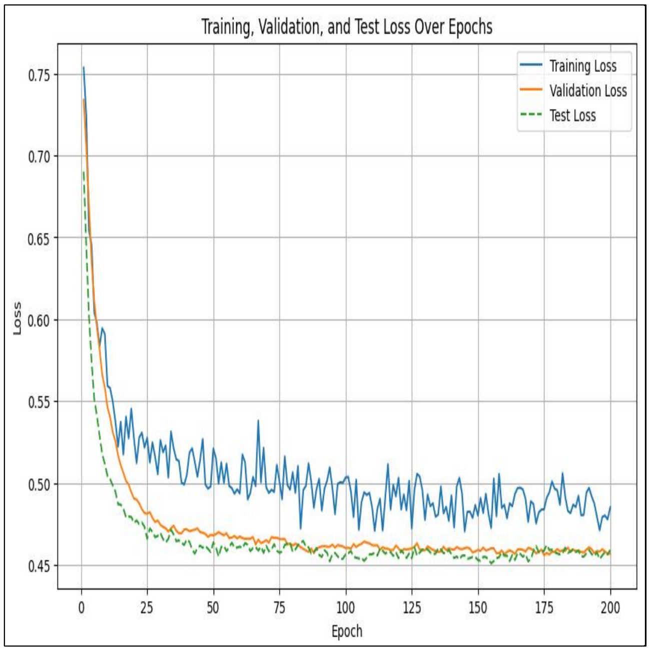


Figure 4: Training, Validation, and Test Loss Over Epochs.

5.2 Training, Validation, and Test Accuracy Over Epochs

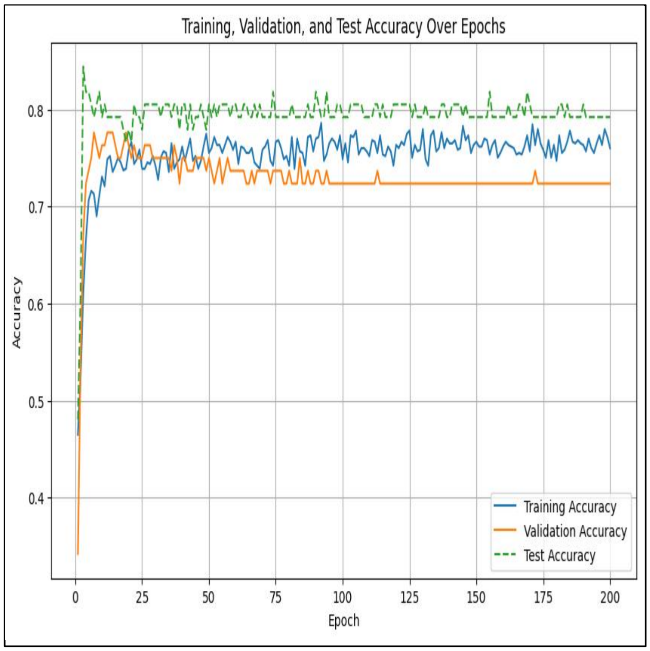


Figure 5: Training, Validation, and Test Accuracy Over Epochs.

5.3 Confusion Matrix

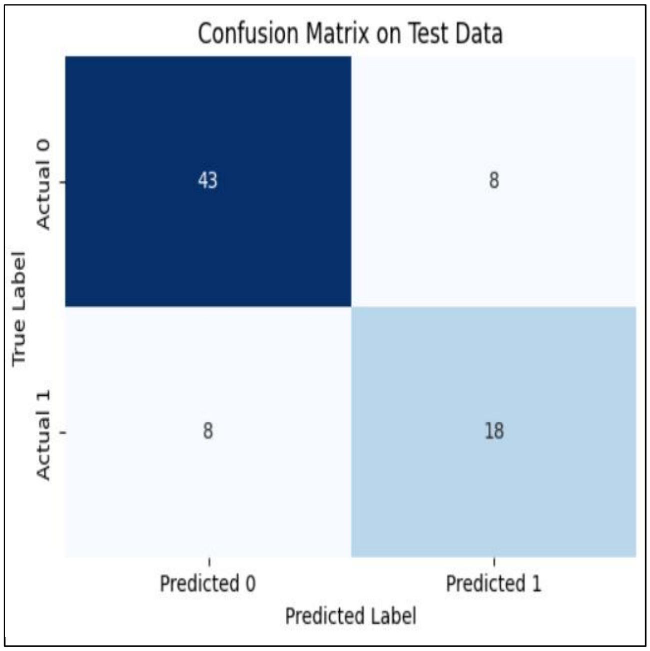


Figure 6: Confusion Matrix for Test Set.

5.4 ROC Curve

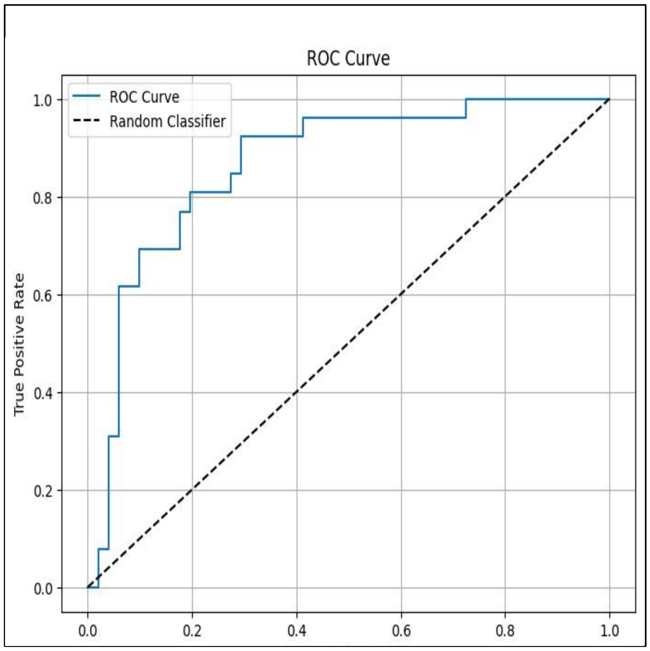


Figure 7: ROC Curve for the model.

The model achieved a balanced performance across precision, recall, and accuracy, making it suitable for binary classification.

PART 2 REPORT

Table 1: Learning Rate Tuning

LEARNING RATE	TEST ACCURACY
0.0001	80.52
0.001	81.82
0.01	81.82

Table 2: Batch Size Tuning

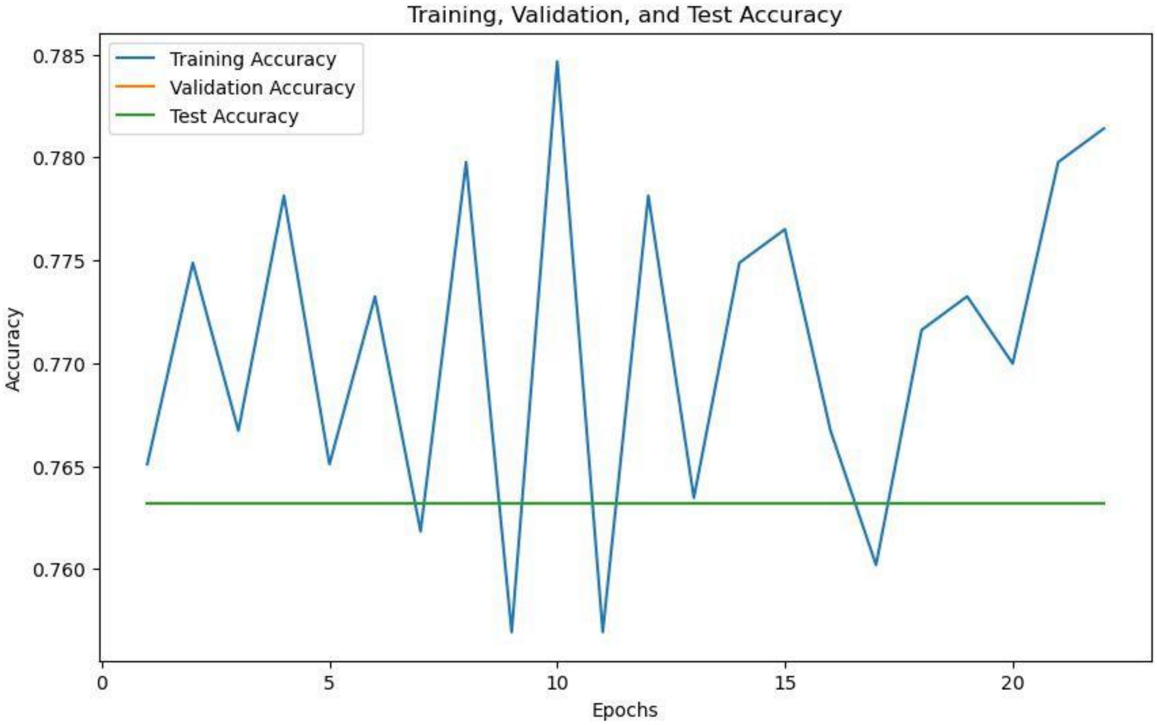
LEARNING RATE	TEST ACCURACY
16	79.22
32	81.82
64	83.12

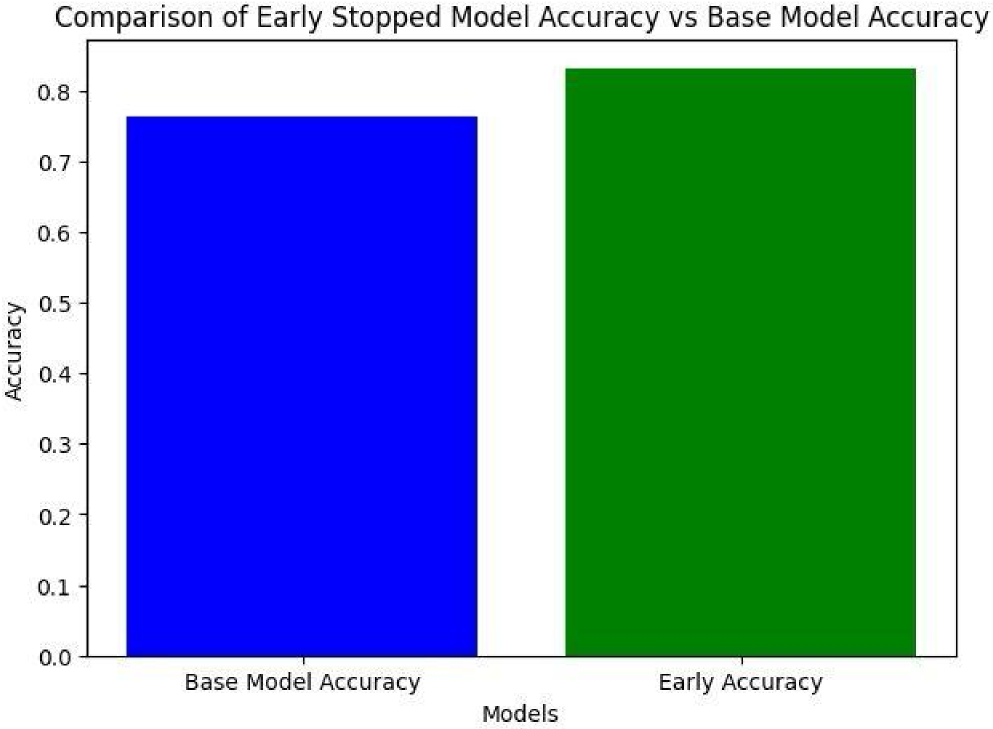
Table 3: Hidden Layer Configuration Tuning

LEARNING RATE	TEST ACCURACY
[64]	85.71
[64, 64]	81.82
[128,64]	83.12

Analysis:

Learning Rate: A rate of 0.001 performed best, indicating a balanced update step.
Batch Size: A size of 64 provided the best accuracy, possibly due to more stable gradient estimates.
Hidden Layers: Configurations with a single layer of size or two layers [64] performed similarly.


Training, Validation, and Test Accuracy
Training Accuracy
Validation Accuracy
Test Accuracy
0.785
0.780
0.775
0.770
0.765
0.760
Accuracy
0
5
10
15
20
Epochs


Comparison of Early Stopped Model Accuracy vs Base Model Accuracy
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Accuracy
Base Model Accuracy
Early Accuracy
Models

Methods for Improvement:

In this I used four different models

1. Early Stopping: Prevents overfitting by halting training when validation loss ceases to improve.

2. Learning Rate Scheduler: Adjusts the learning rate to enhance convergence.

3. Batch Normalization: Stabilizes learning by normalizing inputs to each layer.

4. Gradient Accumulation: Simulates larger batch sizes without increasing memory usage.

Best Model Description

Best Model Configuration:

- Learning Rate: 0.001
- Batch Size: 64
- Hidden Layers: [128, 64]
- Dropout Rate: Adjusted based on best performance

Performance and Analysis:

- The model with a learning rate of 0.001 achieved the highest accuracy of 0.8312.
- Visualizations can include accuracy over epochs to demonstrate improvements with the chosen methods.

CITATIONS:

Pandas: https://pandas.pydata.org/

Sckit-learn: https://scikit-learn.org/stable/

Numpy: https://numpy.org/

Code of Assignment 1

Team Member	Assignment Part	Contribution
Samyak Shah	Part 1, Part 2	50%
Nischal Seemantula	Part 1, Part 2	50%