A Capstone Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

SCHOOL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

by

2203A52163 **MOHAMMAD ABDUL ADNAN**

Under the guidance of

Dr. Ramesh Dadi

Assistant Professor, School of CS&AI.

SR University, Ananthsagar,Warangal,Telagnana-506371

CONTENTS

DATASET

Project-1: F1 2022

The F1 2022 dataset contains detailed data from the 2022 Formula 1 World Championship season, including information on races, drivers, teams, circuits, lap times, pit stops, qualifying results, and final standings. This dataset captures a variety of metrics that are essential for analyzing race performance, driver consistency, and team strategies. It can be used to study trends in driver and constructor performance across the season, identify key factors influencing race outcomes, and predict future race results based on historical performance. With this information, models can be developed for performance forecasting, race simulation, and fan engagement through data-driven insights. This dataset is valuable for motorsport analytics, predictive modeling, and understanding competitive dynamics in Formula 1.

Project-2: Captcha Images

The Captcha Images dataset includes a collection of CAPTCHA images that display randomly generated sequences of alphanumeric characters. These images are used to train models to recognize and interpret distorted text, often designed to prevent automated bots from accessing web services. The dataset poses a challenge due to variations in font style, image noise, distortions, and background clutter. This project focuses on applying Convolutional Neural Networks (CNNs) and sequence prediction techniques to accurately decode the text in CAPTCHA images. The dataset is ideal for developing models in OCR (Optical Character Recognition), enhancing automated form processing, and improving user verification systems. It serves as a practical tool for teaching robust pattern recognition under noise and distortion.

Project-3: Sentiment Analysis – Movie Reviews

The Movie Reviews dataset contains viewer opinions labeled as positive or negative. This project aims to classify the sentiment of each review using natural language processing techniques. By training models to recognize emotional tone and word patterns, we can predict whether a review reflects a favorable or unfavorable opinion. This is useful for understanding audience feedback and improving recommendation systems.
The dataset provides a foundation for building robust sentiment analysis tools used in media analytics, brand monitoring, and user experience evaluation. It also helps in exploring the relationship between language use.

METHODOLOGY

Project 1: F1 2022

Data Collection and Preprocessing: The F1 2022 dataset was loaded into a DataFrame, containing various numerical statistics such as **Points, Pole Positions, Fastest Laps, Wins, Podiums**, and **DNFs** for different drivers. Initial preprocessing involved examining the skewness of these numerical columns, and a **log transformation** (log1p) was applied to reduce high skewness and bring the distribution closer to normal. Visualizations like **box plots** and **pair plots** were used to understand feature distributions and relationships between variables.

Feature Engineering and Outlier Removal: Numerical columns were analyzed using box plots to detect outliers, and the **IQR method** was applied to quantify them. Although outliers were identified, the transformation and standardization steps helped in minimizing their impact. Important features like **Pole Positions, Fastest Laps, Wins, and Podiums** were selected to predict the target variable **Points**.

Exploratory Data Analysis (EDA): Pair plots provided insights into correlations among features. Skewness before and after transformation was printed to verify distribution improvements. These steps helped identify which features most significantly impacted total points scored by a driver across the F1 2022 season.

Model Training: Three regression models—**Linear Regression, Decision Tree Regressor**, and **Random Forest Regressor**—were trained on the preprocessed and scaled data. The dataset was split into training and testing sets using an 80-20 split, and **StandardScaler** was used to normalize the feature values.

Performance Measurement: Each model was evaluated using **MAE (Mean Absolute Error), MSE (Mean Squared Error),** and **R^2 Score (Coefficient of Determination)**. These metrics helped compare the models' ability to predict the final points for drivers based on key performance features.
This structured pipeline enabled effective analysis of driver performance metrics from the 2022 Formula 1 season and offered a comparative view of multiple regression models in predicting race outcomes.

Project 2: Captcha Images

Data Collection and Preprocessing: The CAPTCHA dataset was composed of grayscale images, each labeled with a character extracted from the filename. A DataFrame was created to store image filenames and their corresponding labels. The dataset was split into **training (75%)**, **validation (12.5%)**, and **test (12.5%)** sets. Image augmentation techniques such as rotation, shifting, shearing, and zooming were applied using ImageDataGenerator to artificially expand the dataset and increase robustness.

Model Architecture and Complication: A **Convolutional Neural Network (CNN)** was designed using multiple convolutional layers with ReLU activation, max-pooling, dropout for regularization, and a final softmax output layer. The input images were standardized and resized to **50x200** pixels, and the network was compiled using the **Adam optimizer** and **categorical crossentropy loss**, suitable for multi-class classification.

Training and Evaluation: The model was trained over **10 epochs** using the augmented training set and validated on the validation set. Performance metrics including **loss** and **accuracy** were plotted for both training and validation phases. The final model was saved for later evaluation and reuse.

Performance Measurement: The trained model was tested on the unseen test set, and performance was evaluated using **classification report**, **confusion matrix**, and **ROC curves** for each class. A **Z-test** was performed to compare the model's accuracy against a random baseline (0.5), and **paired t-tests** were used to assess consistency between training and validation accuracy. Additionally, **ANOVA** was applied to compare accuracy distributions across simulated multiple training runs.

Final Evaluation on Subset: The saved model was reloaded and evaluated on a subset of 500 test images to simulate real-world prediction accuracy. The final accuracy was statistically analyzed using another **Z-test** to verify significance.

This project successfully demonstrated the effectiveness of CNNs in solving CAPTCHA image classification tasks, incorporating extensive preprocessing, augmentation, statistical testing, and model evaluation.

Project 3: Sentiment Analysis – Movie Reviews

Dataset Preparation: The dataset consisted of movie reviews and their corresponding numeric ratings. After loading the dataset, the Comments column was treated as the text input, and the Ratings column was used for labeling. Any missing or invalid entries were removed. Ratings were then binned into **5 sentiment categories** using pd.cut() to enable multi-class classification. The data was split into **training (80%)** and **testing (20%)** subsets.

Feature Extraction: The textual reviews were converted to lowercase and tokenized using **Keras Tokenizer** with a vocabulary size of **10,000** words. The resulting integer sequences were padded to a fixed length of **200 tokens** to ensure uniformity. The target labels were one-hot encoded using to_categorical() to fit the multi-class classification task.

Model Architecture: The model was built using a **Sequential LSTM architecture**, comprising:

- An **Embedding layer** that mapped each word token to a dense vector representation.
- An **LSTM layer** with dropout and recurrent dropout for better generalization and to capture temporal dependencies in the review sequences.
- A final **Dense layer** with softmax activation to output a probability distribution across the **five sentiment classes**.

Model Training: The model was compiled using the **Adam optimizer** and **categorical cross-entropy** loss. Training was performed for **5 epochs** with a **batch size of 32**. A **validation set** (20% of the training data) was used to monitor performance on unseen data during training.

Performance Evaluation: After training, the model was evaluated on the test set using various metrics:

- **Accuracy, Precision, Recall, and F1-score** were calculated using the true and predicted sentiment classes.
- A **confusion matrix** was generated to examine how well the model distinguished between different sentiment categories.
- A **multi-class ROC curve** was plotted to visualize the trade-off between true positive rate and false positive rate for each class.
- **AUC (Area Under Curve)** scores were calculated for each class to assess the classification quality.

Visualizations:
Key visualizations included:

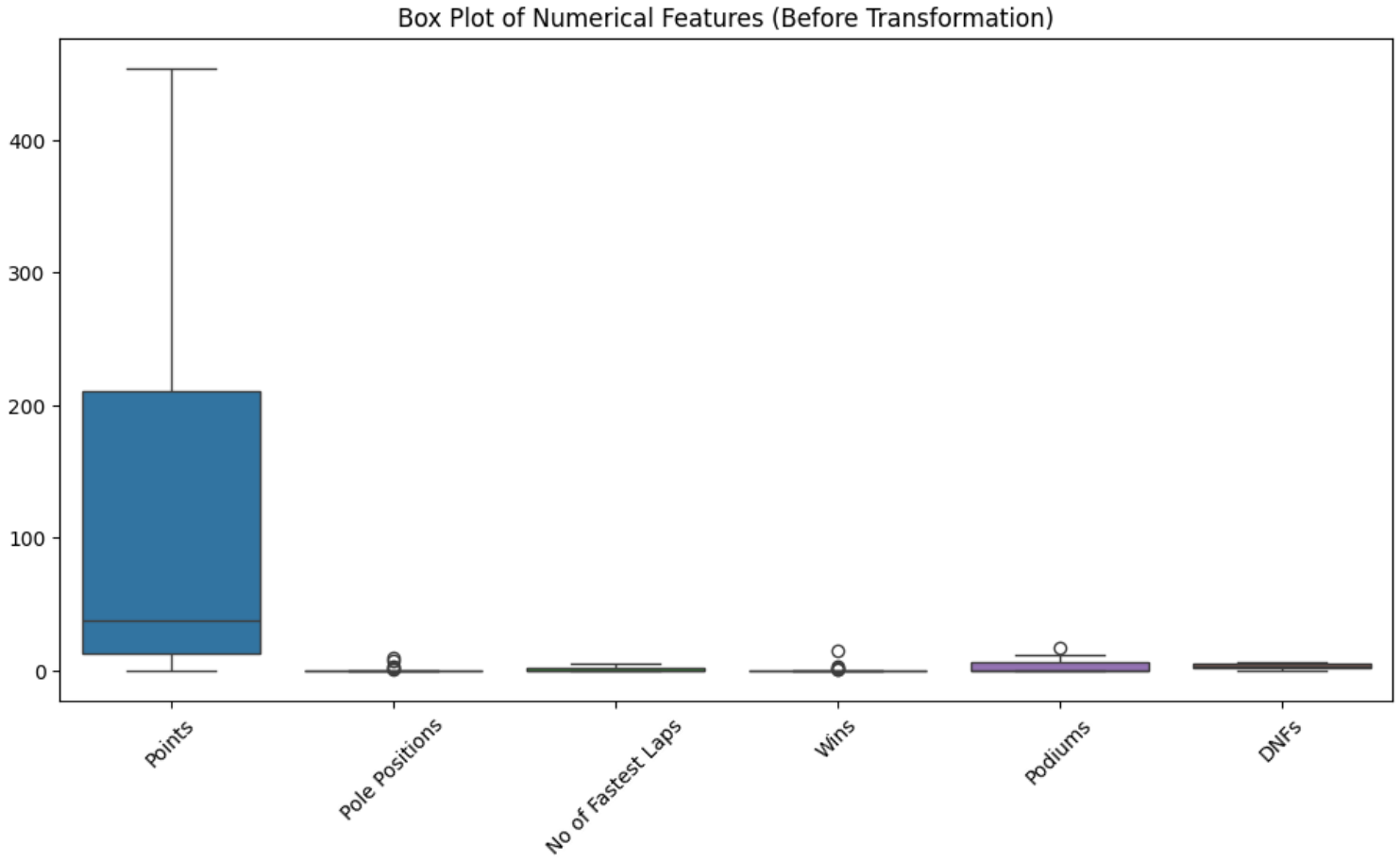
- **Accuracy and Loss Plots:** Displayed model training and validation accuracy and loss over the epochs.
- **Confusion Matrix:** Showed the number of correct and incorrect predictions for each class.
- **Classification Report:** Included precision, recall, and F1-score for each sentiment class.

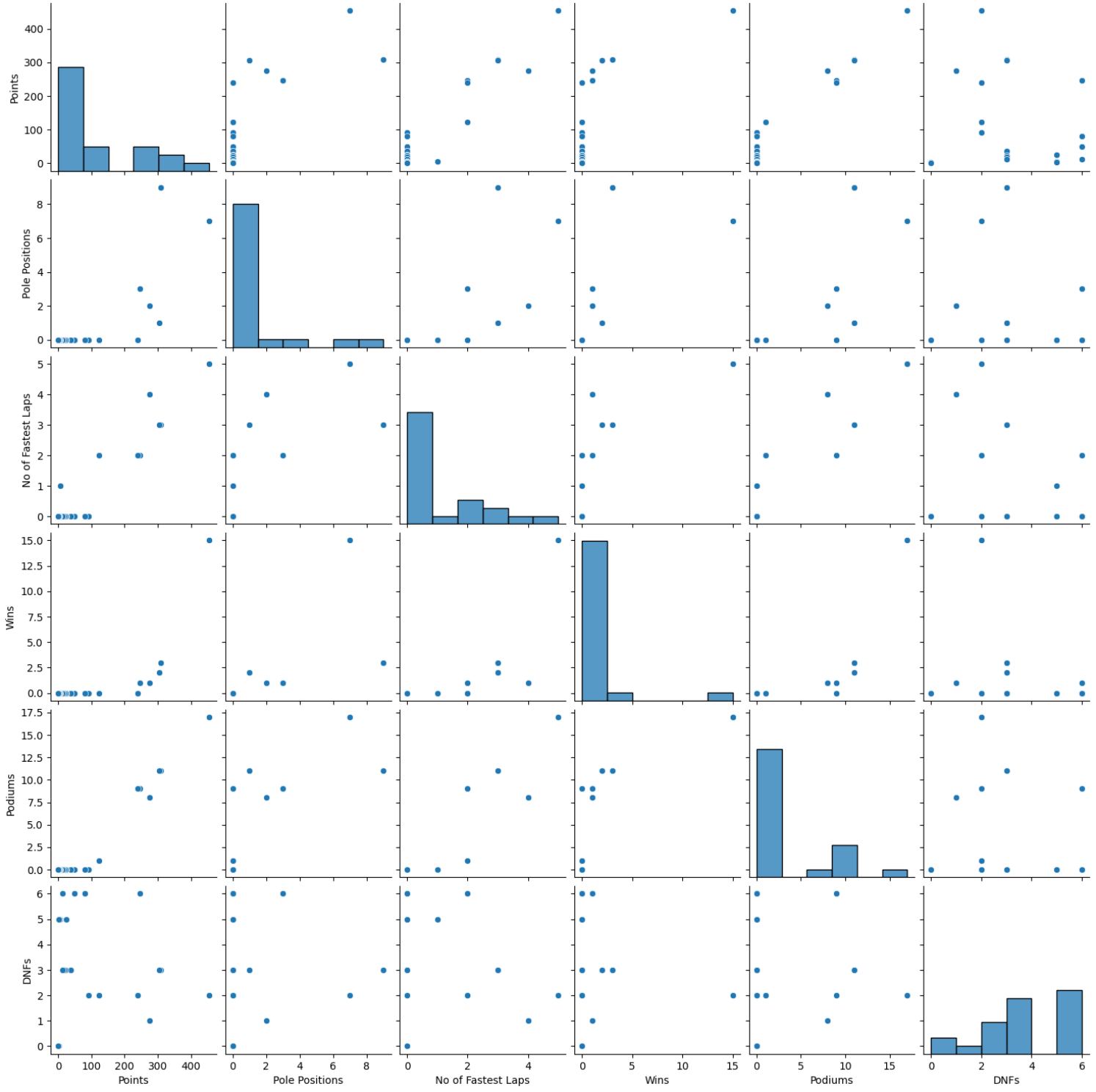
RESULTS

PROJECT-1

Skewness before transformation:
Points 1.181533
Pole Positions 2.506133
No of Fastest Laps 1.273584
Wins 3.986967
Podiums 1.384612
DNFs -0.064696
dtype: float64

Skewness after transformation:
Points -0.308723
Pole Positions 1.847338
No of Fastest Laps 0.835205
Wins 2.535908
Podiums 1.013628
DNFs -1.153847
dtype: float64





Outliers detected:
Points 0
Pole Positions 5
No of Fastest Laps 0
Wins 5
Podiums 1
DNFs 0
dtype: int64

Linear Regression Performance:
MAE: 0.5989082141487044
MSE: 0.8032343780095814
R2 Score: 0.5815541038212204

Decision Tree Performance:
MAE: 0.5708231300106658
MSE: 0.7780324412099794
R2 Score: 0.5946830825082166

Random Forest Performance:
MAE: 0.600299260850443
MSE: 0.7456601693302903
R2 Score: 0.6115474557084857

PROJECT-2

Found 290 validated image filenames belonging to 9 classes.
Found 54 validated image filenames belonging to 9 classes.
Found 54 validated image filenames belonging to 9 classes.

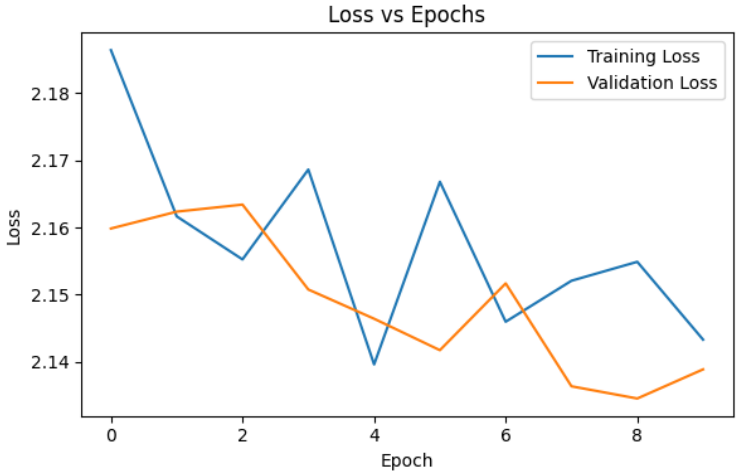
Model: "sequential_5"

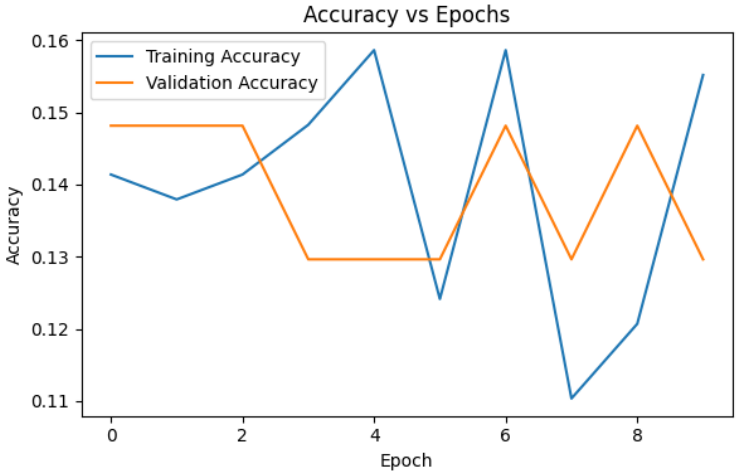
Layer (type)	Output Shape	Param #
conv2d_85 (Conv2D)	(None, 50, 200, 32)	320
conv2d_86 (Conv2D)	(None, 50, 200, 32)	9,248
max_pooling2d_40 (MaxPooling2D)	(None, 25, 100, 32)	0
conv2d_87 (Conv2D)	(None, 25, 100, 64)	18,496
conv2d_88 (Conv2D)	(None, 25, 100, 64)	36,928
max_pooling2d_41 (MaxPooling2D)	(None, 12, 50, 64)	0
conv2d_89 (Conv2D)	(None, 12, 50, 128)	73,856
max_pooling2d_42 (MaxPooling2D)	(None, 6, 25, 128)	0
flatten_13 (Flatten)	(None, 19200)	0
dense_34 (Dense)	(None, 256)	4,915,456
dropout_21 (Dropout)	(None, 256)	0
dense_35 (Dense)	(None, 9)	2,313

Total params: 5,056,617 (19.29 MB)
Trainable params: 5,056,617 (19.29 MB)
Non-trainable params: 0 (0.00 B)

Epoch 1/10
10/10 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ **8s** 401ms/step - accuracy: 0.1222 - loss: 2.1865 - val_accuracy: 0.1481 - val_loss: 2.1598
Epoch 2/10
10/10 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ **1s** 143ms/step - accuracy: 0.1467 - loss: 2.1523 - val_accuracy: 0.1481 - val_loss: 2.1623
Epoch 3/10
10/10 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ **1s** 137ms/step - accuracy: 0.1404 - loss: 2.1515 - val_accuracy: 0.1481 - val_loss: 2.1634
Epoch 4/10
10/10 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ **1s** 143ms/step - accuracy: 0.1388 - loss: 2.1719 val_accuracy: 0.1296 - val_loss: 2.1508
Epoch 5/10
10/10 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ **1s** 139ms/step - accuracy: 0.1939 - loss: 2.1317 - val_accuracy: 0.1296 - val_loss: 2.1464

Epoch 6/10
10/10 ━━━━━━━━━━━━━━━━━━━━ **3s** 170ms/step - accuracy: 0.1492 - loss: 2.1398 - val_accuracy: 0.1296 - val_loss: 2.1417
Epoch 7/10
10/10 ━━━━━━━━━━━━━━━━━━━━ **2s** 152ms/step - accuracy: 0.1676 - loss: 2.1436 - val_accuracy: 0.1481 - val_loss: 2.1517
Epoch 8/10
10/10 ━━━━━━━━━━━━━━━━━━━━ **1s** 137ms/step - accuracy: 0.1217 - loss: 2.1504 - val_accuracy: 0.1296 - val_loss: 2.1364
Epoch 9/10
10/10 ━━━━━━━━━━━━━━━━━━━━ **2s** 158ms/step - accuracy: 0.1041 - loss: 2.1703 - val_accuracy: 0.1481 - val_loss: 2.1346
Epoch 10/10
10/10 ━━━━━━━━━━━━━━━━━━━━ **1s** 139ms/step - accuracy: 0.1493 - loss: 2.1274 - val_accuracy: 0.1296 - val_loss: 2.1389

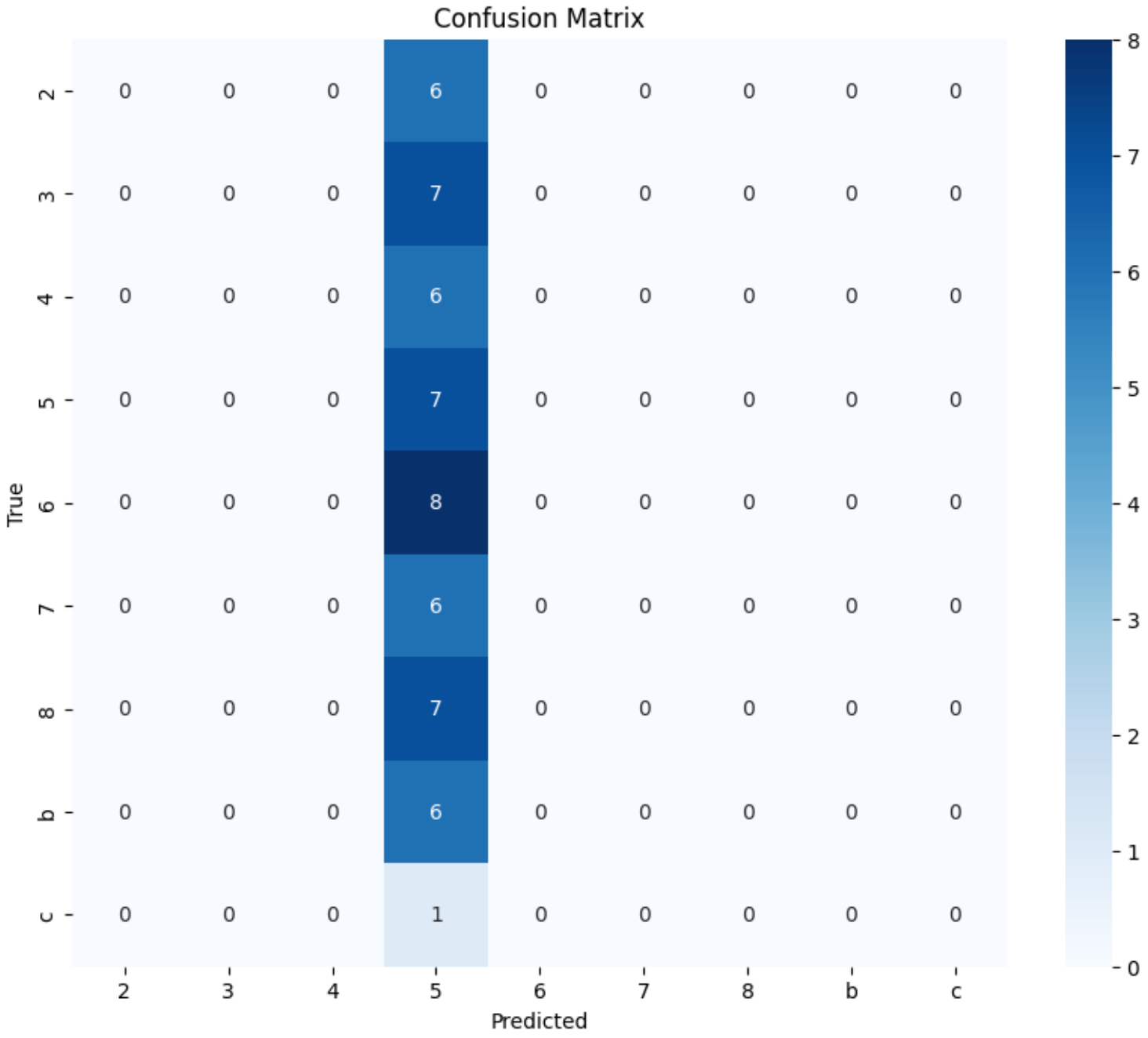


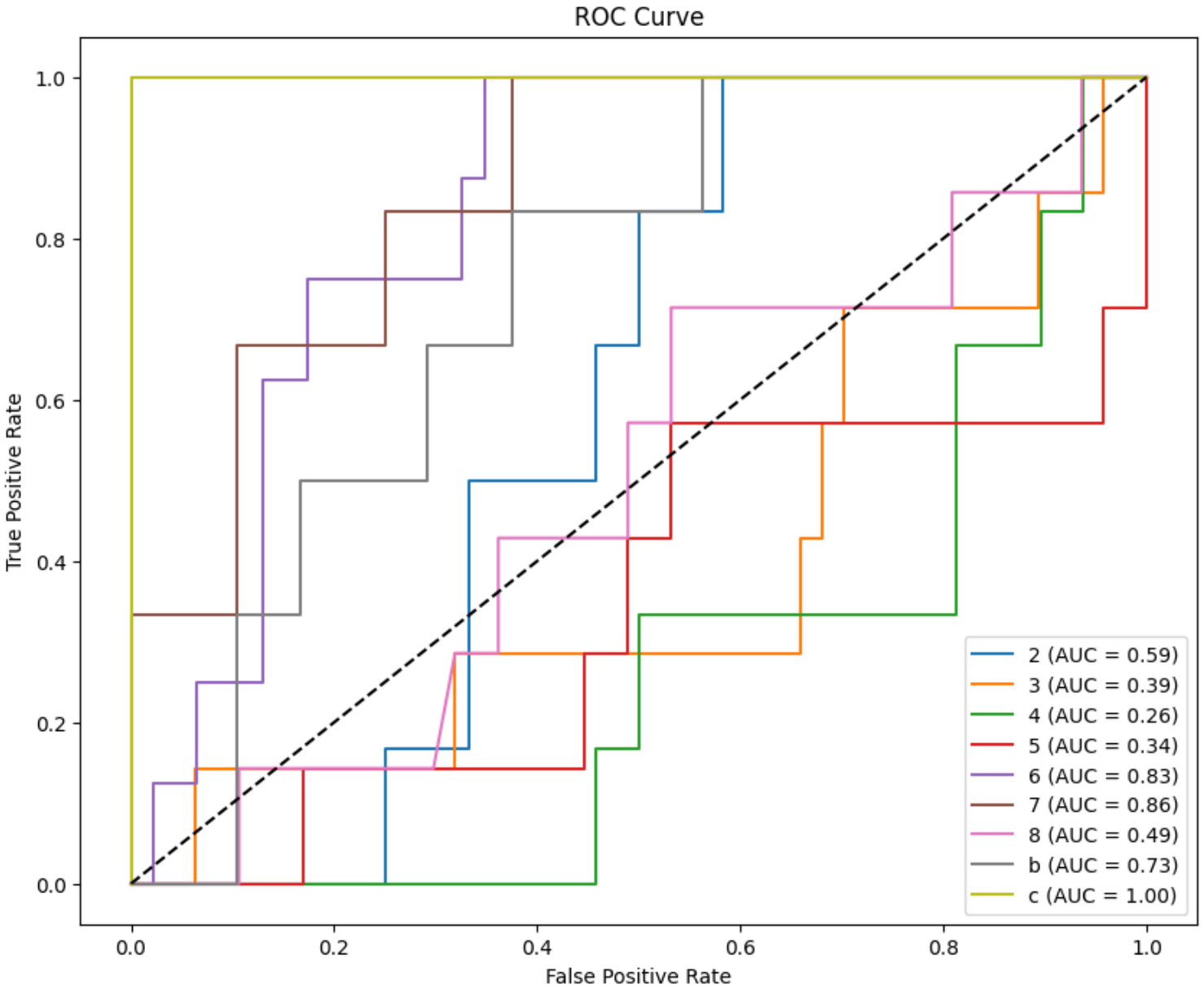


2/2 ━━━━━━━━━━━━━━━━━━━━ **0s** 67ms/step - accuracy: 0.1177 - loss: 2.1505
Test Accuracy: 0.1296, Test Loss: 2.1360
2/2 ━━━━━━━━━━━━━━━━━━━━ **1s** 280ms/step

Classification Report:

	precision	recall	f1-score	support
2	0.00	0.00	0.00	6
3	0.00	0.00	0.00	7
4	0.00	0.00	0.00	6
5	0.13	1.00	0.23	7
6	0.00	0.00	0.00	8
7	0.00	0.00	0.00	6
8	0.00	0.00	0.00	7
b	0.00	0.00	0.00	6
c	0.00	0.00	0.00	1
accuracy			0.13	54
macro avg	0.01	0.11	0.03	54
weighted avg	0.02	0.13	0.03	54


Confusion Matrix
True
Predicted
2 3 4 5 6 7 8 b c
0 0 0 6 0 0 0 0 0
0 0 0 7 0 0 0 0 0
0 0 0 6 0 0 0 0 0
0 0 0 7 0 0 0 0 0
0 0 0 8 0 0 0 0 0
0 0 0 6 0 0 0 0 0
0 0 0 7 0 0 0 0 0
0 0 0 6 0 0 0 0 0
0 0 0 1 0 0 0 0 0
8 7 6 5 4 3 2 1 0



Z-test: Z-stat = -5.9259, p-value = 0.0000
T-test: T-stat = 0.1266, p-value = 0.9020
ANOVA: F-stat = 1.3282, p-value = 0.2817
Final Test on 500 Images: Accuracy = 0.1296, Loss = 2.1360
Z-test on subset: Z-stat = -16.5635, p-value = 0.0000

PROJECT-3

Columns: Index(['Comments', 'Ratings'], dtype='object')

	Comments	Ratings
0	I didnt go in with big hopes, but i was expect...	8
1	A unique genre, a well written story (script) ...	8
2	Majestic at scale, grandeur in VFX, and great ...	9
3	"Kalki 2898" is not just a movie; it's an expe...	10
4	Best Indian movie Nagi combined Hindu mytholog...	10

Shape of X_train_pad: (4000, 200)

Shape of X_test_pad: (1000, 200)

Model: "sequential_14"

Layer (type)	Output Shape	Param #
embedding_14 (Embedding)	(None, 200, 128)	1,280,000
lstm_15 (LSTM)	(None, 64)	49,408
dense_18 (Dense)	(None, 5)	325

Total params: 1,329,733 (5.07 MB)

Trainable params: 1,329,733 (5.07 MB)

Non-trainable params: 0 (0.00 B)

Epoch 1/5

125/125 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ **16s** 93ms/step - accuracy: 0.5615 - loss: 1.1708 - val_accuracy: 0.7110 - val_loss: 0.6737

Epoch 2/5

125/125 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ **12s** 95ms/step - accuracy: 0.6749 - loss: 0.6965 - val_accuracy: 0.8370 - val_loss: 0.4554

Epoch 3/5

125/125 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ **11s** 91ms/step - accuracy: 0.8153 - loss: 0.5128 - val_accuracy: 0.8370 - val_loss: 0.4040

Epoch 4/5

125/125 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ **11s** 89ms/step - accuracy: 0.8490 - loss: 0.3838 - val_accuracy: 0.8370 - val_loss: 0.3954

Epoch 5/5

125/125 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ **12s** 93ms/step - accuracy: 0.8348 - loss: 0.4215 - val_accuracy: 0.8370 - val_loss: 0.3944

32/32 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ **2s** 42ms/step

Classification Report:

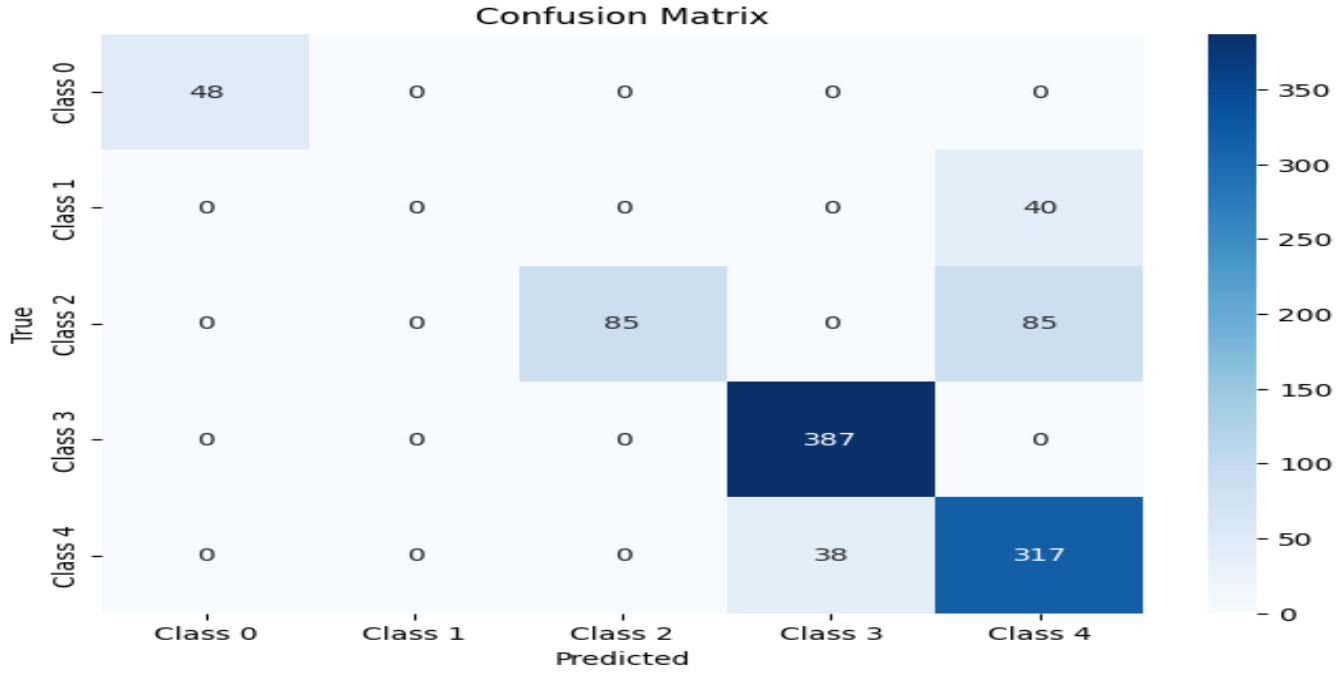
	precision	recall	f1-score	support
Class 0	1.00	1.00	1.00	48
Class 1	0.00	0.00	0.00	40
Class 2	1.00	0.50	0.67	170
Class 3	0.91	1.00	0.95	387
Class 4	0.72	0.89	0.80	355
accuracy			0.84	1000
macro avg	0.73	0.68	0.68	1000
weighted avg	0.83	0.84	0.81	1000

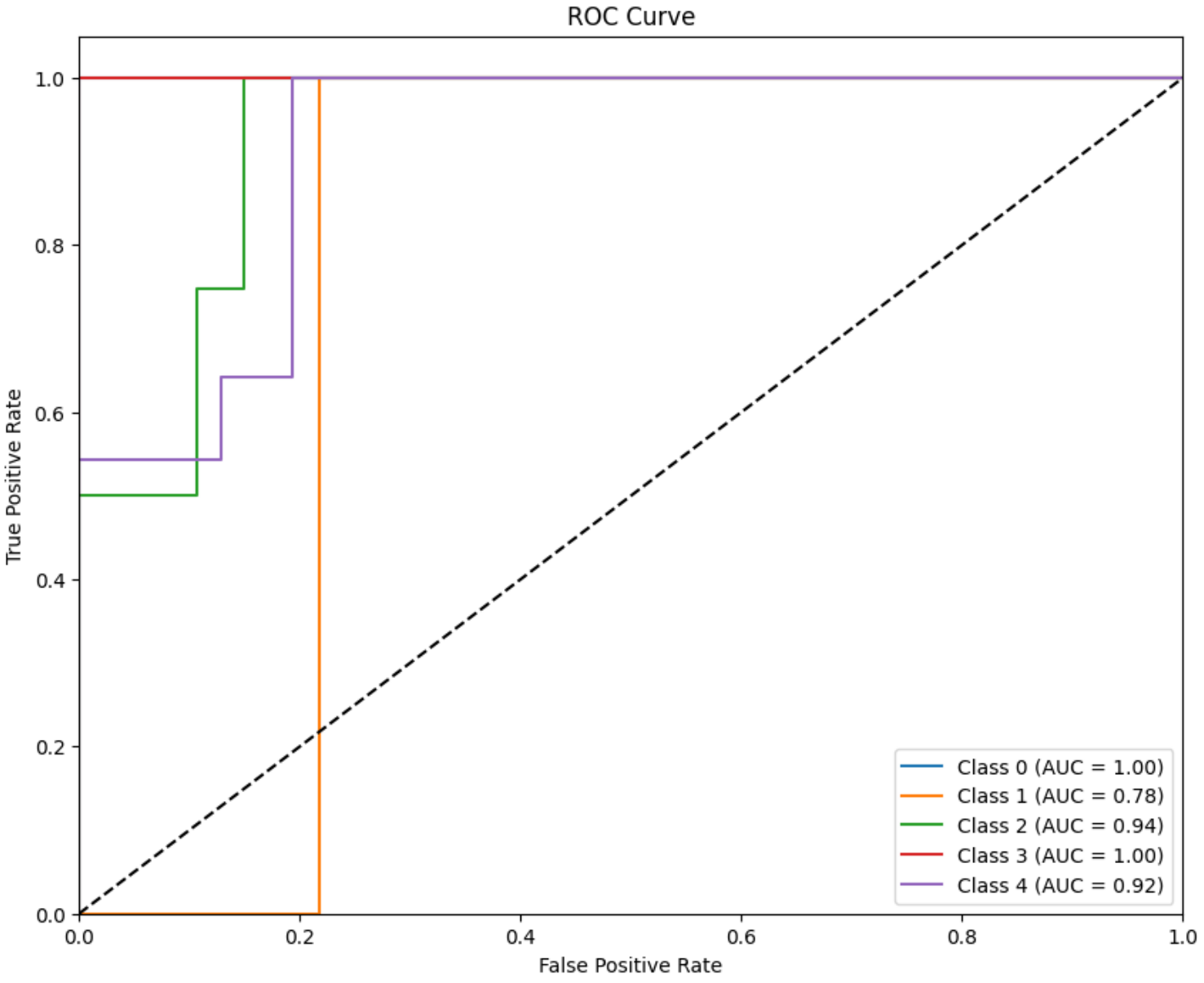
Accuracy: 0.8370

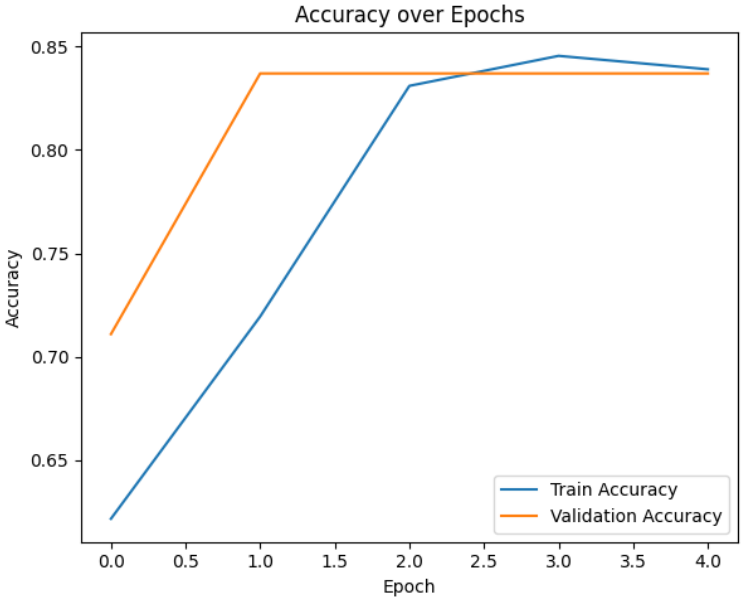
Precision: 0.8250

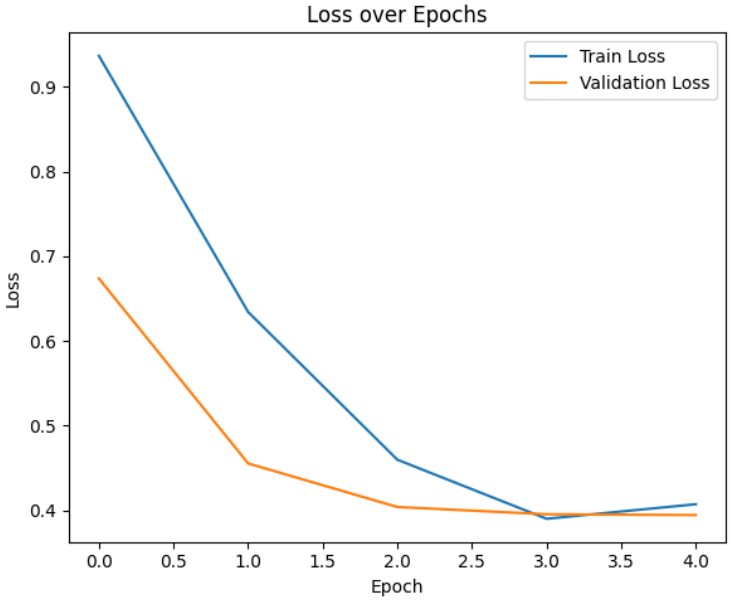
Recall: 0.8370

F1 Score: 0.8126




ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 1.00)
Class 1 (AUC = 0.78)
Class 2 (AUC = 0.94)
Class 3 (AUC = 1.00)
Class 4 (AUC = 0.92)


Accuracy over Epochs
Accuracy
Epoch
0.65
0.70
0.75
0.80
0.85
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Train Accuracy
Validation Accuracy


Loss over Epochs
Loss
Epoch
0.4
0.5
0.6
0.7
0.8
0.9
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Train Loss
Validation Loss