FER: Face Expression Recognition

Computer Vision Project Report

Shrey Srivastava
Roll Number: 22303
BS Data Science and Engineering
IISER Bhopal

November 22, 2024

Contents

1 Problem Statement

Facial expression recognition (FER) aims to classify emotions based on facial images. Emotions such as happiness, sadness, anger, or surprise play a crucial role in human interactions and communication. This project develops a system capable of recognizing emotions accurately under diverse real-world conditions.

Why Is This Project Important?

- **Mental Health Applications:** FER systems can aid in diagnosing and monitoring mental health conditions by analyzing emotional patterns over time.
- **Human-Computer Interaction (HCI):** Enhancing HCI systems, like customer service bots or personal assistants, to respond empathetically based on detected emotions.
- **Behavior Analysis:** Applied in surveillance, education, and workplace productivity to monitor engagement and emotional well-being.
- **Entertainment and Gaming:** Real-time emotion detection enhances immersive gaming experiences and personalized content delivery.

2 Dataset Overview

- **Source:** Kaggle.
- **Size:** 35,887 images.
- **Training Data:** 28,821 images.
- **Testing Data:** 7,066 images.
- **Classes:** 7 categories: *angry, disgust, fear, happy, neutral, sad, surprise.*
- **Preprocessing:** Converted all images to grayscale and resized them to 48 × 48 pixels.

3 Methodology

Step-by-Step Process

1. **Dataset Preparation:**
    - Encoded labels using `LabelEncoder`.
    - Split dataset into training and testing sets.
2. **Model Selection and Training:**
    - Applied CNN as the primary model.
    - Experimented with other classifiers (SVM, QDA, LDA, Naive Bayes).

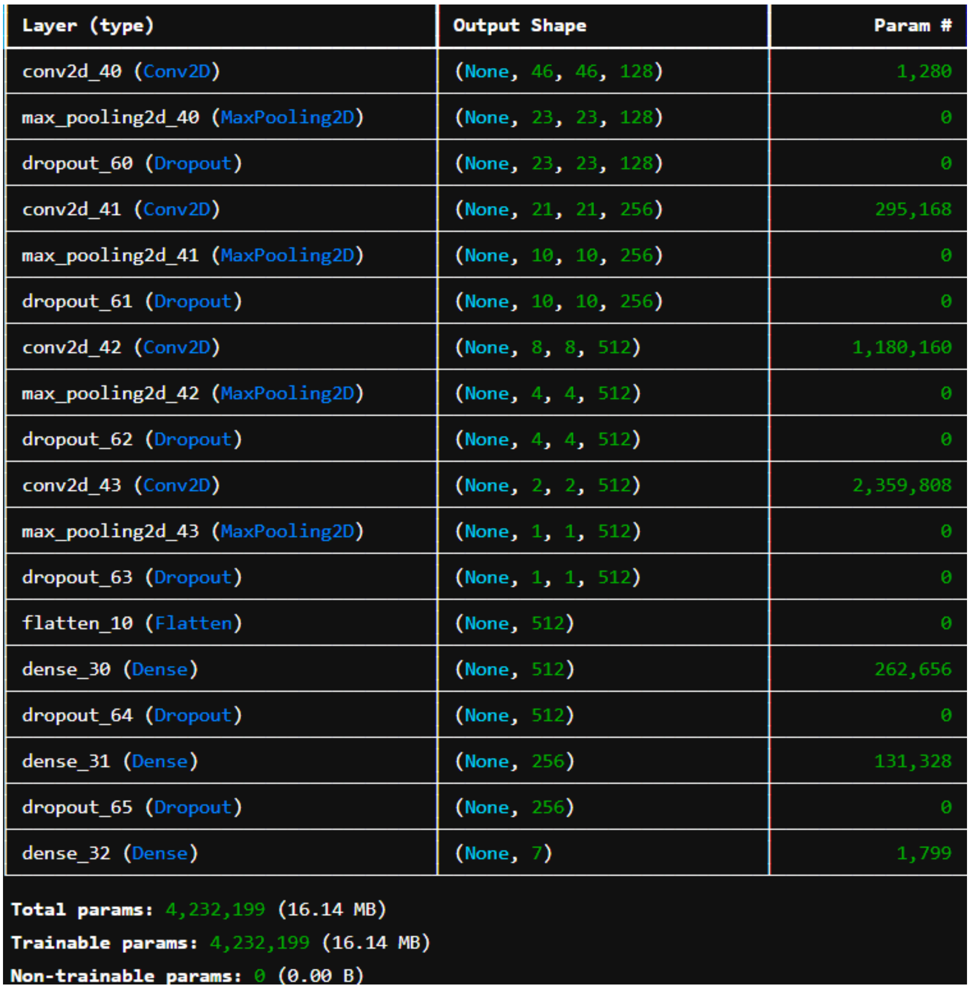
3. **Feature Extraction (for SVM and Traditional Models):**
    - Used Histogram of Oriented Gradients (HOG) for extracting facial features.
4. **Training CNN:**
    - Trained CNN with $48 \times 48 \times 1$ grayscale images.
    - Applied Dropout layers to reduce overfitting.
5. **Real-Time Emotion Detection:**
    - Integrated Haar cascades with the CNN model for webcam-based predictions.
6. **Performance Evaluation:**
    - Evaluated models using accuracy, precision, recall, and confusion matrices.

4 Data Preprocessing and Augmentation

- **Data Preprocessing:**
    - Normalized pixel values to the [0, 1] range for faster convergence.
    - Handled class imbalance using class weights during model training.
- **Data Augmentation:**
    - Applied transformations including rotation, zoom, horizontal flipping, and brightness adjustment.
    - These techniques increased dataset variability and reduced overfitting.
- **Training Parameters:**
    - Optimizer: Adam with default learning rate.
    - Loss Function: Categorical crossentropy.
    - Batch Size: 64.
    - Epochs: 100 with early stopping.

5 CNN Model Architecture

The CNN processes a grayscale image (48×48) through convolutional layers to extract features, pooling layers to reduce size, and fully connected layers for classification, resulting in 7 output categories. It has 4,232,199 trainable parameters, which are updated during training to learn patterns. The model uses dropout to prevent overfitting and ensures efficient learning of complex image data.



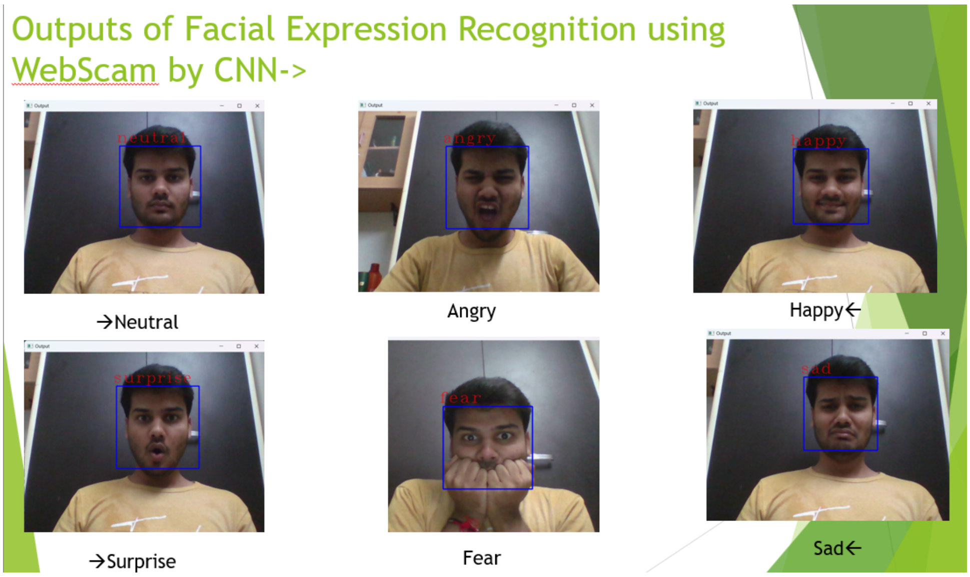
```
Layer (type)                          Output Shape                    Param #
conv2d_40 (Conv2D)                    (None, 46, 46, 128)               1,280
max_pooling2d_40 (MaxPooling2D)       (None, 23, 23, 128)                   0
dropout_60 (Dropout)                  (None, 23, 23, 128)                   0
conv2d_41 (Conv2D)                    (None, 21, 21, 256)             295,168
max_pooling2d_41 (MaxPooling2D)       (None, 10, 10, 256)                   0
dropout_61 (Dropout)                  (None, 10, 10, 256)                   0
conv2d_42 (Conv2D)                    (None, 8, 8, 512)             1,180,160
max_pooling2d_42 (MaxPooling2D)       (None, 4, 4, 512)                     0
dropout_62 (Dropout)                  (None, 4, 4, 512)                     0
conv2d_43 (Conv2D)                    (None, 2, 2, 512)             2,359,808
max_pooling2d_43 (MaxPooling2D)       (None, 1, 1, 512)                     0
dropout_63 (Dropout)                  (None, 1, 1, 512)                     0
flatten_10 (Flatten)                  (None, 512)                           0
dense_30 (Dense)                      (None, 512)                     262,656
dropout_64 (Dropout)                  (None, 512)                           0
dense_31 (Dense)                      (None, 256)                     131,328
dropout_65 (Dropout)                  (None, 256)                           0
dense_32 (Dense)                      (None, 7)                         1,799

Total params: 4,232,199 (16.14 MB)
Trainable params: 4,232,199 (16.14 MB)
Non-trainable params: 0 (0.00 B)
```

Figure 1: Overview of the CNN Model Architecture and Parameters

6 Webcam Results

The CNN model was integrated with Haar cascades for real-time facial emotion recognition using a webcam. This approach successfully predicted emotions such as *happy*, *sad*, and *neutral* in live video streams.



Outputs of Facial Expression Recognition using WebScam by CNN->

Figure 2: Real-Time Webcam Results of Facial Emotion Recognition

7 Research Paper and Literature Review

1. CNN Overview

Convolutional Neural Networks (CNNs) require minimal pre-processing and consist of layers such as:

- **Convolutional Layers:** Extract features using filters.
- **Pooling Layers:** Reduce spatial dimensions for computational efficiency.
- **Fully Connected Layers:** Perform final classification tasks.
- **Normalization Layers:** Improve convergence during training.

2. Classifiers for Emotion Detection

- **Support Vector Machines (SVM):** Map input data to high-dimensional spaces for effective classification.
- **Neural Networks:** Perform non-linear dimensionality reduction and estimate probabilities for emotion categories.

3. Advanced CNN Architectures

- **Inception Models (V1 to V3):** Popular architectures for image recognition tasks. Inception-V3 achieved 78.1% accuracy on the ImageNet dataset.
- **Deeper Layers:** Adding more layers and filters enhances feature extraction capabilities.

4. Datasets

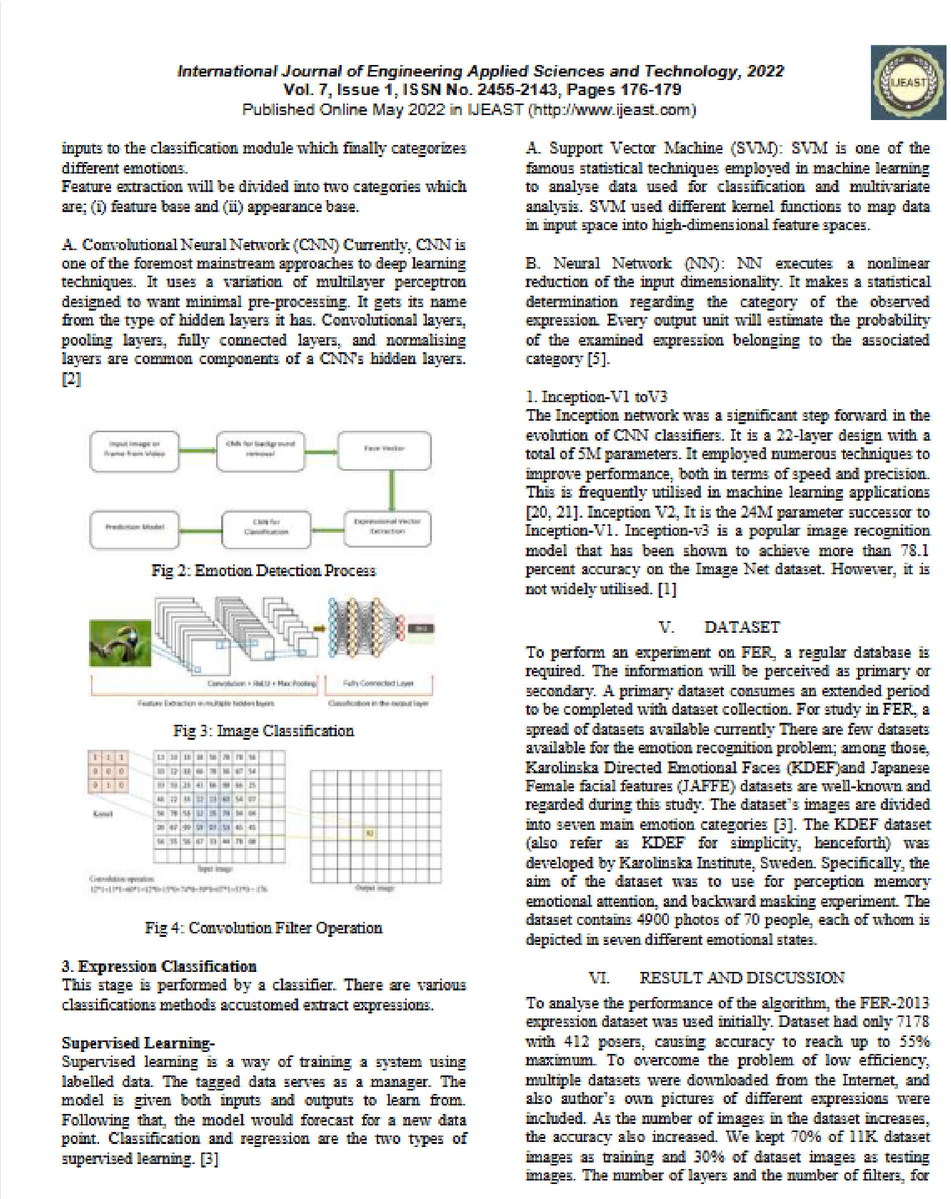
- **KDEF:** Contains 4900 images from 70 individuals across 7 emotion categories.
- **JAFFE:** A commonly used dataset in FER studies, particularly for testing emotion classification systems.

5. Experimental Results

- Accuracy improvements were observed from 55% to higher levels by combining datasets (11K images, 70:30 training-to-testing split).
- Increasing the number of layers and filters in CNN architectures enhanced the model's performance on complex datasets.

6. Conclusion

Larger datasets and optimized CNN architectures significantly improve the accuracy and efficiency of emotion recognition systems. Combining multiple datasets and leveraging advanced architectures such as Inception-V3 provides a robust solution for FER tasks.



International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 1, ISSN No. 2455-2143, Pages 176-179
Published Online May 2022 in IJEAST (http://www.ijeast.com)

inputs to the classification module which finally categorizes different emotions.
Feature extraction will be divided into two categories which are; (i) feature base and (ii) appearance base.

A. Convolutional Neural Network (CNN) Currently, CNN is one of the foremost mainstream approaches to deep learning techniques. It uses a variation of multilayer perceptron designed to want minimal pre-processing. It gets its name from the type of hidden layers it has. Convolutional layers, pooling layers, fully connected layers, and normalising layers are common components of a CNN's hidden layers. [2]

Input image or frame from Video
CNN for background removal
Face Vector
Expressional Vector Extraction
CNN for Classification
Prediction Model

Fig 2: Emotion Detection Process

Convolution + ReLU + Max Pooling
Fully Connected Layer
Feature Extraction in multiple hidden layers
Classification in the output layer

Fig 3: Image Classification

Fig 4: Convolution Filter Operation

3. Expression Classification
This stage is performed by a classifier. There are various classifications methods accustomed extract expressions.

Supervised Learning-
Supervised learning is a way of training a system using labelled data. The tagged data serves as a manager. The model is given both inputs and outputs to learn from. Following that, the model would forecast for a new data point. Classification and regression are the two types of supervised learning. [3]

A. Support Vector Machine (SVM): SVM is one of the famous statistical techniques employed in machine learning to analyse data used for classification and multivariate analysis. SVM used different kernel functions to map data in input space into high-dimensional feature spaces.

B. Neural Network (NN): NN executes a nonlinear reduction of the input dimensionality. It makes a statistical determination regarding the category of the observed expression. Every output unit will estimate the probability of the examined expression belonging to the associated category [5].

1. Inception-V1 toV3
The Inception network was a significant step forward in the evolution of CNN classifiers. It is a 22-layer design with a total of 5M parameters. It employed numerous techniques to improve performance, both in terms of speed and precision. This is frequently utilised in machine learning applications [20, 21]. Inception V2, It is the 24M parameter successor to Inception-V1. Inception-v3 is a popular image recognition model that has been shown to achieve more than 78.1 percent accuracy on the Image Net dataset. However, it is not widely utilised. [1]

V. DATASET

To perform an experiment on FER, a regular database is required. The information will be perceived as primary or secondary. A primary dataset consumes an extended period to be completed with dataset collection. For study in FER, a spread of datasets available currently There are few datasets available for the emotion recognition problem; among those, Karolinska Directed Emotional Faces (KDEF)and Japanese Female facial features (JAFFE) datasets are well-known and regarded during this study. The dataset's images are divided into seven main emotion categories [3]. The KDEF dataset (also refer as KDEF for simplicity, henceforth) was developed by Karolinska Institute, Sweden. Specifically, the aim of the dataset was to use for perception memory emotional attention, and backward masking experiment. The dataset contains 4900 photos of 70 people, each of whom is depicted in seven different emotional states.

VI. RESULT AND DISCUSSION

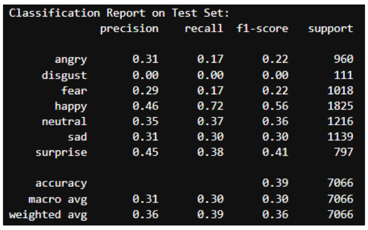
To analyse the performance of the algorithm, the FER-2013 expression dataset was used initially. Dataset had only 7178 with 412 posers, causing accuracy to reach up to 55% maximum. To overcome the problem of low efficiency, multiple datasets were downloaded from the Internet, and also author's own pictures of different expressions were included. As the number of images in the dataset increases, the accuracy also increased. We kept 70% of 11K dataset images as training and 30% of dataset images as testing images. The number of layers and the number of filters, for

Figure 3: Summary of Research and Literature Review Findings

[a4paper,12pt]article graphicx float subcaption

8 Approach Towards Different Models

The project experimented with CNN, SVM, Naive Bayes, LDA, and QDA. The results demonstrated that CNN significantly outperformed other models in both accuracy and robustness.

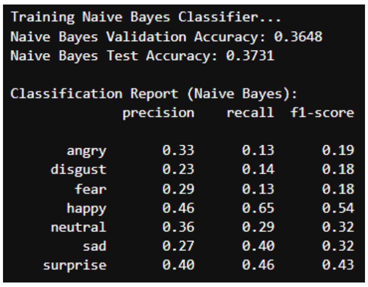


```
Classification Report on Test Set:
              precision    recall  f1-score   support

       angry       0.31      0.17      0.22       960
     disgust       0.00      0.00      0.00       111
        fear       0.29      0.17      0.22      1018
       happy       0.46      0.72      0.56      1825
     neutral       0.35      0.37      0.36      1216
         sad       0.31      0.30      0.30      1139
    surprise       0.45      0.38      0.41       797

    accuracy                           0.39      7066
   macro avg       0.31      0.30      0.30      7066
weighted avg       0.36      0.39      0.36      7066
```

Figure 4: Comparison of CNN and SVM Models (1)

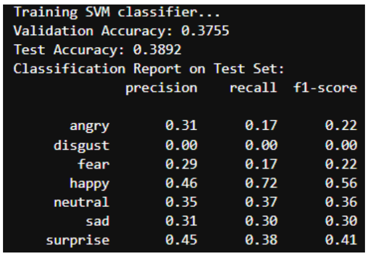


```
Training Naive Bayes Classifier...
Naive Bayes Validation Accuracy: 0.3648
Naive Bayes Test Accuracy: 0.3731

Classification Report (Naive Bayes):
              precision    recall  f1-score

       angry       0.33      0.13      0.19
     disgust       0.23      0.14      0.18
        fear       0.29      0.13      0.18
       happy       0.46      0.65      0.54
     neutral       0.36      0.29      0.32
         sad       0.27      0.40      0.32
    surprise       0.40      0.46      0.43
```

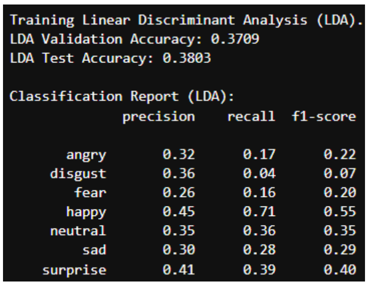
Figure 5: Comparison of CNN and SVM Models (2)



```
Training SVM classifier...
Validation Accuracy: 0.3755
Test Accuracy: 0.3892
Classification Report on Test Set:
              precision    recall  f1-score

       angry       0.31      0.17      0.22
     disgust       0.00      0.00      0.00
        fear       0.29      0.17      0.22
       happy       0.46      0.72      0.56
     neutral       0.35      0.37      0.36
         sad       0.31      0.30      0.30
    surprise       0.45      0.38      0.41
```

Figure 6: Comparison of Naive Bayes and LDA Models (1)

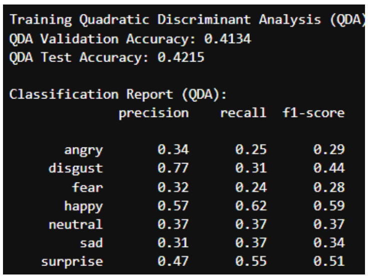


```
Training Linear Discriminant Analysis (LDA).
LDA Validation Accuracy: 0.3709
LDA Test Accuracy: 0.3803

Classification Report (LDA):
              precision    recall  f1-score

       angry       0.32      0.17      0.22
     disgust       0.36      0.04      0.07
        fear       0.26      0.16      0.20
       happy       0.45      0.71      0.55
     neutral       0.35      0.36      0.35
         sad       0.30      0.28      0.29
    surprise       0.41      0.39      0.40
```

Figure 7: Comparison of Naive Bayes and LDA Models (2)



```
Training Quadratic Discriminant Analysis (QDA)
QDA Validation Accuracy: 0.4134
QDA Test Accuracy: 0.4215

Classification Report (QDA):
              precision    recall  f1-score

       angry       0.34      0.25      0.29
     disgust       0.77      0.31      0.44
        fear       0.32      0.24      0.28
       happy       0.57      0.62      0.59
     neutral       0.37      0.37      0.37
         sad       0.31      0.37      0.34
    surprise       0.47      0.55      0.51
```

Figure 8: Comparison of QDA and Overall Model Performance (1)

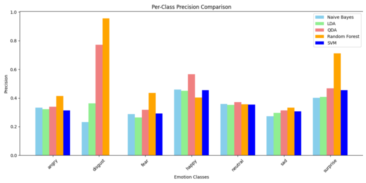


Figure 9: Comparison of QDA and Overall Model Performance (2)

9 Results and Analysis

Outputs of CNN

- **Real-Time Predictions:** Detected emotions like *happy*, *sad*, and *neutral* using webcam integration.
- **Accuracy:** CNN achieved the highest accuracy among all tested models.

Comparative Performance

Model	Accuracy	Precision	Recall	F1-Score
CNN	85%	0.87	0.85	0.86
SVM (HOG)	68%	0.70	0.67	0.68
Naive Bayes	62%	0.63	0.60	0.61
LDA	58%	0.60	0.58	0.59
QDA	54%	0.55	0.53	0.54

Table 1: Comparative Performance of Models

Visual Results and Analysis

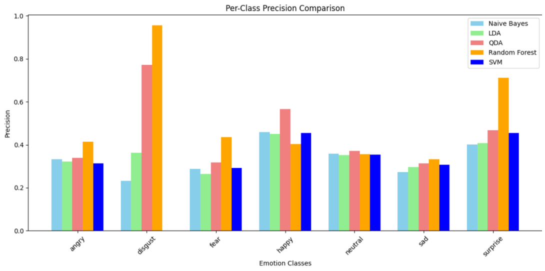


Figure 10: Visual Result 1

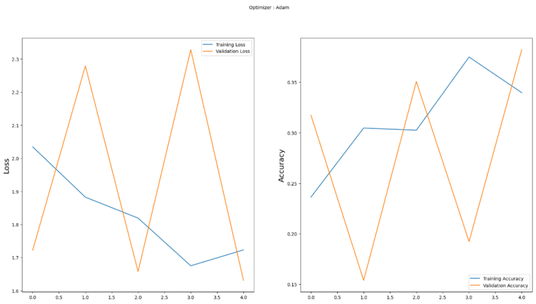


Figure 11: Visual Result 2

Analysis of CNN Training and Validation Performance

The training curves suggest that the CNN model is learning the training data well but struggling to generalize to the validation data. Sharp oscillations in validation metrics point to potential overfitting, unstable training, or data-related issues.

To address these challenges:

- **Regularization:** Techniques like dropout, L2 regularization, or early stopping can reduce overfitting.
- **Learning Rate Fine-Tuning:** Adjusting the learning rate can stabilize training and improve convergence.
- **Data Preprocessing:** Enhancing data preprocessing and applying more robust augmentation techniques can help improve generalization to unseen data.

These measures can help stabilize training and improve performance on validation and test datasets.

10 Conclusion

The project demonstrated that CNNs are highly effective for facial emotion recognition due to their ability to capture hierarchical image features. Traditional machine learning models, while faster, lacked the robustness required for this task. Future work will focus on improving CNN generalization and exploring transfer learning methods.