Deep Learning for Facial Expression Recognition

Dinesh Karnati
North Carolina State University
Raleigh, North Carolina, USA
drkarnat@ncsu.edu

Jake McDavitt
North Carolina State University
Raleigh, North Carolina, USA
jfmcdavi@ncsu.edu

Pratik Thapa
North Carolina State University
Raleigh, North Carolina, USA
pthapa4@ncsu.edu

Sam Ferguson
North Carolina State University
Raleigh, North Carolina, USA
sfergus4@ncsu.edu

Abstract

Facial Expression Recognition (FER) is a key factor in improving human-computer interaction, having applications in industries such as healthcare, automotive, and robotics. However, accurately identifying emotions from facial cues remains a challenge due to the wide variability in facial contours, edging, expressions, and lighting conditions. In this paper, we present an evaluation of deep learning architectures for FER classifications using the FER-2014 dataset. We implement and analyze four different neural network models: multi-layer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN), and residual network (ResNet) to determine their effectiveness in classifying seven emotion categories. We conducted further analysis evaluating the performance of these neural network models in comparison to largely available Large Language Models with image processing capabilities.

1 Problem Statement

Facial expressions play an important role in human social behavior and communication. The recognition of and discrimination between different expressions is essential in our interpretation of one's thoughts and emotional state, with some research suggesting that the face is responsible for 55% of all information transmitted between humans [2] [12]. In the context of artificial intelligence and autonomous systems, this ability is critical for encoding emotional sentiment, which is of practical relevance to a number of applications. These include sociable robotics, medical treatment, driver fatigue surveillance, and human-computer interaction systems [10]. In the past decade, deep learning methods have seen considerable success across a wide array of image classification tasks, along with the continued development of a number of computer vision benchmarks. These include a number of datasets targeted at emotion detection and facial recognition, such as FER2013 [8], RAF-DB [11], AffectNet [13], and EmotioNet [6]. In this project, we will be applying a number of deep learning and machine learning approaches to the facial expression recognition task, comparing and evaluating their performance on our selected data set.

2 Literature

A number of previous works have discussed the application of deep learning approaches to the facial expression recognition task. Li and Deng performed a survey that examined the most relevant approaches [10]. The most prominent architecture described was the convolutional neural network (CNN), which has been used in numerous papers regarding facial expression recognition. Several authors have found that CNNs are robust to positional translation and scale variations [7] [1]. Additionally, the learnable filters in the convolutional layers of CNNs are effective at identifying different types of facial characteristics. Authors such as [17] [16] used deep autoencoders for the FER task, which learn efficient encodings of the facial features to perform dimensionality reduction. Autoencoders were shown to significantly outperform PCA in reducing dimensions while preserving feature recognition rate [17]. For detecting facial emotion and expression in video content, some groups have used recurrent neural networks (RNN) to capture features over long time-series. In [5], facial features were extracted on each frame using a CNN, and these features were fed into an RNN model to predict the response class. Some research has been done on the use of ensemble networks for FER classification, including both at the feature-level [4], and to generate a more robust prediction for the response.

3 Dataset

To investigate our hypothesis, we utilized a widely recognized dataset, FER-2013 [14]. This dataset serves as a benchmark in the field of facial emotion recognition and provides complementary characteristics for model evaluation.

The FER-2013 dataset contains approximately 35,900 grayscale facial images, each of size 48×48 pixels, annotated with one of seven emotion categories. It is split into 28,709 training images and 3,589 test images. These images exhibit a wide range of facial variations and lighting conditions, making the dataset suitable for training robust and generalizable emotion recognition models.

The FER-2013 provides a robust benchmark suite for training and evaluating facial emotion recognition models. It's diversity in emotion categories, expression intensity, subject demographics, and imaging conditions enhances the generalizability and reliability of the models developed in our study.

4 Experiment

We conducted experiments on the FER-2013 dataset. Primarily, our goal is to compare and evaluate the performance of several different deep learning approaches on the facial expression recognition task. For this purpose, we have selected a number of deep learning

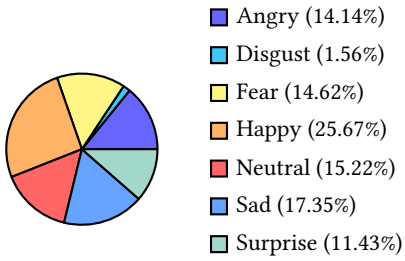


Figure 1: Distribution of Classes in the Train Split

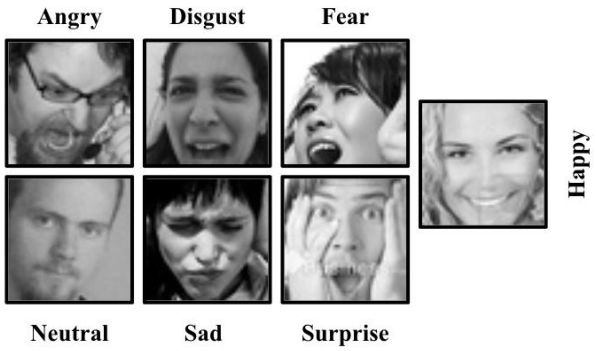


Figure 2: Sample Images from Each Class

architectures that we believe would be appropriate for expression recognition, as well as some simpler architectures to serve as a baseline for our evaluations. These include multi-layer (feed-forward) perceptrons, convolutional neural networks, recurrent neural networks, and a ResNET [9] architecture. The specific details of our model architectures and optimization procedure are outlined in section 4.3. To evaluate the performance of our models after training was completed, we used a number of different performance metrics to obtain a comprehensive report of each model's ability in identifying the different classes present in the dataset. The definitions are detailed in section 5, and include the accuracy, precision, recall, and f1-scores of each model on the test data.

4.1 Hypothesis

Through performing our experiments we seek to answer two main research questions. First, from four specific neural network designs (CNN, RNN, MLP, and ResNET), we determined which one performs the best on emotion classification when provided an image of someone's face. Then, after completing our initial experiment, we evaluated how these models compare to publicly available Large Language Models with visual encoding capabilities. We predict that ResNet will outperform all models including the LLM, due to it's specialization in image classification tasks and extended capabilities from a traditional CNN.

4.2 Rationale

The Facial Expression Recognition (FER) Problem involves models having the ability to effectively process visual data and capture spatial and temporal patterns. To address this classification problem, we chose to implement four neural network models: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Multi-Layer Perceptron (MLP), and Residual Networks (ResNET).

We chose to develop a CNN model because they are effective in extracting spatial features such as edges, shapes, and textures from images, which are crucial for determining emotions. It takes the input image and breaks it down into smaller pieces in order to find underlying patterns. It detects these low-level patterns and progressively captures higher-level features that distinguish emotion-specific textures [15].

We implemented an MLP model to explore how a fully connected feed-forward network could perform on pixel data. MLPs convert input data into a flattened vector, and do not specifically account for spatial relationships. These models tend to be simpler and faster to train, which make them a good starting point for our experiments.

Recurrent neural networks (RNNs) are typically used for sequential data, such as video, audio, time-series, text, etc., as they remember important aspects about each step of the input, helping inform its future predictions. Although RNN's are not traditionally applied to still images, we were curious about how it would perform when applied to image-based FER tasks.

Finally, we incorporated a fine-tuned version of the ResNET architecture. This model extends the capabilities of a CNN through the use of residual connections. Traditional deep networks struggle with the problem of vanishing gradients, leading to a degradation in performance. ResNet addresses these issues by using residual connections that allow for gradients to flow more easily through the network, making it easier to train deep architecture and improving their ability to learn complex features [3].

4.3 Methodology

Models were trained on the FER-2013 train dataset (28,7090) images. Each model was trained for 30 epochs, using the ADAM optimizer with default settings (lr=0.001) and a batch size of 64. Since our raw input was grayscale 48x48 images, no image resizing was necessary or performed. The architectures of our MLP, CNN, and RNN models are described in figure 3, with the ResNet-34 architecture imported through torchvision's models library. Code for the models is available at https://github.com/CSC522NCSU/FacialNN.

After training and evaluating the models, we took a proportional random sample from the testing data for each class and asked ChatGPT-4 Turbo to classify each image into one of the seven emotions. We then took those same samples and had each of our models classify them to directly compare accuracy with the LLM.

5 Evaluation Metrics

To evaluate the performance of our facial expression recognition models, we employed a combination of metrics that offer insights into different aspects of model accuracy. These metrics are the confusion matrix, precision, recall, F1-score, and accuracy. Together, these metrics provided us with a comprehensive overview of model performance across all the classes in the dataset.

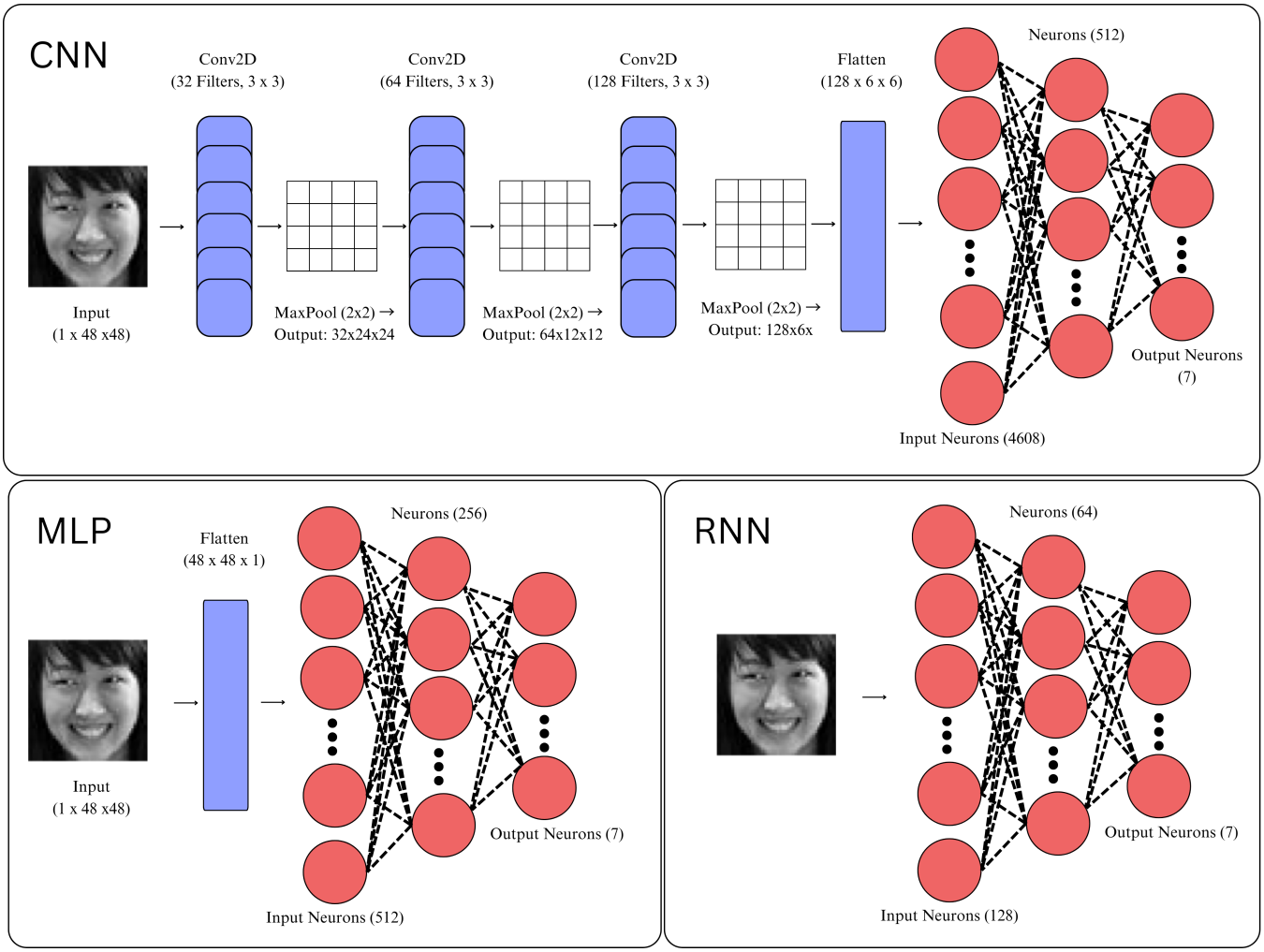


Figure 3: Overview of all the model architectures

Confusion Matrix: The confusion matrix is a table representation of actual versus predicted classifications. Each row represents the instances in an actual class, and each column represents the instances in a predicted class.

From the confusion matrix, we can derive the following labels for each class i:

- **True Positive (TP)**: Correct predictions of class i.
- **False Positive (FP)**: Instances incorrectly predicted as class i.
- **False Negative (FN)**: Instances of class i predicted as another class.
- **True Negative (TN)**: All other correctly predicted instances.

Precision: Precision measures how often positive classifications in a machine learning model are actually positive. It is defined as:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}$$

Recall: Recall measures the proportion of all actual positives that were correctly classified as positives. It is defined as:

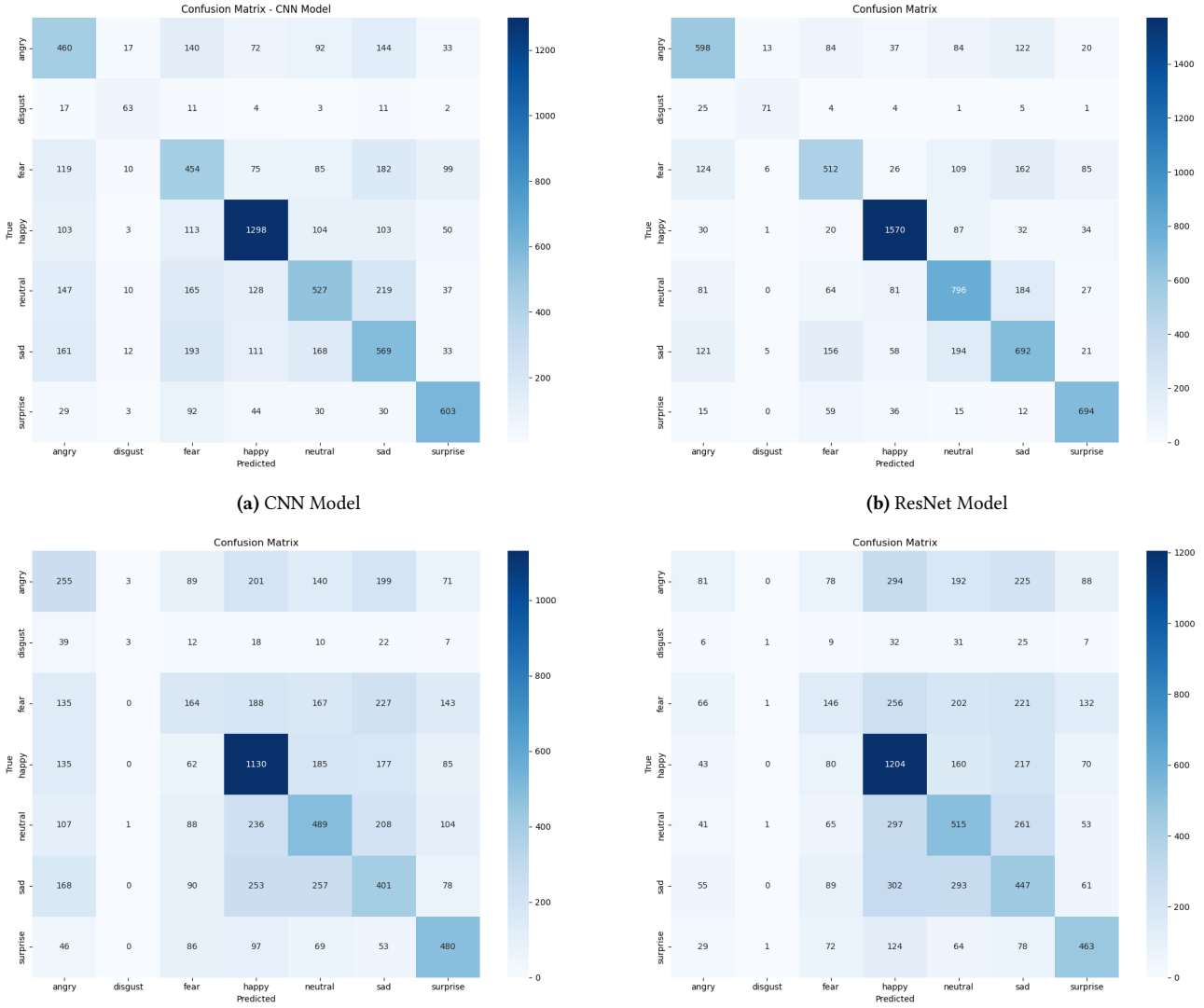
$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}$$

F1-score: The F1-score is the harmonic mean of precision and recall, balancing both false positives and false negatives. It is calculated as follows:

$$\text{F1}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

Accuracy: Accuracy is the proportion of all classifications that were correct:

$$\text{Accuracy} = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$$



(a) CNN Model

(b) ResNet Model

(c) MLP Model

(d) RNN Model

Figure 4: Confusion matrices for the four different models: (a) CNN, (b) ResNet, (c) MLP, and (d) RNN.

6 Neural Networks Results

6.1 CNN

Figure 4 (a) displays the confusion matrix for the CNN model, highlighting its performance in classifying emotions. The diagonal values represent correct predictions, with "Happy" achieving the highest accuracy at 1,298 correct classifications, indicating the model's strong ability to classify this emotion. Conversely, "Disgust" is the lowest performing class with only 63 correct classifications, likely due to insufficient training data. Misclassifications are all the values off-diagonal. The model has some confusion between classes "Sad" and "Neutral" (219 misclassifications) and between "Fear" and "Sad" (193 misclassifications). While the model performs well for expressive emotions like "Happy" and "Surprise," it struggles with subtle or underrepresented emotions, suggesting a need for more balanced datasets and fine-tuned feature extraction processes to improve classification accuracy.

Figure 5 (a) presents a bar graph illustrating the performance of the CNN model across seven emotion classes using three evaluation metrics: precision, recall and f1 score. Among these classes, happy and surprise exhibit higher precision and recall values, resulting in stronger F1 scores compared to other emotions. This result indicates that the model is more effective at accurately identifying these two emotions, both in terms of correctly predicting them when they occur and minimize false positives.

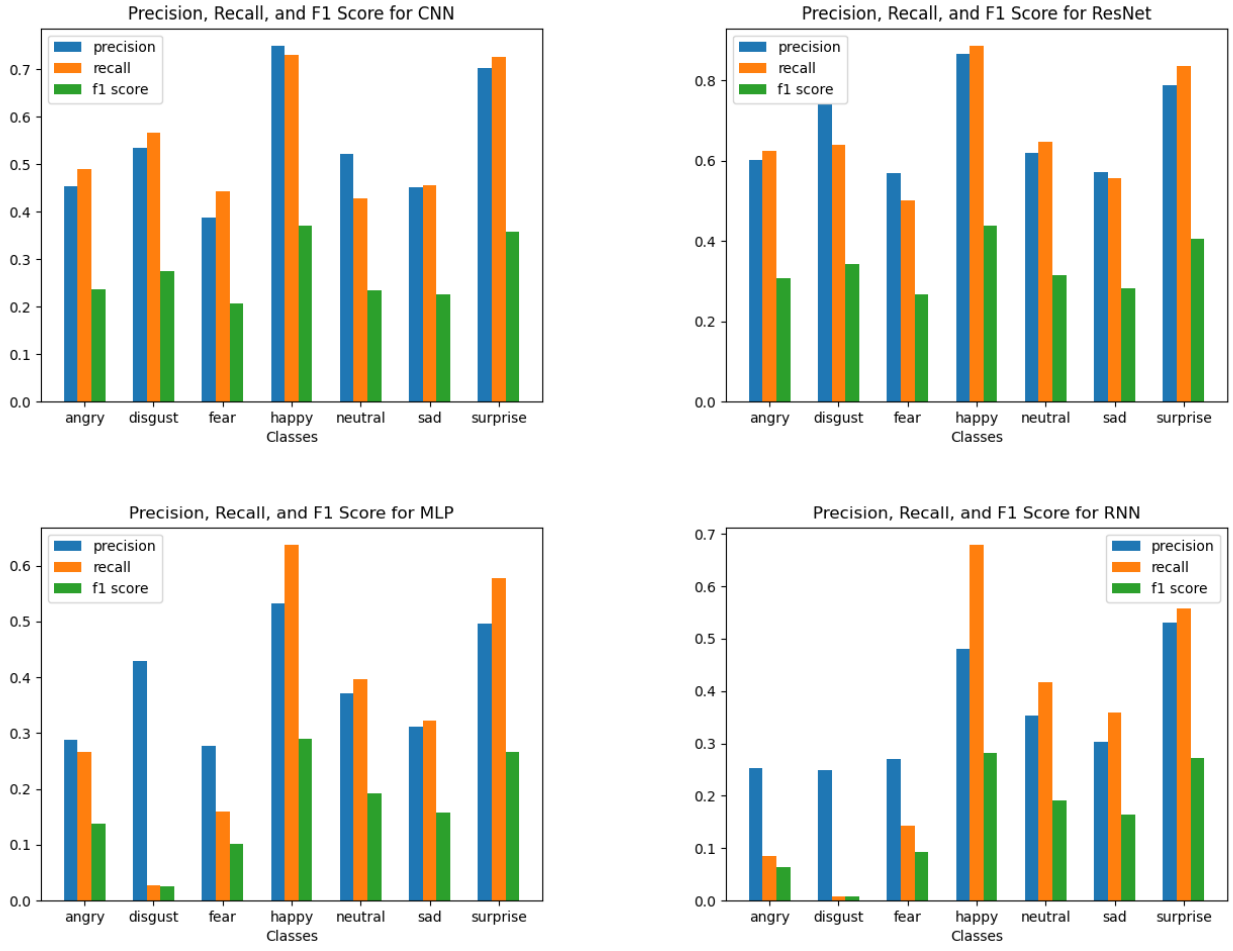


Figure 5: Per class metric for the four different models: (a) CNN, (b) ResNet, (c) MLP, and (d) RNN.

6.2 ResNET

Figure 4 (b) shows the confusion matrix for the ResNET model. Similar to the CNN model, the ResNET model showed the highest proficiency in correctly classifying "Happy" images with 1570 correct classifications. In general, this model had a better classification accuracy as shown through the higher values in the diagonal compared to the CNN's confusion matrix. "Disgust" remains the lowest-performing class, with only 71 correct classifications. Misclassifications also occur in roughly the same classes suggesting overlapping features in these emotion categories.

Figure 5 (b) presents a bar graph illustrating the performance of ResNet-34 model across seven emotion classes used for evaluation. The result is quite distinctive, as ResNet outperformed CNN, achieving higher precision, recall and f1 scores across most classes. Notably, the model surpasses the 0.8 threshold in several cases, as indicated on the y-axis. Similar to the CNN, ResNet-34 performs particularly well on happy and surprise classes. This improvement can be attributed to the ResNet-34's pre-trained architecture, which enables it to capture more complex and abstract patterns within the data.

6.3 MLP

Figure 4 (c) shows the results of the multi-layer perception. Similar to the other models, the MLP was successfully able to distinguish the 'happy' class from the other labels in the dataset. Looking at the performance metrics in Figure 5, we can see that the model had difficulty identifying the 'disgust' class, achieving a recall and f1-score of less than 10%. Looking at the confusion matrix, we can see that the model predicted the 'disgust' class only 7 times, 2 of which were correct. Overall, the MLP achieved an accuracy of 40.71% on the validation data.

Figure 5 (c) presents a bar graph illustrating the performance of the MLP model across seven emotion classes used for evaluation. In comparison to CNN and ResNet, the MLP model demonstrates significantly lower overall performance. Unlike CNNs, MLPs do not utilize convolutional layers, which are essential for extracting

spatial features from images and refining them before classification. This lack of feature extraction capability likely contributes to the MLP's inferior performance. Nevertheless, the model shows relatively better accuracy for the happy and surprise classes. It is also worth noting that the precision for most classes is lower than the recall, indicating that while the model is able to identify many relevant instances (high recall), it also generates more false positives, reducing its precision.

6.4 RNN

Figure 4 (d) shows the results of a recurrent neural network. The RNN model was most successful at identifying the 'happy' classes and 'surprise' classes, with the high recall score indicating it performs well at identifying true positives for those classes. However it particularly struggled with the 'angry', 'disgust', and 'fear' classes compared to the other models.

Figure 5 (d) presents a bar graph illustrating the performance of the RNN model across seven emotion classes used for evaluation. When compared to the MLP model, the RNN shows similar trends in precision, recall, and F1 score across most classes. This similarity in performance may be attributed to the fact that both models lack convolutional layers, which are critical for capturing spatial features in image data.

Table 1: Model accuracy comparison across architectures

Model	Accuracy
ResNet	0.6872
CNN	0.5549
MLP	0.4071
RNN	0.3980

6.5 Overall Accuracy

Table 1 summarizes the final accuracy of each model on the FER-2013 test set. Among the four models we have developed, the ResNet model achieved the highest accuracy of 68.72%, followed by the CNN at 55.49%. The MLP and the RNN models lagged significantly behind.

The results align with some of our assumptions about convolutional architectures (CNN, ResNet). ResNet's superior performance can be attributed to its residual connections, which help mitigate vanishing gradient issues and allow for more effective training of deep networks. The CNN model also performed strongly, validating its effectiveness at extracting hierarchical spatial features from facial images.

In contrast, the MLP and RNN architectures underperformed. The MLP lacks spatial awareness since it treats the pixel data from the images as flat vectors, hindering the ability to recognize facial patterns. The RNN is known to be well-suited for sequential data, so its application to static images likely limited its ability to capture meaningful patterns.

7 Neural Networks vs LLM

In this section we explore how our trained Neural Network models compare at classifying emotions compared to a widely available large language model, ChatGPT (specifically GPT-4 Turbo). Figure 6 shows a matrix of the 35 image samples we used for the LLM and our models for the FER image classifiation task.

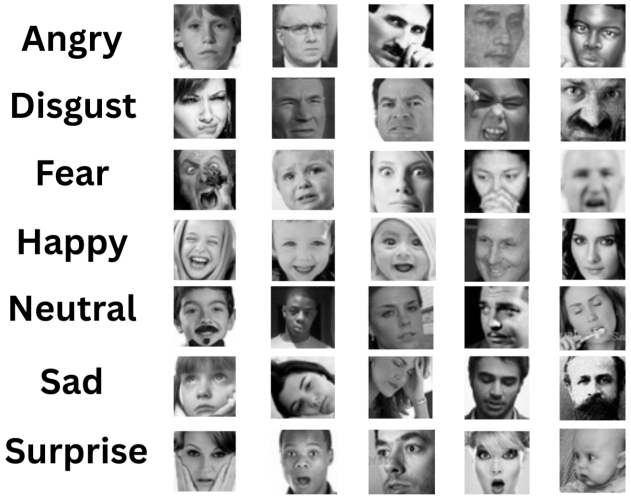


Figure 6: Image Samples used to test LLM and Generated NN Models

When testing ChatGPT-4 Turbo's FER capabilities, we uploaded the 35 images and prompted the LLM with the following prompt for each image: "From the image provided, what emotion does the person in it seem to have? Your options to choose from are: angry, disgust, fear, happy, neutral, sad, and surprise. Keep your response to only the emotion you choose". The LLM would then respond with a single word identifying the response class of the emotion it believed was associated with the image. Figure **??** presents a confusion matrix of the results from our testing. Immediately, we noticed that the LLM performed well at correctly identifying images that were labeled as 'Sad' or 'Surprise', predicting 4 out of the 5 samples correctly for each of those label categories. It also performed well in identifying images labeled 'Happy' and 'Neutral', correctly identifying 3 out of 5 samples for those labels. However, the LLM struggled with the last three labels: 'angry', 'disgust', and 'fear', only correctly labeling 2 out of 5 samples for 'disgust' and 'fear' and only getting 1 out of 5 samples correct for the 'angry' label.

7.1 Comparison

With the rapid advancement of large language models, especially those with multimodal capabilities like GPT-4 Turbo, we were initially curious and optimistic about how well they could perform on image-based tasks like facial expression recognition (FER). Given their success across a range of tasks like reasoning, vision, and natural language, we speculated that the LLM might outperform our custom-trained neural networks.

However, our results were surprising. As shown in Table 2, while GPT-4 Turbo demonstrated strong performance in certain emotion categories such as Sad, Surprise, Happy, and Neutral, its overall classification accuracy (54.29%) fell short of both the ResNet (57.14%)

Model	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Overall
MLP	0%	0%	20%	20%	20%	0%	80%	20%
CNN	40%	60%	60%	100%	0%	60%	100%	60%
Resnet	40%	60%	0%	80%	40%	80%	100%	57.14%
RNN	0%	0%	80%	0%	0%	0%	20%	14.29%
ChatGPT-4 Turbo	20%	40%	40%	60%	60%	80%	80%	54.29%

Table 2: Per-class and overall accuracy across different models

and CNN (60%) models. Notably, the LLM struggled with recognizing more subtle or less frequently occurring emotions like Angry, Disgust, and Fear, where even our CNN showed significantly higher classification accuracy.

One important distinction lies in the training objectives. Our deep learning models were trained specifically on facial expression classification using tens of thousands of labeled images. However, GPT-4 Turbo, although capable of interpreting visual input, has not been fine-tuned for emotion classification in the same manner. This lack of targeted training likely contributes to its weaker performance than the other CNN-based models.

Overall, the GPT-4 Turbo showed a promising baseline performance without any task-specific training, reinforcing the value of purpose-built neural networks for visual classification problems. These findings show that, despite the versatility of LLMs, they are not yet a holistic replacement for specialized models when it comes to detailed tasks such as FER - at least not without fine-tuning.

8 Final Conclusion / Discussion

We report that convolutional architectures (CNN and ResNet) achieved the best performance on the FER task, with the MLP and RNN models displaying a significantly lower accuracy. In terms of the class difficulty, our confusion matrix diagrams in figure 4 show that correct identification of the 'happy' class was consistent throughout the models, while recognition of the 'disgust' class was more challenging, particularly for the MLP and RNN architectures. This is also reflected in the accuracy report from figure 5. We have successfully shown that deep learning models are capable of learning the features required for facial expression recognition, reiterating the findings from many of the papers cited in our literature review. Additionally, we have evaluated which architectures perform the best on this task, and have provided insight and possible explanations as to why certain models perform better than others.

References

[1] 2002. Head-pose invariant facial expression recognition using convolutional neural networks. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*. IEEE Computer Society, 529.

[2] Sharmeen M Saleem Abdullah and Adnan Mohsin Abdulazeez. 2021. Facial expression recognition based on deep learning convolution neural network: A review. *Journal of Soft Computing and Data Mining* 2, 1 (2021), 53–65.

[3] Babina Banjara. 2025. Deep Residual Learning for Image Recognition (ResNet Explained). https://www.analyticsvidhya.com/blog/2023/02/deep-residual-learning-for-image-recognition-resnet-explained/ Accessed: 2025-04-13.

[4] Sarah Adel Bargal, Emad Barsoum, Cristian Canton Ferrer, and Cha Zhang. 2016. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 433–436.

[5] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 467–474.

[6] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. 2016. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5562–5570.

[7] B. Fasel. 2002. Robust face analysis using convolutional neural networks. In *Proceedings of the 16th International Conference on Pattern Recognition (ICPR)*, Vol. 2. IEEE, 40–43.

[8] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *Neural information processing: 20th international conference, ICONIP 2013, daegu, korea, november 3-7, 2013. Proceedings, Part III 20*. Springer, 117–124.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV] https://arxiv.org/abs/1512.03385

[10] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE transactions on affective computing* 13, 3 (2020), 1195–1215.

[11] Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2852–2861.

[12] Wafa Mellouk and Wahida Handouzi. 2020. Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science* 175 (2020), 689–694.

[13] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (2017), 18–31.

[14] M Msambare. 2021. FER-2013 Facial Expression Recognition Dataset. https://www.kaggle.com/datasets/msambare/fer2013. Accessed: 2025-04-02.

[15] Mohd Sanad Zaki Rizvi. 2020. Learn Image Classification on 3 Datasets using Convolutional Neural Networks (CNN). https://www.analyticsvidhya.com/blog/2020/02/learn-image-classification-cnn-convolutional-neural-networks-3-datasets/ Accessed: 2025-04-13.

[16] Muhammad Usman, Siddique Latif, and Junaid Qadir. 2017. Using deep autoencoders for facial expression recognition. In *2017 13th International Conference on Emerging Technologies (ICET)*. IEEE, 1–6.

[17] Nianyin Zeng, Hong Zhang, Baoye Song, Weibo Liu, Yurong Li, and Abdullah M Dobaie. 2018. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing* 273 (2018), 643–649.