Real Time Facial Emotion Recognition using Deep Learning

Likitha Santhapalli,Nitisha Reddy Gaddam,Vasudha Maddi

December 6, 2024

Abstract

Facial Emotion Recognition (FER) plays a vital role in computer vision with applications ranging from mental health monitoring to personalized education tools and virtual assistants. This project evaluates the performance of two deep learning models one is a basic Convolutional Neural Network (CNN) and the other is a more sophisticated architecture ResEmoteNet using the FER2013 dataset. ResEmoteNet includes Squeeze-Excitation (SE) and Residual blocks enabling it to focus on critical facial features and enhance feature extraction.Both models were trained and tested under standardized conditions including consistent data preprocessing, fixed hyperparameters and identical train-test splits ensuring a fair comparison. To validate the models' real-time performance, they were tested on video inputs and optimized for real-time emotion recognition.

The results indicate that while both models performed well, ResEmoteNet demonstrated superior accuracy particularly in real-time and video-based scenarios. This highlights its potential for developing robust and efficient FER systems. Future research could focus on enhancing the models' robustness across diverse datasets and improving scalability to support broader real-world applications.

1 Introduction

Facial emotion recognition (FER) represents a fascinating and rapidly evolving domain within computer vision with the potential to transform fields like healthcare, education, security, and human-computer interaction. Emotions are at the heart of human communication and systems capable of interpreting these emotions accurately can pave the way for more personalized and impactful interactions between people and technology.

In education, for example , facial emotion recognition could help educators assess student engagement or identify those facing challenges enabling personalized teaching strategies. Healthcare applications might include diagnosing mental health conditions or tracking patients' emotional states over time. In marketing, emotion recognition could provide valuable insights into customer responses to products or campaigns, while in security it might help detect suspicious or dangerous behavior. These diverse applications highlight the transformative impact of emotion recognition technology.

Although it shows great potential, emotion detection still faces some challenges.Emotional expressions differ significantly across individuals and external factors like lighting, facial orientation, and subtle variances in expressions complicate accurate detection. Early approaches depended on manually engineered features like Local Binary Patterns (LBP) and Gabor filters, coupled with machine learning models such as Support Vector Machines (SVM).While these methods worked to some extent, they were time-consuming and had difficulty adapting to diverse datasets.

The advent of deep learning has redefined the field, with Convolutional Neural Networks (CNNs) playing a key role. CNNs learn patterns directly from data, offering improved accuracy and adaptability.Pre-trained architectures such as ResNet and VGG have further advanced the field by effectively capturing subtle emotional cues across diverse datasets.

This project aims to develop a robust facial emotion recognition system utilizing both a basic CNN and the advanced ResEmoteNet architecture for static and real-time applications. Our system will classify seven core emotions which are happiness, sadness, anger, fear, surprise, disgust, and neutrality. The FER2013 dataset,a benchmark suited for real-world scenarios is used for training and testing.

To ensure high performance, we will apply data augmentation techniques and integrate architectural enhancements including Squeeze-Excitation (SE) and Residual blocks in ResEmoteNet, to focus on critical facial features. Ultimately, this project aims to deliver a reliable FER system with applications in mental health support, interactive technologies, and a deeper understanding of human behavior.

2 Method

This project implemented two deep learning models for facial emotion recognition: a baseline Convolutional Neural Network (CNN) and a more advanced architecture named ResEmoteNet. The CNN provided a foundational benchmark, whereas ResEmoteNet included advanced features to improve performance in both static image analysis and dynamic applications like real-time and video-based emotion detection. This section details the methodology, preprocessing steps and deployment process for these models.

Model 1: Convolutional Neural Network

The baseline CNN was implemented to classify 48×48×1 grayscale images. Its architecture consists of three stages, each comprising two convolutional layers with increasing filter counts.In the initial layers, a smaller number of filters such as 64 are used to identify simple patterns like edges and textures. As the network deepens, the number of filters increased to 128 and 256 in subsequent layers enabling the model to capture more complex patterns and abstract features such as shapes and high-level structures. These layers progressively extracted simple to complex features using 3×3 filters and the ReLU activation function. To improve training stability, batch normalization was applied after each layer, followed by max pooling to downsample spatial dimensions while keeping essential features. Dropout was introduced after each pooling operation, with rates incrementally rising from 0.25 to 0.45 to mitigate overfitting.

After the convolutional layers were followed by a global average pooling layer that reduces feature maps into a compact representation. The fully connected dense layer with 128 neurons further refined these features, using batch normalization and dropout for regularization. The final output layer using a softmax activation function classified the data into seven emotion categories.This balanced architecture effectively captured feature complexity while minimizing risk of overfitting.

Model 2:ResEmoteNet

ResEmoteNet expanded upon the baseline CNN, integrating advanced features like residual connections and Squeeze-and-Excitation (SE) mechanisms. Residual Blocks helps address the issue of vanishing gradient issues and allowed for the training of deeper networks. Each block consists of two convolutional layers with batch normalization and ReLU activation.The layers are connected through a skip connection allowing the input to be added directly to the output. If there are any dimensional mismatches between the input and output, 1×1 convolutional layers are used to adjust the size.

The Squeeze-and-Excitation (SE) mechanism helps the model focus on important features by adjusting the feature maps. It does this by summarizing spatial information across different channels and applying scaling factors to highlight the most important features.This process involves two steps: "squeeze" achieved via global average pooling to capture the overall spatial information, and "excitation" which used dense layers to generate channel-specific scaling factors.

ResEmoteNet's architecture starts with three convolutional layers for initial feature extraction, followed by batch normalization, ReLU activation, and max pooling. An SE block then enhances the significant features before passing the data through three Residual Blocks.These blocks progressively increase filter sizes to capture more complex patterns. Strided convolutions in these blocks efficiently reduce spatial dimensions.

The final stages of ResEmoteNet included a Global Average Pooling layer, followed by three dense layers with 1024, 512, and 256 neurons, respectively. Dropout, set at a rate of 0.5, was applied after each dense layer to enhance generalization. A softmax activation layer classified inputs into seven emotion categories.

Preprocessing

Preprocessing played a key role in ensuring consistent input quality and making the model more robust.The images were resized to 48×48 pixels and normalized to a range of [0, 1] for stable training.Several data augmentation techniques such as rotation,zooming ,width and height shifts and horizontal flipping were applied to diversify the dataset and help prevent overfitting.Additionally,the nearest neighbor method is used to fill any new pixels created during these transformations.Grayscale images were used to reduce computational complexity while keeping important features. The dataset was split into training and validation subsets with 20% reserved for validation which helped evaluate the model's performance on unseen data. These preprocessing steps were essential in enabling the model to learn effectively and generalize to new data.

Real-Time and Video Input Adaptations

Real-time and video-based emotion detection required specialized adaptation pipelines.For real-time detection, OpenCV was used to capture live video feeds and apply a Haar cascade classifier to detect faces within each frame from live streams or video files. These detected faces were then processed by the trained models to predict emotions, ensuring low-latency inference for each frame. In the case of video analysis,individual frames were extracted and preprocessed in a similar manner to still images and predictions averaged across the frames to determine the overall emotion. This methods helped improve the the models' reliability in dynamic environments.

Both models underwent identical training conditions to ensure a fair comparison. The training process used the Adam optimizer with an initial learning rate of 0.001, categorical cross-entropy as the loss function, a batch size of 64, and a total of 50 epochs. TensorFlow and Keras frameworks were utilized for efficient model development and training.

3 Experiments

3.1 Dataset

The FER2013 dataset which is available on Kaggle, serves as a prominent benchmark for facial expression recognition tasks. It consisits grayscale images of faces each labeled with one of seven emotion categories: anger, disgust, fear, happiness, neutrality, sadness, and surprise. These images represent a wide range of facial expressions, providing a solid foundation for training and evaluating emotion detection models.With its extensive variety of emotional categories, FER2013 provides a robust basis for evaluating how well models can identify and classify human emotions.Figure 1 illustrates sample images from the dataset.

Figure 1: Sample Images

In real-time emotion detection,preprocessing plays a crucial role in achieving consistent and reliable model performance. To detect and extract facial regions from static images or video frames, the `haarcascade_frontalface_default.xml` file from OpenCV was used. This pre-trained Haar cascade model is effective at detecting frontal facial features, allowing for accurate localization and cropping of facial regions.These cropped images were then fed as input into the trained emotion classification model. By standardizing the input dimensions and maintaining consistent image quality, this preprocessing pipeline greatly improved the system's accuracy and reliability for real-time applications.

One of the main challenges with the FER2013 dataset is the class imbalance which can be observed in the uneven distribution of images across different emotion categories. For example, the training set has 7215 images labeled "Happy" which makes it the most dominant class, while "Disgust" is severely

underrepresented with only 436 images.This isssue also exists in the test set where "Happy" dominates with 1774 images compared to just 111 images for "Disgust".Other categories like "Sad","Fear" and "Neutral" have a more balanced representation, while "Surprise" and "Angry" are less common.

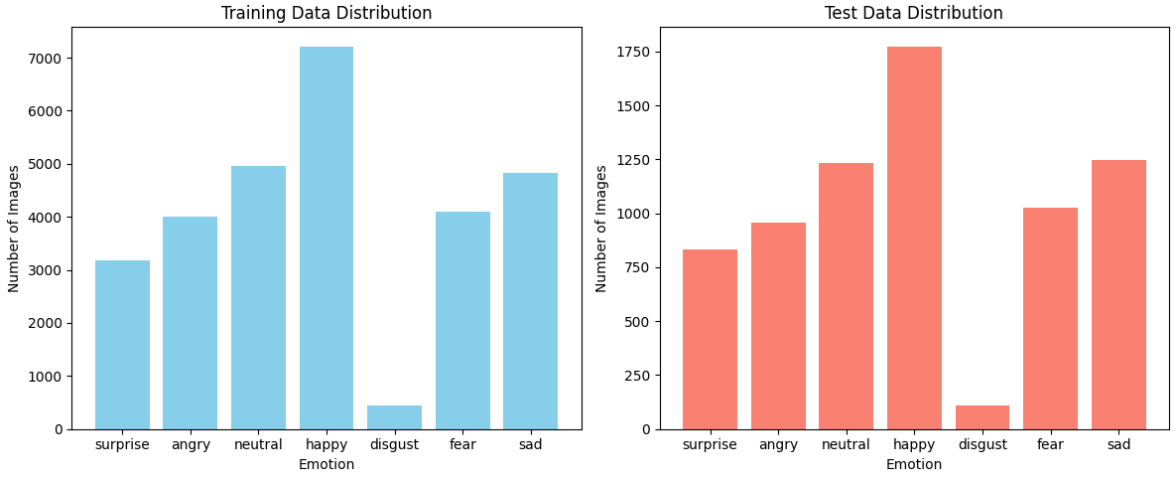

Figure 2: Class Distribution

The class imbalance in the FER 2013 dataset introduces complications in training of emotion detection models. When certain emotions are overrepresented it can lead model to favor those emotions , resulting in higher accuracy for dominant categories but poor performance on those with fewer samples. Addressing this issue is important ensuring balanced and fair model performance across all emotion classes.

3.2 Evaluation metrics

The proposed models were evaluated using categorical accuracy and categorical cross-entropy loss as the primary metrics. Categorical accuracy, ideal for multi-class classification tasks like facial emotion recognition, measures the proportion of correct predictions among all predictions. Categorical cross-entropy loss on the other hand helps us understand the difference between predicted and actual probability distributions, serving as a key guide during the optimization process.

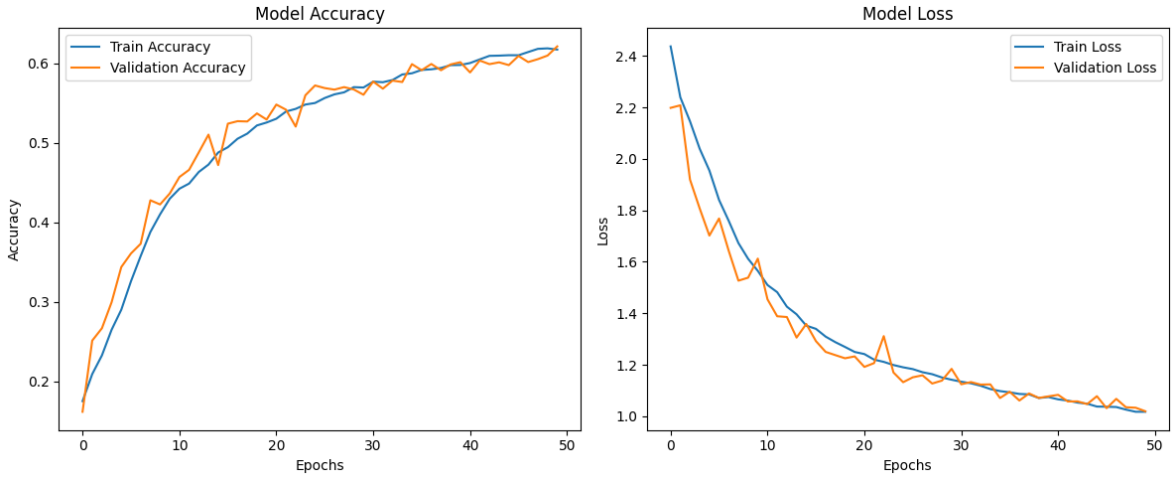

Figure 3: CNN Model Accuracy vs Loss Plot

Figure 3 shows us the train and validation accuracy and loss plots.The CNN model gave us a good starting point reaching a train accuracy of 61.9 and the validation accuracy of 62.14% with a relatively

simple architecture. This indicates that model can generalize well on unseen data.We can also observe the training and validation accuracy were constantly improved without the signs of overfitting and fluctuations. It's performance remained stable which makes it suitable choice for applications where computational efficiency and ease of implementation are important. However,the model struggled a bit in real-time applications as it had trouble handling the complexity of facial expressions that changes quickly.

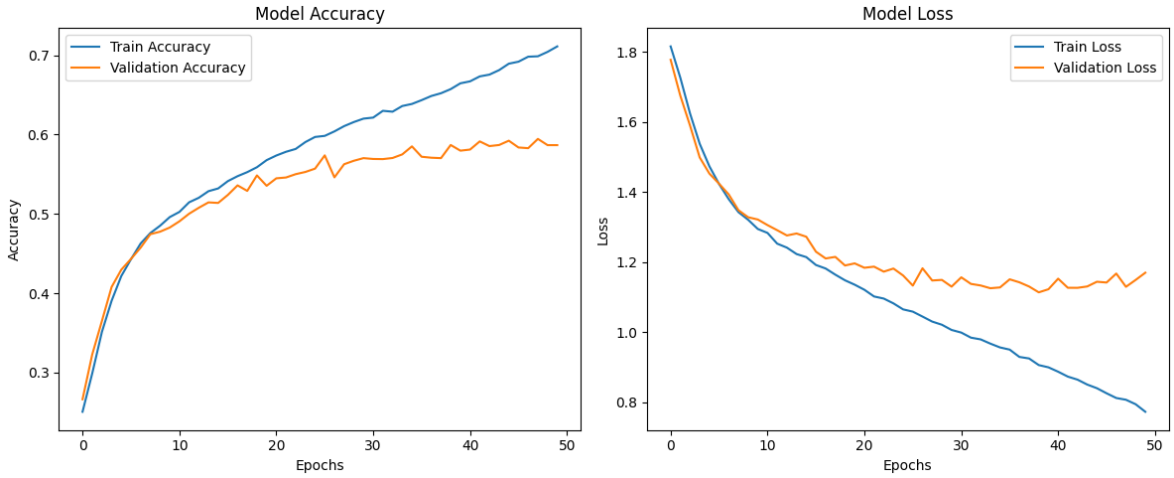

Figure 4: ResEmoteNet Model Accuracy vs Loss Plot

ResEmoteNet, with its deeper network architecture and use of residual connections showed stronger learning capabilities, achieving a training accuracy of 71.1%. Although its validation accuracy was a bit lower at 59.4% the model did show some signs of overfitting which suggests that further fine-tuning could help improve its ability to generalize and perform better on new data.However,it still outperformed the CNN model in real-time scenarios making it a better choice for applications requiring dynamic emotion detection.

The accuracy and loss plots for both the models showed clear difference in their learning patterns. The CNN model exhibited stable convergence with the training and validation metrics staying close to each other throughout the process indicating a balanced learning approach.On the other hand,ResEmoteNet showed a more noticeable gap between training and validation performance highlighting the challenges in maintaining generalization despite its higher training accuracy.

3.3 Results

The evaluation of the CNN and ResEmoteNet models on the FER2013 test dataset, consisting of 7,178 images across seven classes, revealed distinct performance trends.

The CNN model achieved a test accuracy of 62.6% and a test loss of 1.0056. This demonstrates that the model correctly predicted the emotions in about 62.6% of the test images. The test loss value of 1.0056 represents how far the model's predictions are from the actual emotions, with lower loss values indicating better accuracy and performance.Overall, the CNN model was pretty consistent across the seven emotion categories performing equally well with emotions like happiness, sadness, and anger. This makes it a solid baseline for facial emotion detection tasks.

When we looked at predictions on test images,it was clear the model could classify emotions accurately in most cases. However,the model struggled with subtle or mixed emotions, differentiating between anger and disgust or fear and surprise. Although the model worked fine with static images,it wasn't as effective in real-time situations. This shows that the CNN model is good for analyzing static images but it falls short when it comes to dynamic and fast-changing facial expressions in real-world settings.

On the other hand, ResEmotenet model achieved a test accuracy of 61.9% and a test loss of 1.0984 which is just a bit lower than the CNN model's performance.It performed almost similar to CNN when it comes to static images. However,because of its advanced architecture ResEmoteNet consistently performed better than CNN when it comes to detecting emotions in real-world scenarios such as live

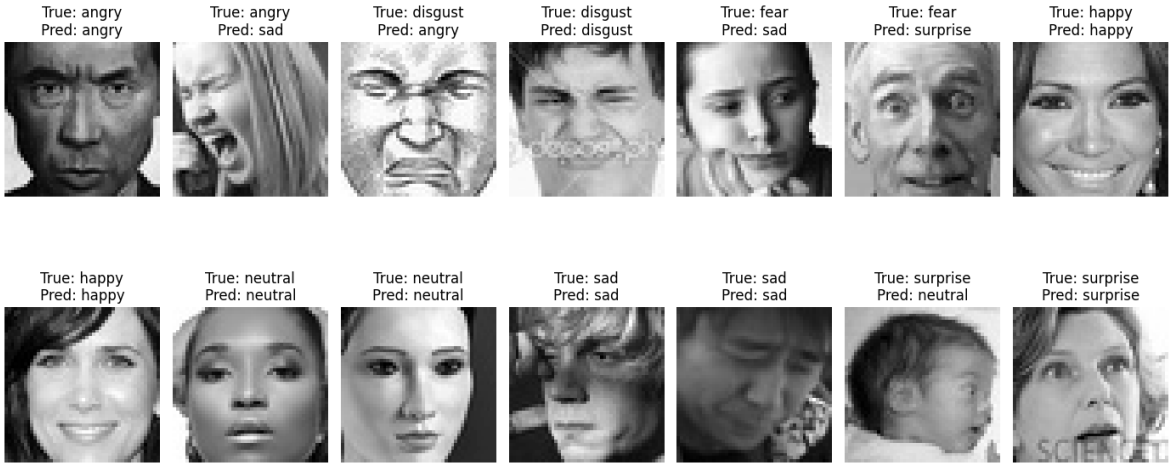

Figure 5: CNN Test Results

webcam streams.This indicates this model is better at adjusting to different situations, especially when things were more dynamic. Its deeper architecture helped it to identify subtle emotional details and maintain stable performance even under challenging conditions.

Figure 6: ResEmoteNet Test Results

The comparison between the CNN and ResEmoteNet models shows the balance between simplicity and flexibility in the model designs. To conclude,the CNN model was effective and provided consistent results on static test data which makes it a solid choice for simpler and more controlled tasks.However,it was less adaptable to real-world scenarios where emotions can be more subtle and fleeting emphasize the need for more advanced architectures like ResEmoteNet which can better handle practical implementations like real-time video or live webcam streams.This makes it a suitable choice for applications that require faster and accurate emotional recognition particularly when the data is more unpredictable and varied such as surveillance and health care.

3.4 Analysis and discussions

The project provides valuable insights into the strengths and limitations of models used for facial emotion recognition.

The CNN model which is best known for its simpler design established a solid baseline with training and test accuracies of 61.9% and 62.6%, respectively. Its consistent performance across training, validation, and test datasets shows that it effectively generalizes without significant overfitting.However,the

models performance accuracy in real-time based situations was somewhat limited particularly when interpreting dynamic facial expressions in video-based contexts. This limitation is likely due to the model's relatively shallow architecture which struggles to capture the intricate, complex patterns necessary for accurately recognizing subtle emotions.

In contrast, ResEmoteNet, inspired by the deeper pretrained ResNet architecture showed a stronger learning capacity with a training accuracy of 71.1%. By including residual connections and a more complex structure, it was better equipped to identify subtle patterns in facial expressions. Despite this, its validation and test accuracies were 59.4% and 61.9%, slightly trailing the CNN model. This indicates the potential overfitting where the model became overly specialized to the training data, leading to reduced generalization for unseen examples.

ResEmoteNet performed well in real-time applications due to its deeper architecture, which handled changes in facial expressions more effectively. This made it more reliable for live analysis, showing that a more complex model is important for recognizing subtle features. While the model might need some adjustments to reduce overfitting, its real-time performance is promising. It has great potential for practical use in areas like surveillance, healthcare, and human-computer interaction.

Model	Test Accuracy
CNN	62.6
ResEmoteNet	61.9

Table 1: Test Accuracy Results

The training and validation accuracy and loss curves of the model can be observed in Figure 3 and Figure 4. For CNN model,We can observe as the number of epochs increases both training and validation accuracy steadily improve, while the loss decreases, indicating that the model is learning effectively and generalizing well to unseen data.

For ResEmoteNet model,the training accuracy steadily increases, while the validation accuracy shows some fluctuations, suggesting the model learns but may struggle with generalization. Both training and validation loss decrease over time, indicating that the model is improving, though some instability is observed in validation performance.

Both models were trained under consistent hyperparameter settings to ensure a fair comparison with learning rate of 0.001 ,for 50 epochs and batch size of 64.The learning rate of 0.001 worked better in our case because it allowed the model to converge faster while showing minimal sign of overfitting.So,there is still chance for improvement indicating that further fine-tuning of the learning rate or other hyperparameters could lead to even better results.

Some of the challenges we faced in the project,the models were not able to detect the faces which has accessories such as spectacles,as they weren't trained on enough examples of such cases.Additionally, due to class imbalance in the dataset, the model struggled to detect emotions like disgust which had very few samples compared to other emotions.

These findings emphasize the trade-offs between simplicity and complexity in model design for emotion recognition, underscoring the need for careful selection and tuning based on specific application demands.

4 Conclusion

This project compared the performance of two deep learning models, CNN and ResEmoteNet, for Facial Emotion Recognition (FER) in both static images and real-time applications. Because of its simpler design, the CNN model was reliable, efficient and performed well on static images making it an excellent option for simple tasks with little resources. While it showed strong performance with static data, it had room for improvement in handling real-time emotion detection from video streams as it is better suited for less dynamic scenarios.

ResEmoteNet,on the other hand included advanced features like residual connections and Squeeze-Excitation mechanisms, which helped it focus on important facial details and identify emotions more accurately. It performed much better in real-time scenarios, especially with video data, and was able to detect subtle emotional changes effectively. However, it sometimes overfitted the training data, meaning its performance on new data was slightly lower, and it needs further adjustments to make it more reliable.

Overall, the project showed that while CNN is suitable for simpler tasks, ResEmoteNet is better for real-time applications but needs more fine-tuning to work consistently. In the future, improving ResEmoteNet's ability to handle different datasets and fixing issues like data imbalance could help make it more practical for real-world applications.

In this project, we have learned how important it is to choose the right model and also use proper preprocessing techniques to improve accuracy in facial emotion recognition. We have also observed how advanced features can help in real-time tasks but need careful adjustments to work well. Working as a team also taught us how collaboration makes solving complex problems easier.

5 Contribution

Our project was a collaborative effort, with all team members contributing equally and supporting each other as needed.

5.1 Code

In this project, my role included developing and implementing the preprocessing pipeline, training and evaluating the ResEmoteNet model, and performing in-depth data analysis. My responsibilities included normalizing and augmenting the FER2013 dataset to optimize training, fine-tuning the model's hyperparameters, applying early stopping to mitigate overfitting, and conducting rigorous evaluations on test data to assess model performance.

5.2 Report

Additionally, I contributed significantly to the project report. I summarized the project objectives in the Abstract and provided detailed explanations of data preprocessing and analysis in the Methods and Datasets sections. I also defined and described the Evaluation Metrics, ensuring the documentation clearly communicated our methodology and results.

References

[1] A. K. Roy, H. K. Kathania, A. Sharma, A. Dey, and M. S. A. Ansari, "Resemotenet: Bridging accuracy and loss reduction in facial emotion recognition," sep 2024. 1 Sep 2024.

[2] A. Vulpe-Grigoraşi and O. Grigore, "Convolutional neural network hyperparameters optimization for facial emotion recognition," 2021. 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE), 2021.

[3] N. T. Singh and Ritu, "Facial emotion detection using haar cascade and cnn algorithm," 2023. 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT).

[4] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual masking network," 2021. Proceedings of the International Conference on Pattern Recognition (ICPR).

[5] T. K. Kumar, "Artificial intelligence-based real-time facial emotion monitoring system," 2023. Proceedings of the 2023 9th International Conference on Computer and Communication Engineering (ICCCE).

[1]. [2] [3] [4] [5]