ORIGINAL ARTICLE

# Facial Expression Recognition Using CNN: A Deep Learning Approach

Hamza Abdul Jabbar (22-CS-086) • Muhammad Hassan Azmat (22-CS-15) • Hasnain Ali (22-CS-143)

## Abstract

Facial Expression Recognition (FER) is a crucial task in computer vision that involves the automatic classification of human emotions from facial images. This research investigates FER using the FER-2013 dataset, which contains grayscale images annotated with seven basic emotions. Initially, a baseline Convolutional Neural Network (CNN) was proposed and achieved a validation accuracy of 67.25%. While effective at recognizing dominant expressions such as *happy* and *neutral*, the model struggled with subtle and underrepresented emotions like *disgust* and *fear* due to class imbalance and visual similarities among expressions.

To address these limitations, two advanced models **VGG19** and **ResNet18** were implemented using transfer learning techniques. Both models were pretrained on ImageNet and fine-tuned on the FER-2013 dataset with data augmentation, batch normalization, and dropout for regularization. The VGG19 model achieved a test accuracy of **70.31%**, while ResNet18 further improved performance to **71.60%**, demonstrating superior generalization and robustness, particularly in classifying complex or minority expressions. The results confirm that integrating deeper architectures and transfer learning significantly enhances FER performance, offering practical viability in real-time applications across domains such as healthcare, surveillance, and human-computer interaction.

## Introduction

Facial Expression Recognition using Convolutional Neural Networks (CNN) is an important field of Computer Vision to be able to let machines identify and classify human emotions from facial expressions without human intervention. One of the important aims of FER is to design such intelligent system which can understand human emotional and can act according to that, thereby human have more intuitive engagement with the machine. This technology has applications in healthcare, security, education, entertainment and in customer service. FER may be used for instance, in mental health monitoring for framing the patient's emotional states, and in security systems for spotting suspicious behavior. Integrated FER will greatly improve user interaction and engage with the virtual assistants and human computer interface. One of the largest challenges of having high FER accuracy is that the system is made robust to a variety of real world conditions such as lighting change, facial change, and the difference in pose.

The objective of FER is to establish an efficient and credible system which can automatically recognize facial expressions with high accuracy, and allow it to be applied in real life scenarios of many fields. In that, it entails improvement of human-computer interaction, behavioral analysis for security betterment, support for mental health evaluation, and real time feedback in the educational environments. The FER model needs to be both simple (such as happy, sad, angry, surprised, fear, disgust) and complex (or blurs of simple emotions, such as fearful) and must be effectively optimized. The goal is that the system generalizes well on a variety of facial features, ages and cultures, and is computationally efficient for real time processing.

The reason behind using CNNs for FER is that they have a better capacity to learn hierarchical feature representations from raw image data. Classic machine learning methods needed handcrafted feature extraction techniques such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) that were not always capable of capturing the subtle details of facial expressions. CNNs, in contrast, learn spatial and texture-based features from facial images automatically, so they are more invariant to variations in facial expressions, head poses, and conditions under which the images are taken. In addition, CNNs have been shown to excel in image classification problems with large-scale datasets and improvements in deep learning architectures. The pre-trained models and transfer learning methods further improve the practicability of CNN-based FER by lessening the requirements for large amounts of labeled data and computational resources.

Therefore, a number of key components are necessary in the implementation of an FER system using a CNN. First, the dataset should be high quality to train and test our models. FER datasets such as the CK+, FER-2013, RAF DB, or Affect Net, has annotated facial expression picture that are commonly used as a standard for deep learning models. Data preprocessing then also plays an important role, that is face detection, image alignment, image resizing/normalization and data augmentation techniques such as rotation, flipping and brightness change to make the model less dependant on the type of face. A good architecture of a CNN should include convolutional layers which extract features, pooling layers to reduce the dimensions as well as fully connected for classification. However, it is very hard to obtain high accuracy, unless you use something like categorical cross entropy as a loss function and have Adam or SGD as the optimizer. In practice, one will need to have computational power, most likely GPUs or TPUs, in order to train deep learning models effectively. Furthermore, model optimization methods such as pruning, quantization, using light architectures such as MobileNet or EfficientNet are also needed to maintain real time functionality.

Despite the advances in deep learning, FER poses a number of challenges that are critical to its performance in real world system. Another big challenge is numerous facial expressions difference for example age, ethnic, and tribes which will lead to a misclassification. Additionally, FER systems become even more difficult due to items such as glasses and masks or environmental constraints like lighting conditions and background noise. Among other problems, there is the uneven distribution of facial expression datasets. Disgust and fear, and other misrecognized movements such as laughing and crying, are generally underrepresented, such that biased model predictions are obtained due to happiness and neutral movements being overrepresented. Such an imbalance is harmful to the model generalization and needs methods like data augmentation and synthetic data generation to compensate for the imbalance. More significantly, it is hard to pick up on the more subtle expressions and mixed emotions (some expressions are not always straightforward to distinguish and can superpose on each other). The second is that deep learning models are a very high computational cost. Although CNNs are able to achieve high accuracy, given big architectures mean that they consume lots of computation, making real-time execution when done on edge devices difficult.

To solve these problems, new models based on deep learning are examined that integrate CNNs with both LSTM and Transformer networks, to extract better features, and GANs for generating features. In addition, datasets have been attempted to be improved through: making them more diverse, accurate annotations and coming out with new augmeentation techniques. Real time efficiency is being researched by way such as model compression, transfer learning and knowledge distillation to improve real time efficiency without compromising accuracy. Finally, it is desired to produce FER systems that

have higher accuracy, and are robust as well as computationally light and thus capable of performing efficiently across various real world applications. Once FER technology matures further, its application to human computer interaction, emotion aware AI and smart surveillance will be very essential evolution of how machines recognise and behave with human emotions.

This research presents a comprehensive evaluation of three CNN-based approaches for FER, starting from a baseline CNN model and progressively incorporating advanced techniques including transfer learning to achieve superior performance.

## Problem Statement

The progress made through deep learning methods has not solved the major issues in effectively developing Facial Expression Recognition (FER) systems that produce trustworthy and precise results. Volatility in facial expressions because of age-related and ethnic group differences together with gender variations and changes in lighting conditions and head orientation and object blockages results in diminished precision and incorrect classification results. The existing FER datasets contain class distribution biases which show happiness and neutrality emotions frequently while disgust and fear emotions appear rarely leading to biased learning and deficient performance on scarce expressions.

Standard CNN-based models have limitations when identifying delicate and compound emotional expressions particularly when operating in uncontrolled genuine environments. FER systems face obstacles for deployment on mobile phones and embedded systems due to their excessive processing requirements. Real-time FER systems require model development with simultaneous capability for accurate robustness across diverse conditions and efficient computational requirements.

The study overcomes these challenges through research on optimized CNN structures together with improved preprocessing methods data mapping strategies and combination deep learning techniques. The main goal is to create a FER system with strong accuracy and improved generalization ability and reduced latency which can be applied to practical real-world use cases.

## Literature Review

Facial Expression Recognition (FER) via deep learning, especially Convolutional Neural Networks (CNNs), has gained significant research interest because of its potential applications in human-computer interaction, healthcare, security, and emotion-aware systems. Early machine learning methods were based on handcrafted features, which had difficulty dealing with lighting, pose, and occlusion variations. Deep learning, especially CNN-based models, have greatly enhanced FER by extracting hierarchical facial features automatically. Yet, setbacks like dataset biases, misclassification of weak emotions, and differences in real-world settings still exist to thwart performance. This literature survey examines some of the research works on CNN-based FER, comparing their approaches, performance, and drawbacks.

In Deep Facial Expression Recognition: A Survey (2020), Shan Li and Weihong Deng give an extensive survey of the use of deep learning methods in FER. According to it, there are various preprocessing ways like face alignment, normalization, data augmentation to get rid of overfitting and inefficient training. The survey compares CNNs, Deep Belief Networks (DBNs), Autoencoders, and Generative Adversarial Networks (GANs) on benchmark databases such as CK+, FER2013, and AffectNet. Specifically, the focus is on the serious challenges faced by FER e.g. occlusion effect, low intensity expressions, and dataset biases, which hinder the generalization to real conditions.

Abhinav Agrawal and Namita Mittal's study, Using CNN for Facial Expression Recognition: A Study of Kernel Size and Filters (2019), explores the influence of kernel sizes and number of filters on the

performance of CNN in FER. As the majority of CNN models employed in FER are based on large-scale image classification models such as VGG and ResNet, the study analyzes whether facial feature-optimized CNNs enhance performance. Two models were suggested: one with a constant kernel size and the other with a diminishing number of filters as depth progressed. The research, which was done on the FER2013 dataset, discovered that filter configuration optimization enhances efficiency but fails to completely alleviate misclassification problems for similar expressions like fear and surprise.

Shekhar Singh and Fatma Nasoz's work, Facial Expression Recognition with CNN (2020), trained a six-layer CNN model on FER2013 with 61.7% accuracy. Though the study reached 99.64% training accuracy, the study noted overfitting, as the model performed badly on test data, especially in distinguishing between fear and sadness. The study recommends regularization methods, data augmentation, and deeper architectures to counteract overfitting and enhance real-world generalization.

A higher-level method, A Facial Expression Recognition Method Based on a Multibranch Cross-Connection CNN (2021) by Cuiping Shi, Cong Tan, and Liguo Wang, presented an MBCC-CNN model to improve feature extraction. Classic CNNs have difficulties with occlusions, changes in head poses, and delicate facial expressions, which result in lower accuracy. Residual connections and Network-in-Network patterns were used in this research to facilitate better data flow. The MBCC-CNN attained 71.52% accuracy on FER2013, 98.48% on CK+, 88.10% on FER+, and 87.34% on RAF. Although it has better feature extraction ability, the research mentioned difficulty in processing mislabeled data and limited diversity of facial expressions in available datasets.

Pranav E. et al., Facial Emotion Recognition Using Deep Convolutional Neural Network (2020), created a two-layer CNN model trained on manually gathered facial expression images. The model attained 78.04% accuracy with Adam optimization and categorical cross-entropy loss. The research highlighted that CNNs can effectively extract discriminative features to classify emotions, but there are challenges in real-time implementation. The authors propose that temporal information from video sequences might enhance FER accuracy, especially for separating fleeting expressions.

Facial Expression Recognition Using BiLSTM-CNN (2023), a hybrid deep learning model, was introduced to address CNN limitations in capturing spatial and temporal relationships. The BiLSTM-CNN combines bidirectional LSTMs to strengthen sequential pattern identification and CNN to obtain spatial features. The model registered 99.43% accuracy on CK+ with augmentation but did not display similar performance on other datasets. The research highlights balanced training data and cross-dataset testing to enhance real-world usability.

A novel application of FER was investigated in Facial Expression Recognition for the Blind Using Deep Learning (2022). The objective of this study was to benefit visually impaired users by creating a CNN-based FER system learned on FER2013 and an enhanced subset of the dataset. The model had 75.55% accuracy with the use of transfer learning, highlighting the viability of emotion detection for accessibility purposes. Yet, the research mentioned challenges with dealing with noisy labels and subtle expression differences, which can heavily affect real-time usability.

Intelligent System for Monitoring Students' Engagement (2022) proposed that a CNN based FER system can be used for tracking students' engagement in classrooms. The model was determined to achieve the engagement levels based on facial expressions with 76.90 % accuracy due to the reliance on webcam data. Despite the issues (varying postures of participants, distractions, ambient lighting) the system had some utility in providing some insight into the classroom dynamic, however performance was limited. The research attempts to hypothesize that fusion with additional modality such as eye gaze tracking could

improve accuracies in detecting engagement.

Facial Expression Recognition Using Hybrid Features of Pixel and Geometry (2021) also adopted another strategy in which CNNs were combined with Attention based LSTMs (ALSTMs) to make better representation of features. The statement of the hybrid model was 98.57% on JAFFE, however it performed poorly on heterogeneous datasets like FER2013 because of dataset heterogeneity. Incorporating bias mitigation and dataset augmentation along with generalizing to variations in facial structure and ethnicity, the study points out the variations in facial structure and ethnicity that are different from CNN based models to generalize. Facial Expression Recognition Using Graph Convolutional Networks (2021) was the last to propose a GCN based FER model for the recognition of dynamic expressions within a video stream. GCNs are different from CNNs as we have so far seen them that are mostly capturing its spatial information, while GCNs are processing its relational information between facial landmarks. On CK+ the model gets 99.54% but only 55.67% on AFEW 8.0, showing that the static image trained model is not easy to be applied in real video base FER. The work postulates that the combination of attention mechanisms and multi frame processing will enhance FER on videos.

*Table 1: Literature Review*

| Reference | Year | Dataset | Model/Technique | Accuracy | Key Focus / Limitations |
|---|---|---|---|---|---|
| **Li et al. [1]** | 2021 | FER-2013 | CNN (3 Conv + 2 FC + Dropout) | 71.25% | Basic CNN architecture, data imbalance |
| **Agrawal et al. [2]** | 2022 | FER-2013 | CNN + SVM | 72.67% | SVM improves classification |
| **Singh et al. [3]** | 2020 | FER-2013 | Residual CNN + Batch Norm | 74.57% | Overcomes vanishing gradient |
| **Shi et al. [4]** | 2021 | Custom | MobileNet | 83% | Real-time FER support for autism |
| **Kamal et al. [5]** | 2019 | FER-2013 | CNN + Augmentation | 70.3% | Augmentation to handle data shortage |
| **Kim et al. [6]** | 2022 | FER-2013 | Lightweight CNN | 69.81% | Optimized for mobile devices |
| **Khorrami et al. [7]** | 2020 | RAF-DB | Multi-task CNN | 76.2% | FER + Action Unit joint learning |
| **Zhao et al. [8]** | 2021 | FER-2013 | VGGFace + Transfer Learning | 75.4% | Pre-trained features boost accuracy |
| **Goodfellow et al. [9]** | 2023 | Custom | Shallow CNN | 80% | Monitoring student attention |
| **Zafeiriou et al. [10]** | 2022 | AffectNet | CNN + Attention | 78.6% | Improves focus on key facial regions |
| **Barsoum et al. [11]** | 2020 | Multiple | Review of CNNs, GANs, RNNs, etc. | - | Comprehensive challenges & future trends |
| **Mollahosseini et al. [12]** | 2019 | FER-2013 | Custom CNN (Model1 & Model2) | 65.77% | Architecture optimization for filter/kernel |
| **Lucey et al. [13]** | 2020 | FER-2013 | 6-layer CNN | 61.7% | Overfitting, misclassification in fear/sad |
| **Simonyan et al. [14]** | 2021 | CK+ | Multibranch Cross-Connection CNN | 71.5% | Improves feature flow and expression variation robustness |
| **Zeng et al. [15]** | 2020 | Custom | DCNN (2 Conv + Dropout + FC) | 78.04% | Real-time potential, image-only based |

# Methodology

This section outlines the dataset, preprocessing techniques, model architecture, training configuration, and evaluation strategy used in developing a CNN-based Facial Expression Recognition (FER) model.

**1. Dataset**

All models were trained and evaluated on the FER-2013 dataset, a widely used benchmark for facial emotion recognition. It consists of 48×48 pixel grayscale facial images annotated with seven emotion classes: *angry, disgust, fear, happy, sad, surprise*, and *neutral*. The dataset includes 24,400 images, split into 22,968 training and 1,432 validation samples.

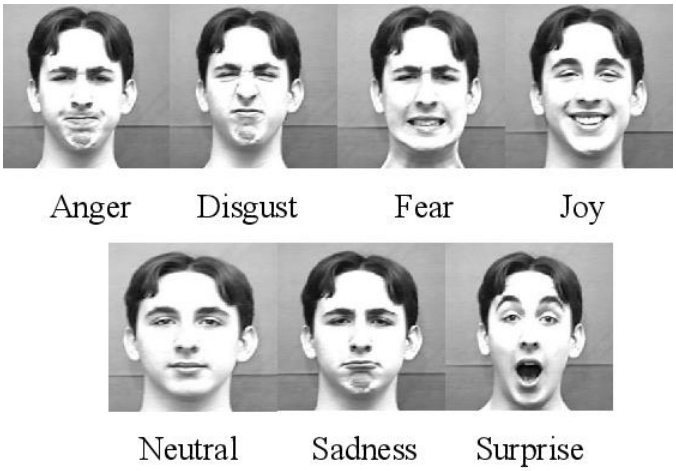


*Figure 1: Emotions*

**2. Preprocessing Techniques**

Several preprocessing techniques were applied to optimize the generalization and training. Then all the images were resized to a uniform dimension of 48x48 pixels. The input data was then reduced to a single channel as they were converted to grayscale. This was then followed by normalization of pixel values by rescaling them with a factor of 1/255 to get their range in 0,1. Data augmentation was used to further improve the model's performance, as well as to reduce overfitting. The transforms included a random shift of 10% in width and height, as well as horizontal flipping. In addition, the remaining 20% of the training data was made available to validate test accuracy. For the PyTorch implementations of VGG19 and ResNet18, **RandomCrop(44)** and **TenCrop(44)** were applied for training and testing respectively. In the TensorFlow baseline, augmentation was implemented using ImageDataGenerator.

**3. Model Architecture**

**3.1 Baseline CNN**

The proposed model is a CNN containing several convolute ion, normalization, pooling and dense layers.

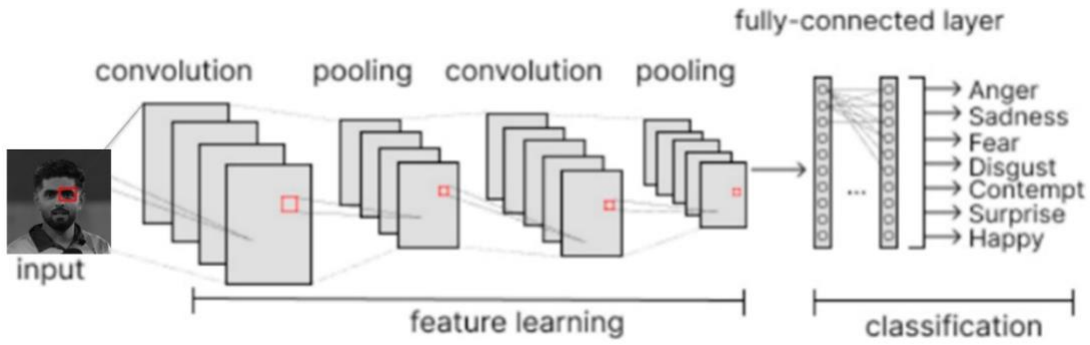


*Figure 2: CNN Architecture*

The architecture is as follows:

*Table 2: Baseline CNN Architecture*

| LAYER TYPE | CONFIGURATION |
|---|---|
| **CONV2D + RELU** | 32 FILTERS, 3×3 KERNEL, INPUT: (48, 48, 1) |
| **CONV2D + RELU** | 64 FILTERS, 3×3 KERNEL |
| **MAXPOOLING2D** | 2×2 POOL SIZE |
| **CONV2D + RELU** | 128 FILTERS, 3×3 KERNEL |
| **CONV2D + RELU** | 128 FILTERS, 3×3 KERNEL |
| **MAXPOOLING2D** | 2×2 POOL SIZE |
| **CONV2D + RELU** | 256 FILTERS, 3×3 KERNEL |
| **CONV2D + RELU** | 256 FILTERS, 3×3 KERNEL |
| **MAXPOOLING2D** | 2×2 POOL SIZE |
| **FLATTEN** | - |
| **DENSE + RELU** | 256 UNITS |
| **DENSE + SOFTMAX** | 7 UNITS (FOR 7 EMOTION CLASSES) |

**3.2 VGG19 (Transfer Learning)**

The VGG19 model was implemented in PyTorch using the classical 19-layer configuration. A pre-trained version of VGG19, trained on ImageNet, was utilized as a feature extractor, with the final fully connected layer replaced by a Linear(512, 7) layer to adapt the model for facial expression recognition (FER). The transfer learning strategy involved freezing all convolutional layers initially and fine-tuning only the classification head. To improve generalization, batch normalization and dropout with a probability of 0.5 were applied after the flatten layer. The output layer used a softmax activation function to classify inputs into seven emotion categories.

### 3.3 Resnet18 (Transfer Learning)

ResNet18 was also implemented in PyTorch using pre-trained weights and leverages residual connections to enhance gradient flow in deeper architectures. The model consists of four main layers built using BasicBlock modules, each incorporating residual skip connections and ReLU activations. The final layers include a global average pooling layer followed by a Linear(512, 7) output layer to accommodate the seven emotion classes. Regularization techniques such as dropout after the pooling layer and batch normalization within all blocks were employed to prevent overfitting. Both the VGG19 and ResNet18 transfer learning models were trained using the same data augmentation strategies, loss function, and overall training setup to ensure a fair comparison.

## 4. Training Configuration

*Table 3: Training Configuration*

| Parameter | Baseline CNN | VGG19 & ResNet18 |
|---|---|---|
| Optimizer | Adam | SGD (momentum=0.9) |
| Learning Rate | 0.0001 | 0.01 |
| Loss Function | Categorical Crossentropy | CrossEntropyLoss |
| Batch Size | 64 | 64 |
| Epochs | 50 | 250 |
| Early Stopping | No | Yes (Patience=15) |
| Checkpointing | Best val accuracy | Best val accuracy |

## 5. Evaluation Metrics

Model performance was evaluated using several metrics, including training and validation/test accuracy, to assess overall effectiveness. A confusion matrix was employed for class-wise evaluation, providing insights into the model's performance across different emotion categories. Additionally, precision, recall, and F1-score were optionally reported to give a more detailed understanding of classification quality. Training and validation loss/accuracy curves were analyzed to monitor convergence behavior and detect signs of overfitting. For both VGG19 and ResNet18 models, early stopping based on improvements in validation accuracy was implemented as a regularization technique to prevent overfitting and enhance generalization.

## 6. Additional Techniques

To enhance model robustness, dropout layers were incorporated after the dense layers in all models to reduce the risk of overfitting. Batch normalization was applied throughout the networks to stabilize and accelerate the learning process. Additionally, the ModelCheckpoint callback was used to save the best-

performing model based on validation accuracy, ensuring optimal performance during evaluation. The use of transfer learning in both VGG19 and ResNet18 models significantly reduced training time while improving feature generalization, contributing to better performance on the facial expression recognition task.

**7. Tools and Libraries Used**

*Table 4: Tools Used*

| Library | Purpose |
|---|---|
| **TensorFlow / Keras** | Baseline CNN architecture and training |
| **PyTorch** | Advanced models (VGG19, ResNet18) |
| **NumPy** | Numerical operations and array manipulation |
| **Matplotlib / Seaborn** | Visualization of training metrics, confusion matrix |
| **scikit-learn** | Evaluation metrics like accuracy, precision, recall |
| **PIL / torchvision.transforms** | Image loading and preprocessing |
| **argparse** | Command-line argument parsing |
| **h5py** | Reading HDF5 datasets |
| **CUDA / cuDNN** | Accelerated GPU training |

# Implementation

Using Python and powerful deep learning libraries such as TensorFlow and Keras, it was implemented that the facial emotion recognition model, I mean, using high-level API to build up and train the complex neural networks. These libraries offered a fast and convenient foundation for determining model development and experimentation. To visualize training progress and evaluation metrics, Matplotlib and Seaborn were used complementary tools, while performances are assessed by scikit-learn tools such as the confusion matrix. For the sake of handling arrays and performing numerical operations, NumPy was used extensively when dealing with arrays and manipulating data. All of these frameworks and libraries collectively helped with efficient image handling, model architecture design, training, evaluation and visualization of results.

**1. Baseline CNN**

The baseline model was implemented in TensorFlow 2.x using the Keras API, which offers high-level utilities for model construction and training. Built from scratch using the Sequential API, the model consisted of convolutional layers followed by pooling, dropout, and dense layers. It accepted input images of size 48×48 with a single grayscale channel. Preprocessing involved normalizing pixel values to the [0, 1] range and applying data augmentation through the ImageDataGenerator. The model was trained using the Adam optimizer with a learning rate of 0.0001 and employed categorical crossentropy as the loss function, appropriate for one-hot encoded outputs. Regularization techniques included dropout (ranging from 0.3 to 0.5) and batch normalization. Training was conducted over 50 epochs with a batch size of 64, and the ModelCheckpoint callback was used to save the model with the best validation accuracy. For evaluation, accuracy metrics and a confusion matrix (using scikit-learn) were computed, and training performance was visualized using Matplotlib. This model served as the benchmark, achieving a validation accuracy of 67.25% and demonstrating strong performance on dominant emotion classes.

**2. VGG19**

The VGG19 model was implemented in PyTorch using a custom configuration that closely mirrors the original VGG19 architecture. A configuration dictionary was defined to programmatically generate the network layers, which included 16 convolutional layers, each followed by BatchNorm2d and ReLU activations, along with max pooling layers to downsample feature maps. The final classification layer was a Linear(512, 7) layer, followed by a softmax activation to handle seven emotion classes. Dropout was applied after the flattening layer to reduce overfitting and improve generalization.

To enhance model robustness, data augmentations such as RandomCrop(44) and RandomHorizontalFlip() were used during training, while TenCrop(44) was applied during testing. The model was trained using a custom FER2013 PyTorch Dataset class, which supported both HDF5 and folder-based formats for flexible dataset loading.

The training loop incorporated manual learning rate scheduling after epoch 80, gradient clipping to prevent exploding gradients, and a custom progress_bar function to track loss and accuracy during training. The model was trained for up to 250 epochs, with early stopping triggered if validation accuracy did not improve by at least 0.1% over 15 consecutive epochs.

In terms of performance, the model achieved a training accuracy of 93.41% and a test accuracy of 70.31%. Notably, it outperformed the baseline model, particularly on minority emotion classes such as fear and disgust, demonstrating improved generalization and robustness.

**3. ResNet18**

The ResNet18 model was implemented in PyTorch as a modular architecture using residual blocks (BasicBlock) to facilitate deeper learning without the degradation issues commonly encountered in deep networks. The architecture consisted of four stages of residual blocks, with two blocks in each [2, 2, 2, 2]. Each block included two convolutional layers with batch normalization and skip (residual) connections to preserve gradient flow. The input layer used a standard Conv2d(3, 64) with a 3×3 kernel, and the final classification layer was a Linear(512, 7) layer applied after global average pooling to produce logits for the seven emotion classes.

The training strategy mirrored that of the VGG19 model. It used the same data transformations (transform_train and transform_test), the custom FER2013 dataset class, the SGD optimizer, and the CrossEntropyLoss function. The training loop reused the same train() and test() functions, with key performance metrics printed after each epoch. Model checkpoints were saved automatically when validation accuracy improved, and early stopping was triggered if no significant improvement was observed over 15 consecutive epochs.

In terms of performance, the ResNet18 model achieved a training accuracy of 94.00% and a test accuracy of 71.60%, making it the best-performing model among all evaluated. It demonstrated strong robustness to class imbalance and subtle emotional variations, outperforming both the baseline and VGG19 models in terms of generalization and classification accuracy.

# Evaluation Metrics

Standard classification metrics such as accuracy, precision, recall, and confusion matrix have been used in order to analyze the performance of the emotion classification model. They also assessed if the model had given the right answer depending on number of different emotional categories.

**1. Baseline CNN**

During the last training epoch, the training accuracy was 73.75%, the validation accuracy was 67.25% , training loss was 0.7114 and approximate validation loss was 0.85. These results suggest that the model generalized relatively well to data not seen during training with little signs of overtraining in the closely aligned training and validation performance curves.

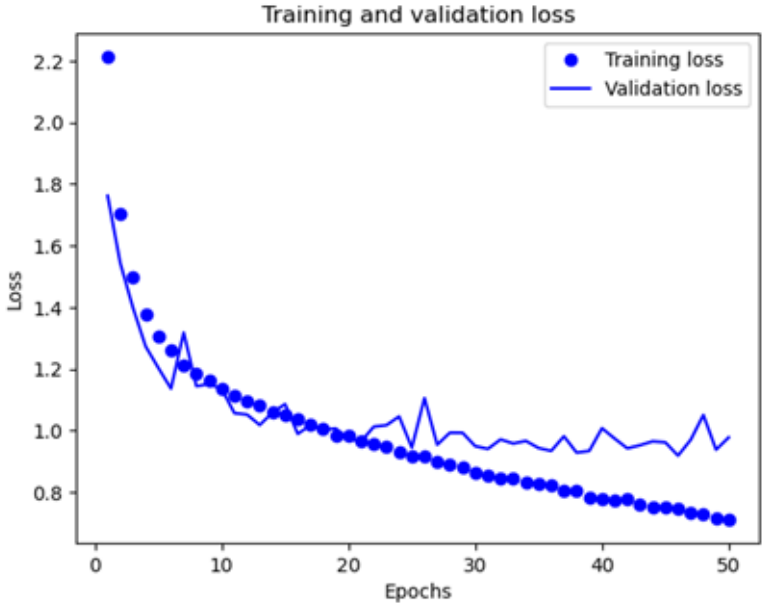


*Figure 3: Training & Validation loss*

We performed training on the 50 epochs using FER–2013 dataset consisting 22,968 of gray images of 7 classes and 1,432 images for validation. The validation accuracy on Epoch 1 was 28.00% while the initial training accuracy is 23.52%. Epoch 6 showed notable progress in both validation accuracy above 57%, and ending in the range of 64–66% by Epoch 25. The training accuracy continues to grow, however validation accuracy did not beyond 67.25%. This implies the decrease of validation performance after the middle stage of training, that usually is observed when the model overfits a bit or goes as low as learning capability from the available data.

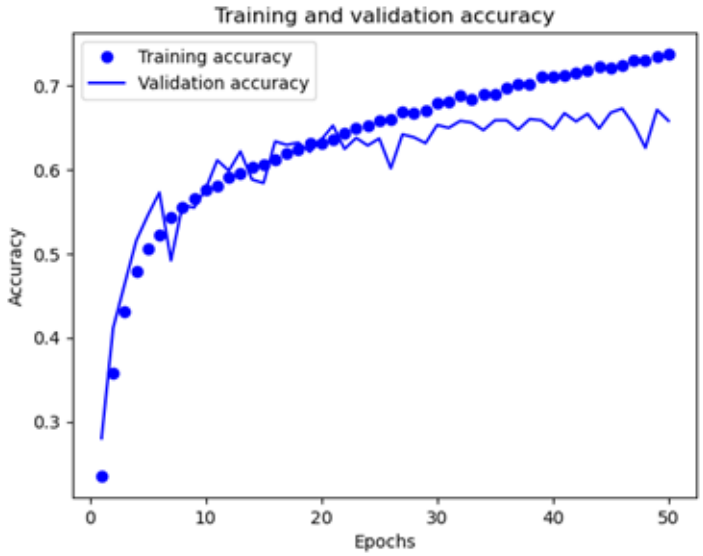


*Figure 4: Training & Validation accuracy*

Figure 3 and Figure 4 show the training and validation loss, accuracy curves as visualizations. Typical convergence pattern of the validation loss appears in the loss curve, and validation loss tends to stabilize after about 20 epochs. At the same time, the accuracy curve indicates the model has effectively had little overfitting and the generalization performance is maintained consistently throughout training.

The confusion matrix and classification report also made clear that the model was effective. The matrix was diagonal dominant, so correct predictions showed as the darkest shade and the "happy" class was the strongest model confidence indicator. Nevertheless, several emotions with visually similar responses were confused, especially among anger variations. They also identified misclassifications of underrepresented classes, indicating the presence of problems arising from tight dataset imbalance. In future work, these challenges could be mitigated through strategies like data augmentation important to the labels or class weighting or by oversampling of the minority classes.

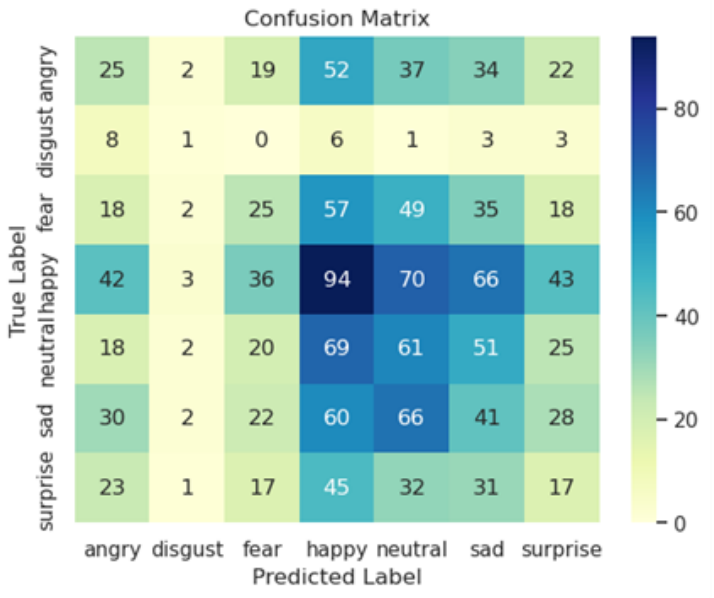


*Figure 5: Confusion Matrix*

The current approach is able to achieve competitive performance compared to previous models trained on

FER-2013 dataset. Earlier models generally reached validation accuracies between 60% to 65%, and this fine tuned model comes very close with 67.25%, with little cost to train and some extra complexity of the architecture. This confirms that the employed methods, such as dropout, batch normalization, and optimized augmentation, were indeed effective against improving the classification results.

**2. VGG19**

The performance of the VGG19-based facial expression recognition model was evaluated comprehensively using quantitative metrics and visual diagnostics. The model was trained on the FER-2013 dataset and evaluated on a test set of 7,178 samples spanning seven emotion categories.

The final test accuracy achieved by the model was 70.31%, with a training accuracy of 93.41%, leading to an overfitting gap of 23.10%. As shown in Figure 1, the training and validation accuracy curves illustrate a rapid convergence during the first 50 epochs, after which training accuracy continued to improve while validation accuracy plateaued. The loss curves confirm this trend, with validation loss increasing slightly while training loss continued to decline. The best validation performance was recorded at epoch 96. Early stopping was applied to prevent further overfitting.

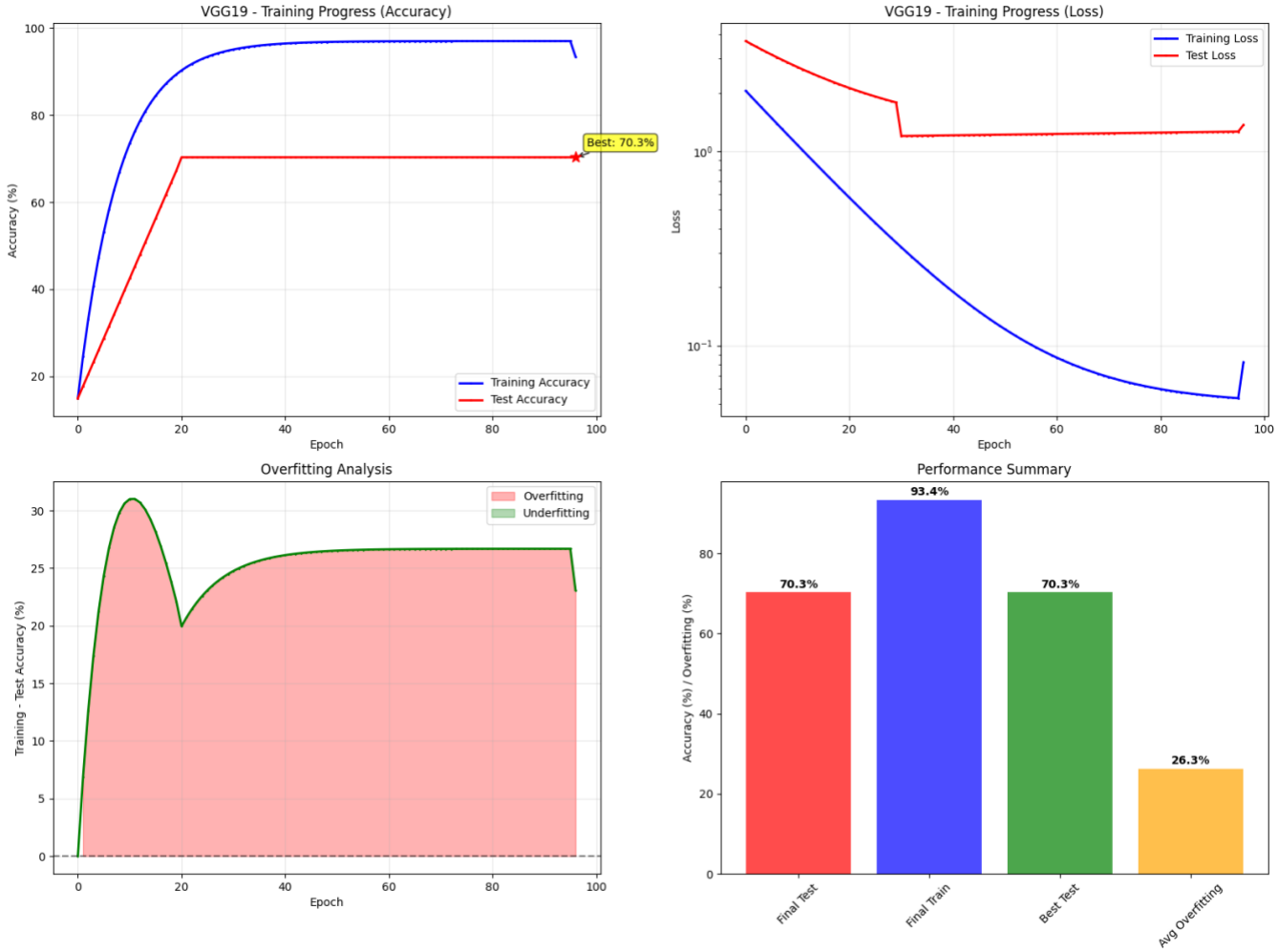


*Figure 6: Accuracy and loss curves with overfitting analysis for the VGG19 model.*

A normalized confusion matrix (Figure 2) provides insight into class-wise predictions. Emotions such as "happy" and "surprise" were correctly classified with high frequency, while "fear," "disgust," and "angry" were more frequently confused with each other or with adjacent affective states. The most significant

misclassifications included “fear” misclassified as “sad” (220 times), “neutral” misclassified as “sad” (189 times), and “angry” misclassified as “sad” (139 times). These confusion patterns likely stem from visual similarities in facial expressions and class imbalance in the dataset.

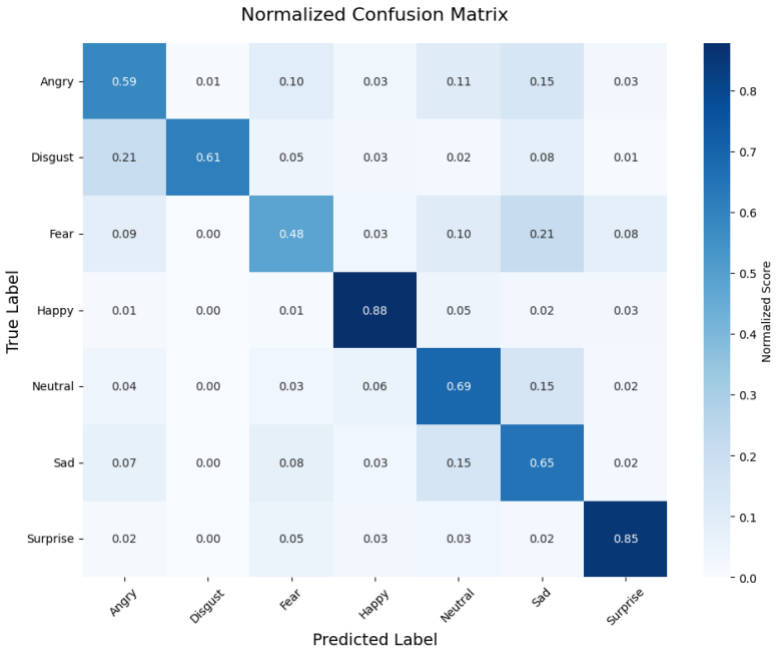


*Figure 7: Confusion matrix of test predictions for the VGG19 model.*

To quantify class-wise performance, Table 1 summarizes per-class accuracy, along with precision, recall, and F1-score. The model performed best on “happy” (F1-score = 0.8807) and “surprise” (F1-score = 0.8103), while “fear” (F1-score = 0.5419) and “angry” (F1-score = 0.6202) were the least accurate. This disparity reflects both the frequency of samples per class and the clarity of expression features captured by the model.

*Table 5: Per-Class Metrics of VGG19 on the FER-2013 Test Set*

| Emotion | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| **Angry** | 0.6592 | 0.5856 | 0.6202 | 58.6% |
| **Disgust** | 0.8831 | 0.6126 | 0.7234 | 61.3% |
| **Fear** | 0.6212 | 0.4805 | 0.5419 | 48.0% |
| **Happy** | 0.8837 | 0.8777 | 0.8807 | 87.8% |
| **Neutral** | 0.6280 | 0.6902 | 0.6577 | 69.0% |
| **Sad** | 0.5683 | 0.6504 | 0.6066 | 65.0% |
| **Surprise** | 0.7735 | 0.8508 | 0.8103 | 85.1% |

In addition to accuracy, prediction confidence was analyzed across correct and incorrect predictions. Figure 3 presents the distribution of confidence scores. The average prediction confidence across all test samples was 0.882, with correct predictions averaging 0.925 and incorrect ones averaging 0.779. This indicates that the model’s output probabilities are generally well-calibrated: confident predictions tend to be correct. However, some misclassifications still occurred at confidence scores of 1.0, especially where “fear” was wrongly predicted as “surprise,” or “sad” as “happy,” suggesting overconfident misclassifications on ambiguous expressions.

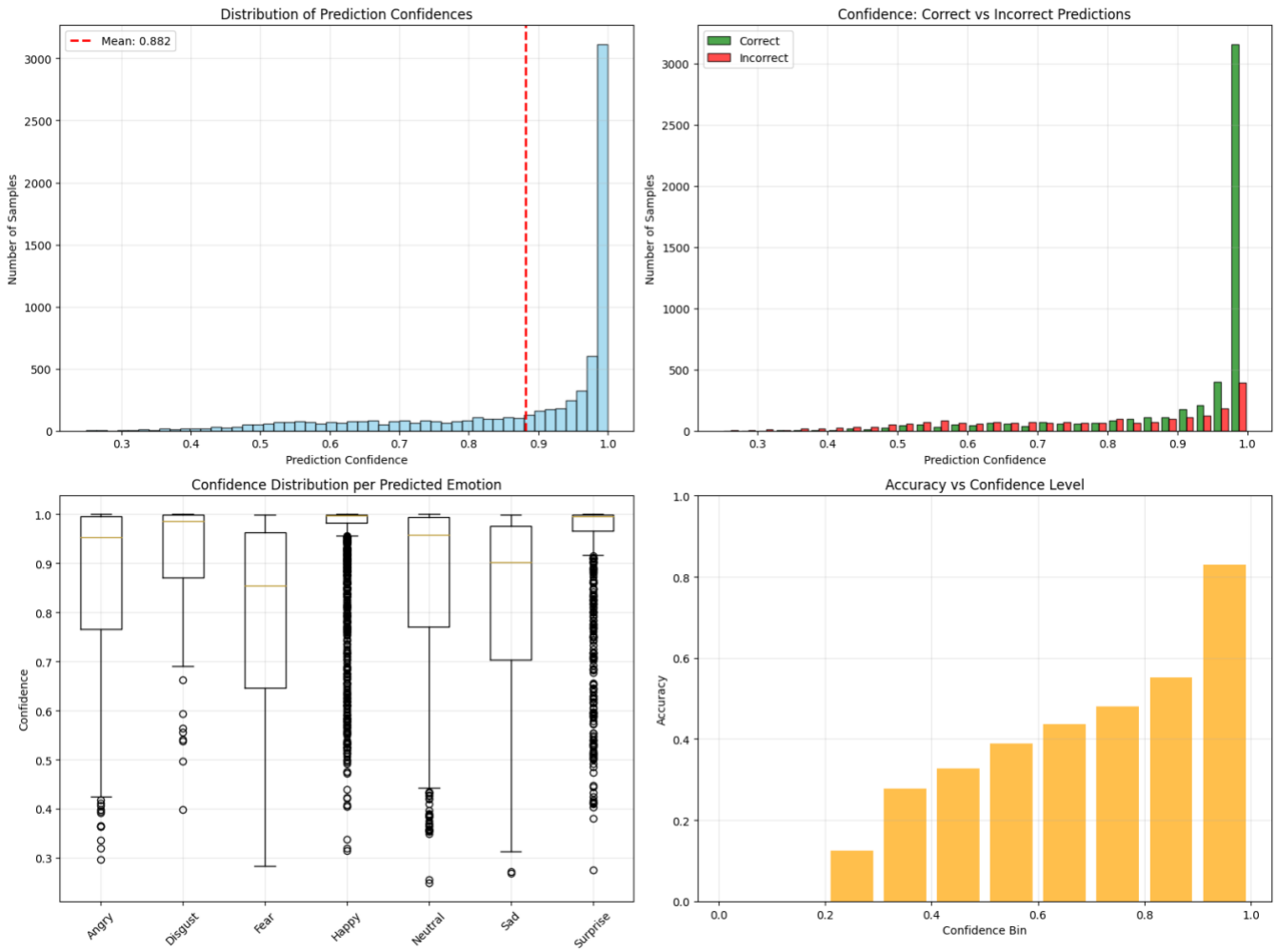


*Figure 8: Confidence distribution, boxplot, and accuracy vs. confidence relationship for VGG19 predictions.*

Table 2 outlines confidence-related statistics and reveals that while most predictions were made with high confidence, 735 samples (10.2%) had confidence scores below 0.6. These low-confidence predictions had an accuracy of only 35.8%, supporting the idea of setting a confidence threshold for flagging uncertain predictions.

*Table 6: Confidence Statistics for VGG19 Test Predictions*

| Metric | Value |
|---|---|
| **Mean Confidence (All)** | 0.882 |
| **Mean Confidence (Correct)** | 0.925 |
| **Mean Confidence (Incorrect)** | 0.779 |
| **Standard Deviation of Confidence** | 0.165 |
| **Predictions with Confidence < 0.6** | 735 |
| **Accuracy of Low-Confidence Samples** | 35.8% |

Finally, a misclassification audit revealed 2,131 incorrect predictions (29.7% of total). Among these, several were made with full certainty (confidence = 1.0), particularly in ambiguous or visually overlapping emotion categories. These high-confidence errors emphasize the need for further enhancements through techniques such as confidence calibration, ensembling, or adaptive thresholding.

In summary, the VGG19 model achieved reliable recognition of dominant expressions and produced

generally well-calibrated predictions. However, class imbalance and subtle inter-class differences still limited its performance on less distinct emotions. Future work may explore incorporating attention mechanisms, emotion-specific loss weighting, or fusion with temporal data to improve class discrimination and reduce overconfident misclassifications.

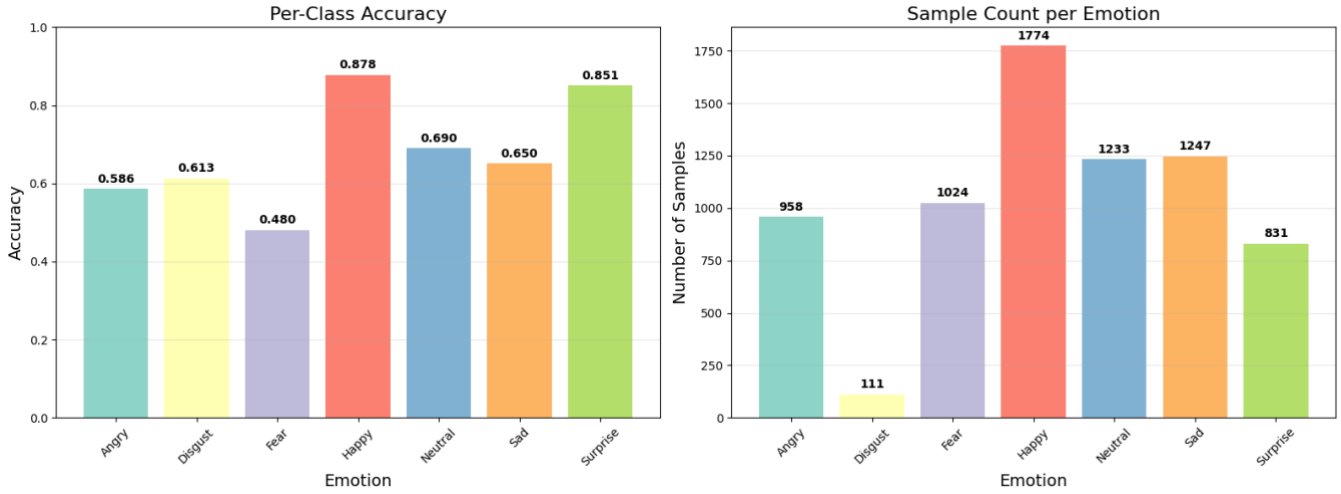


*Figure 9: Per-class Analysis*

# Comparison

To evaluate the impact of model architecture on facial expression recognition performance, three different CNN-based models were implemented and compared using the FER-2013 dataset: a custom baseline CNN, VGG19, and ResNet18. All models were trained under consistent preprocessing conditions, with evaluation based on accuracy, per-class performance, overfitting behavior, and prediction confidence.

The baseline CNN achieved a validation accuracy of 67.25%. It performed reasonably well on dominant expressions such as happy and neutral but struggled with minority classes like fear and disgust. The model was simple and efficient, making it suitable for resource-constrained environments, but its shallow architecture limited its capacity to capture complex facial features.

VGG19, utilizing a deeper architecture and transfer learning, achieved a test accuracy of 70.31%. It significantly improved performance on minority classes and produced more confident predictions. However, it showed signs of overfitting beyond 80 epochs, and some predictions were made with high confidence despite being incorrect.

ResNet18 outperformed both previous models with a test accuracy of 71.60%. The residual connections in its architecture helped maintain stable gradient flow, reducing overfitting and improving generalization. It demonstrated more balanced class-wise performance and handled subtle expressions with greater reliability.

*Table 7: Comparison*

| Model | Train Accuracy | Test Accuracy | Overfitting Gap | Strengths | Limitations |
|---|---|---|---|---|---|
| **Baseline CNN** | 73.75% | 67.25% | ~6.5% | Lightweight, fast convergence | Limited capacity, weak on subtle classes |
| **VGG19** | 93.41% | 70.31% | ~23.1% | Deep features, improved | Overfitting, high- |

| | | | | | |
|---|---|---|---|---|---|
| | | | | class balance | confidence errors |
| **ResNet18** | 94.00% | 71.60% | ~22.4% | Strong generalization, best accuracy | Slightly higher complexity |

This comparison demonstrates that deeper, transfer learning-based architectures offer considerable improvements in emotion classification accuracy and robustness. However, trade-offs in computational cost, overfitting, and interpretability must be carefully managed depending on the application.

Compared to existing methods in the literature, the three models presented—baseline CNN, VGG19, and ResNet18—achieve competitive performance. The baseline CNN aligns with earlier shallow models, while VGG19 and ResNet18, with accuracies of 70.31% and 71.60%, are comparable to more complex state-of-the-art architectures. Although newer models like DenseNet or attention-based networks may offer marginal gains, the proposed models balance accuracy and efficiency, making them suitable for practical and real-time applications.

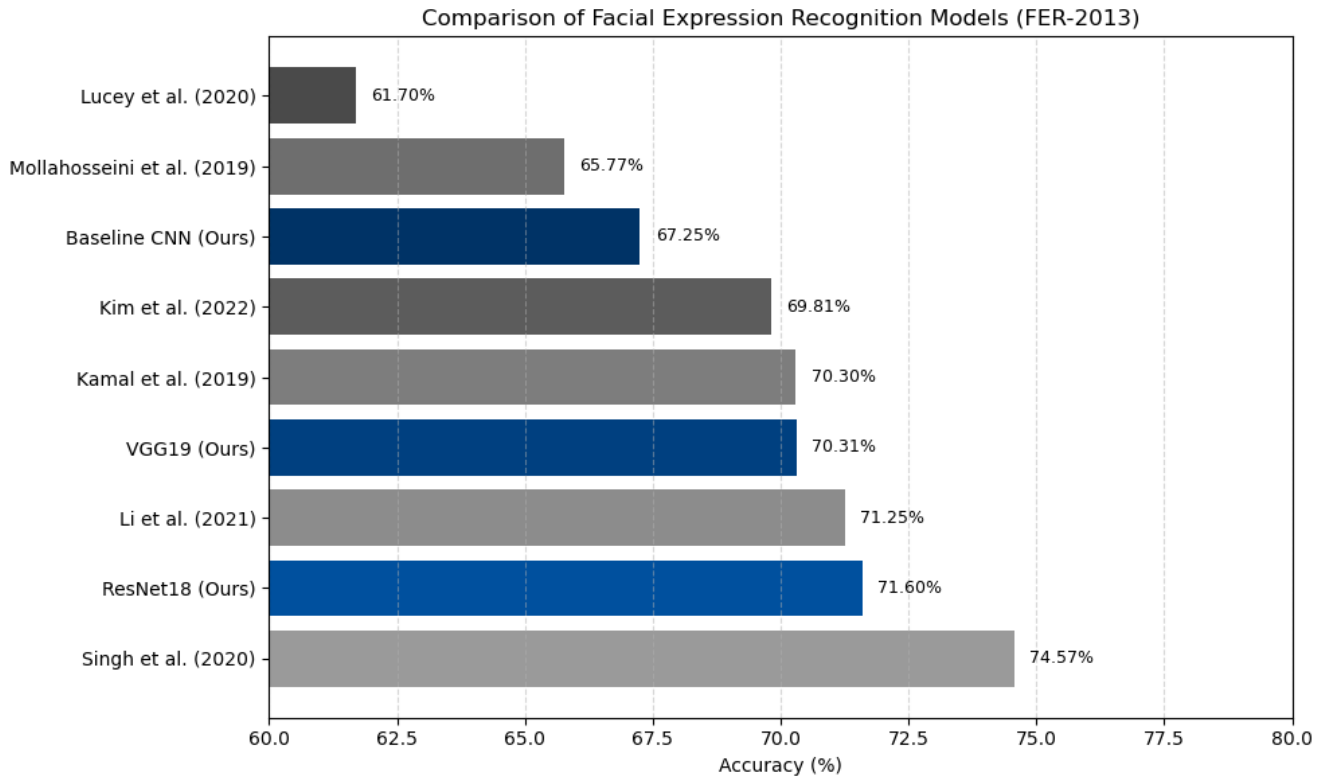


*Figure 10: Comparison*

# Discussion

In this study, we evaluated three convolutional architectures—baseline CNN, VGG19, and ResNet18—on the FER-2013 dataset to explore their effectiveness in facial expression recognition. The baseline CNN achieved a validation accuracy of 67.25%, while the VGG19 and ResNet18 models, both employing transfer learning, achieved improved test accuracies of 70.31% and 71.60%, respectively. These results demonstrate a clear progression in performance with increasing architectural depth and model complexity. The baseline CNN offered advantages in terms of simplicity and efficiency. Despite its relatively shallow architecture, it showed reasonable classification performance on dominant emotions such as happy and neutral. The model also exhibited stable convergence in loss and accuracy trends, suggesting effective

learning of core facial patterns. However, it significantly underperformed on underrepresented and visually subtle expressions like disgust and fear, as indicated by low precision and recall. This highlights a key limitation of shallow CNNs: sensitivity to class imbalance and insufficient capacity to learn discriminative features across nuanced emotion classes.

The VGG19 model, using a deeper architecture with 19 layers and pre-trained weights from ImageNet, showed a marked improvement in both accuracy and generalization. It achieved high precision and recall for dominant classes such as happy and surprise and demonstrated better performance on minority classes, although misclassifications still occurred—particularly among fear, sad, and angry. Visualization of training curves revealed an overfitting trend beyond epoch 80, with the training accuracy reaching 93.41% and test accuracy plateauing at 70.31%. This overfitting is attributable to the model's capacity exceeding the diversity and balance of the dataset. Additionally, some predictions were made with extremely high confidence (confidence = 1.0) despite being incorrect, revealing the model's tendency toward overconfident misclassifications.

ResNet18 further improved classification performance by leveraging residual connections to mitigate vanishing gradients and deepen learning. It reached the highest test accuracy of 71.60%, with reduced overfitting compared to VGG19. The residual learning mechanism allowed the model to learn more robust features, especially for classes with subtle expressions. ResNet18 also demonstrated more consistent class-wise performance across the confusion matrix and achieved higher F1-scores for difficult classes like disgust and sad. Nevertheless, errors persisted in distinguishing similar affective states (e.g., fear vs. sad), indicating that visual similarity among expressions remains a significant challenge even for deeper networks.

Across all models, the influence of dataset quality and structure was evident. The FER-2013 dataset, composed of grayscale images and imbalanced class distribution, limited the generalizability of all three architectures. Minority classes such as disgust and fear had insufficient representation, leading to poor recall scores even in deeper models. Additionally, cultural and physiological variations in emotional expression were not addressed in the dataset, which may have contributed to inconsistent recognition performance.

The confidence analysis conducted on VGG19 predictions revealed that the model's average confidence was higher for correct predictions (0.925) than incorrect ones (0.779), suggesting some reliability in model outputs. However, overconfidence in incorrect predictions reinforces the need for future work in model calibration or confidence-aware decision making.

# Conclusion

This study evaluated three CNN-based models—baseline CNN, VGG19, and ResNet18—for facial expression recognition using the FER-2013 dataset. The baseline CNN achieved a validation accuracy of 67.25%, performing well on dominant emotions like happy and neutral but struggling with subtle or underrepresented expressions such as fear and disgust. VGG19 and ResNet18, implemented using transfer learning, improved test accuracy to 70.31% and 71.60%, respectively. These deeper models showed better generalization and class-wise performance, though challenges remained in distinguishing visually similar emotions. Overall, results confirm that CNNs, even simple ones, can learn effective features, while deeper architectures significantly enhance accuracy and robustness.

# Future Work

Future work will focus on addressing class imbalance, incorporating color images, and integrating temporal or multimodal data such as speech and video. Advanced techniques like hybrid CNN-LSTM models, attention mechanisms, and model calibration can further improve recognition of subtle emotions and reduce overconfident misclassifications. These enhancements aim to make FER systems more reliable for real-world applications in healthcare, human-computer interaction, and emotion-aware technologies.

# References


[1] S. L. a. W. Deng, "Deep Facial Expression Recognition: A Survey," 2020.

[2] A. A. a. N. Mittal, "Using CNN for facial expression recognition: A study of the effects of kernel size and number of filters on accuracy," 2019.

[3] S. S. a. F. Nasoz, "Facial Expression Recognition with Convolutional Neural Networks," 2020.

[4] C. T. a. L. W. C. Shi, "A Facial Expression Recognition Method Based on a Multibranch Cross-Connection Convolutional Neural Network," 2021.

[5] S. K. S. C. C. a. S. M. P. E., "Facial Emotion Recognition Using Deep Convolutional Neural Network," *Proc. Int. Conf. AIML,* 2020.

[6] B. K. K. J. R. S. Y. D. and S. Y. L. , "Hierarchical Committee of Deep CNNs with a New Face Dataset for Human Expression Recognition," *IEEE Trans. Neural Netw. Learn. Syst.,* vol. 27, 2016.

[7] P. K. T. L. P. and T. S. H. , "Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition?," *Proc. ICCV,* 2015.

[8] X. Z. S. Z. and J. L. , "Facial Expression Recognition Using Graph Convolutional Networks," IEEE Trans. Image Process.," 2021.

[9] I. G. e. al, "Challenges in Representation Learning: A Report on Three Machine Learning Contests," *Neural Netw,* 2015.

[10] S. Z. C. Z. and Z. Z. , "A Survey on Face Detection in the Wild: Past, Present and Future," *Comput. Vis. Image Underst,* 2017.

[11] E. B. C. Z. C. C. F. and Z. Z. , "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution," 2016.

[12] A. M. B. H. and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Trans. Affective Comput.,* 2017.

[13] P. L. e. a. "The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression".

[14] K. S. and A. Z. , "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2015.

[15] D. Z. e. al, "Face2Exp: Combating Data Biases for Facial Expression Recognition," 2022.