Facial Emotion Recognition Based on Resnet-18 Model

Hamed M Suliman[1]

High Institute of Science and Technology, Misrata, Libya
hamedmmsuliman@gmail.com

Article information

Key words

Face emotion recognition; accuracy;

training.

Received 13 12 2025, Accepted 28 12 2025, Available online 28 12 2025

Abstract

This paper introduces an efficient computational method capable of real-time facial expression analysis, extracting features and classifying seven distinct emotional states from static images. To bridge the gap between traditional recognition and robust automated analysis, we propose an optimized deep learning framework based on the ResNet18 architecture. The model employs a hybrid transfer learning approach and enhances generalization through strategic data augmentation. Furthermore, the architecture leverages the skip connections inherent in ResNet to maintain feature integrity. Trained on the CK+ dataset, the model achieved a state-of-the-art accuracy of 98.4%.

I. Introduction

Facial expressions play a crucial role in both emotion recognition and individual identification, serving as a vital means of nonverbal communication. While speech is the primary channel for verbal contact, facial expressions convey internal sensations, allowing individuals to externalize their emotional states. Consequently, these states can be accurately perceived in others through their facial expressions (Tarnowski & Kołodziej, 2017). In the medical field, convolutional neural networks (CNNs) are employed to analyze facial expressions, enabling the automated detection of emotions based on facial cues. One such application assists clinicians in monitoring the impact of antidepressants by capturing daily facial images. By examining these daily changes, doctors can more accurately assess a patient's progress and refine treatment strategies accordingly (Zhang & Alazab, 2019). Facial emotion recognition mechanisms can largely be classified into two major methods. The first involves grouping predefined emotions using direct algorithmic means, while the second relies on detecting and describing specific facial features (Mehendale, 2020).

Traditional approaches for facial feature extraction—such as geometric and texture-based methods like Local Binary Patterns (LBP), Facial Action Units (FAUs), Local Directional Patterns (LDP), and Gabor wavelets—have been widely utilized. Recently, however, a new approach has emerged: deep learning. This method has demonstrated high levels of accuracy and reliability. The development of powerful convolutional neural networks (CNNs) capable of detecting and classifying features both accurately and efficiently has led researchers to increasingly adopt deep learning for facial emotion detection (Mellouka & Handouzia, 2020).

A Convolutional Neural Network (CNN) is a specialized architecture primarily composed of convolutional layers designed to extract features from images. These hidden layers perform a convolution operation, transforming input data before passing it to the subsequent layer—a process that makes CNNs exceptionally effective for image analysis.

In addition to convolutional layers, CNNs utilize fully connected layers. In these stages, every neuron is linked to every neuron in the previous layer, forming a dense network. These layers are responsible for high-level reasoning and are ultimately used to classify images into specific categories (Saravanan, Perichetla, & Gayathri, 2019).

hamedmmsuliman@gmail.com

II. Literature Review

Multiple factors impact the efficiency of existing facial emotion recognition models, including poor image quality, extremely small image sizes, varying head pose angles, diverse illumination conditions, camera lens distortion, and low image resolution. These factors play a vital role in determining the performance of a model. For example, extremely small images can significantly degrade accuracy, while poor image quality can restrict effective feature extraction, thereby decreasing algorithmic performance (El-Hag1 & Shafai, 2025). Furthermore, the size of the training dataset, the choice of feature extraction technologies, and the specific learning system approaches profoundly affect the accuracy of deep learning models. Latest researches have suggested many approaches to tackle these impedences. Kuruvayil et al. (El-Hag1 & Shafai, 2025) suggested a facial emotion recognition system consists of a feature embedding network, came after by a nearest neighbour Classifier. Their classification procedure enhanced by using meta-learning.

Arushi and Vivek (Haghpanah and Saeedizade, 2022) implemented multiple CNNs to classify static facial images into seven categories. They designed three different classifiers, including a five-layer CNN and a deeper parameterized CNN. By utilizing VGG16 and VGGFace to fine-tune multi-parameter models, they recorded a resulting accuracy of 48% for their facial emotion recognition model.

Siyue and Xie (Lakshman & Yadlapalli, 2022) developed a CNN architecture termed Deep Comprehensive Multi-patch Aggregation Convolutional Neural Networks (DCMA-CNN). This model consists of two branches: the first identifies local features from sample patches, while the second extracts holistic features from the entire facial image. In a different approach, Dawood et al. ((Lakshman & Yadlapalli, 2022) introduced the Viola-Jones framework for frontal face detection. For profile or side-view identification, they employed a skin-detection methodology, followed by the Histogram of Oriented Gradients (HOG) algorithm for the feature extraction stage. In recent years, most models for detecting facial emotional features have tended to omit essential emotional data. For instance, Ali et al. suggested the use of Support Vector Machines (SVM), while Evans proposed the Haar Wavelet Transform (HWT) method. Furthermore, the generalization and robustness of these network algorithms often exhibited low performance. In other studies, Lu employed biorthogonal wavelet entropy as a feature extractor with a fuzzy SVM classifier, while Phillips utilized the Jaya algorithm for facial emotion recognition. Similarly, Yang used cat swarm optimization to recognize emotions, while Li opted for Biogeography-Based Optimization (BBO) for the same task (Li & Lima, 2021).

Despite the advancements in deep learning, automated Facial Emotion Recognition (FER) remains a challenge due to the high variability of real-world data. Traditional feature extraction methods, such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG), often fail to capture the complex, non-linear patterns of human expressions under varying illumination and head poses. Furthermore, existing deep learning attempts often suffer from lack the generalization necessary for sensitive applications like medical patient monitoring. There is a critical need for a robust, high-accuracy framework that can minimize information loss and maintain performance across diverse image qualities. This study proposes an optimized deep learning framework based on a ResNet18 architecture to fill the gap between traditional recognition and robust automated analysis. The key contributions of this work are as follows:

1-Hybrid Transfer Learning Approach**:** Unlike models trained from scratch, our approach utilizes a pre-trained ResNet18 backbone, specifically fine-tuning the deep residual blocks (Layers 3 and 4) to capture high-level emotional small difference while keeping general feature knowledge.

2-Enhanced Generalization via Data Augmentation: To address the "Environmental Robustness" gap, we implemented a sophisticated augmentation pipeline including random rotation, and horizontal flipping. This ensures the model remains invariant to lighting distortions and camera angles.

3-Targeted Feature Extraction: By leveraging the global average pooling and skip connections inherent in ResNet, the model overcomes the "Information Loss" common in traditional networks, allowing for the detection of accurate expressions.

III. Building the Model

A. Dataset

The presented model was trained and tested using the Extended Cohn-Kanade (CK+) dataset, a well-known resmyce for emotion recognition. This dataset contains grayscale, face-cropped images resized to 48×48 pixels. The classification task requires predicting one of seven emotions—Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Contempt—from each facial expression.

The training partition consists of 789 examples, while the public test partition contains 192 examples. The CK+ dataset exhibits significant class imbalance: Anger (135 images), Disgust (177), Fear (75), Happiness (207), Sadness (84), Surprise (249), and Contempt (54). Unlike 'in-the-wild' datasets, CK+ includes posed expressions, resulting in clear images with high-intensity facial features. Examples of the CK+ dataset are shown in Figure 1. Notably, the Contempt class has significantly fewer examples (54) than others, whereas the Surprise class exhibits a substantial data bias due to its large sample size (249).

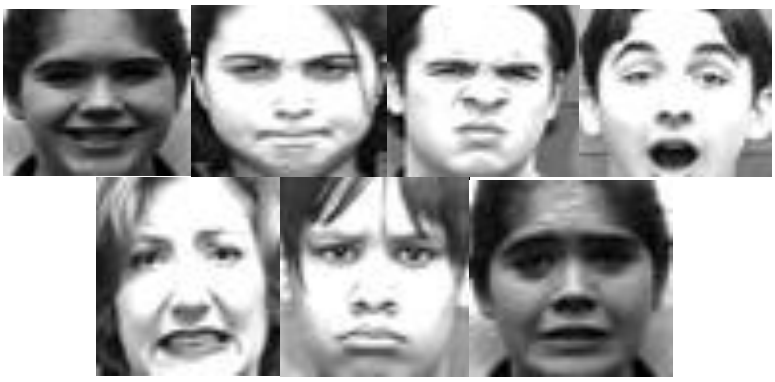

Figure 1. CK PLUS Expressions.

B. Data Processing

Input images undergo preprocessing to optimize them for subsequent feature extraction. An essential initial step involves converting all images to grayscale, regardless of their original format. This reduction in color information decreases computational complexity and allows the model to focus on facial muscle movements rather than skin tone variations.

Subsequently, pixel values are normalized to a range of [0, 1] to facilitate faster model convergence and accelerate training. The CK+ images are then resized to 48 x 48 pixels to remain compatible with standard architectures such as ResNet and VGG. Finally, data augmentation (rotation, shifting, zooming, horizontal flipping, and brightness/contrast adjustments) is applied via Keras' ImageDataGenerator. Training and validation streams are established using batch generators (batch size = 32) operating at a fixed image resolution. This augmentation pipeline enhances the model's generalization capability, ensuring more robust performance in real-world deployments.

C. Model Selection

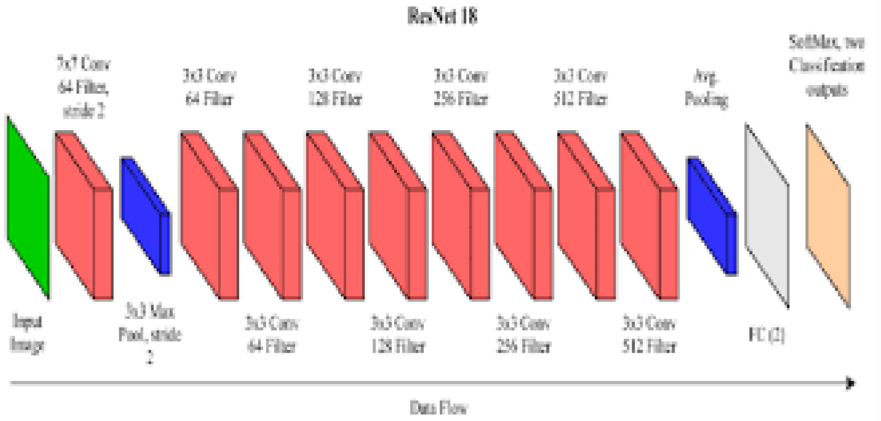

Figure 2. ResNet-18 model architecture.

The proposed emotion recognition system utilizes a ResNet-18 CNN architecture. This model is 18 layers deep and is renowned for its ability to mitigate the vanishing gradient problem through shortcut connections. The ResNet-18 architecture is depicted in Figure 2. ResNet-18 performs the following tasks:

1. Initial Feature Extraction (The "Stem").
The process begins with a 7 x 7 Convolution layer followed by Max Pooling. This layer investigate the raw image pixels to detect low-level features for example the simple edges and orientation of lines which represent the outlook of the jaw or bdidge of the nose.

2. Hierarchical Learning via Residual Blocks.
ResNet-18 contains eight residual blocks ordered into four stages. Each stage captures more abstract information than the last. Stage 1-2 (Low-Level Patterns). These layers identify local textures and shapes, such as the curves of the lips or the wrinkling around the eyes. Stage 3-4 (High-Level Semantics). These deeper layers detect complete facial components and their relative positions—for example, a wide-open mouth (surprise/happiness) or furrowed eyebrows (anger/sadness).

3. The Power of "Skip Connections".
The defining feature of ResNet is the Identity Shortcut (Skip Connection). The Problemi in standard deep networks, as you add more layers, the "gradient" (the signal used for learning) can vanish, making it hard for the model to train. The ResNet skip connection allows the input of a block to "jump" over the convolutional layers and be added directly to the output. This ensures that the model can learn the residua*l* (difference) between the input and output, preventing the loss of important facial information

4. Global Feature Aggregation.
After the final residual block, the model has a high-dimensional feature map (typically7 x7 x 512). Global Average Pooling (GAP). Instead of using a traditional flattened layer which has millions of parameters, ResNet-18 uses GAP. It takes the average of each feature map, condensing the entire face into a single feature vector (often 512 dimensions).

5. Final Classification.
The 512-dimensional vector is passed into a Fully Connected (FC) Layer which perfprm mapping and softmax operations.

1- Mapping: This layer maps the abstract features to the number of classes (e.g., 7 emotions: happiness, sadness, etc.).
2- Softmax: A Softmax function is applied to the output, converting the raw scores into probabilities. For example, if the model detects "raised cheeks" and "upturned lips," it will assign a 0.98 probability to the "Happiness" class. Table 1 shows the tasks performed by ResNet-18 CNN architecture.

Table 1. The tasks performed by ResNet-18 CNN architecture.

Stage	Component	Primary Function	Facial Feature Examples
1. Stem	7×7 Conv + Max Pooling	Low-level feature extraction and spatial reduction.	Jawline outline, bridge of the nose, basic edges.
2. Local Learning	Residual Stages 1 & 2	Detection of textures and local geometric shapes.	Lip curves, eye wrinkles, skin textures.
3. Semantic Learning	Residual Stages 3 & 4	Identifying complex components and spatial relationships.	Wide-open mouth, furrowed brows, eye positioning.

Stage	Component	Primary Function	Facial Feature Examples
4. Aggregation	Global Average Pooling (GAP)	Dimensionality reduction into a 512-D vector.	Summarized facial representation.
5. Prediction	Fully Connected + Softmax	Classification and probability calculation.	Mapping features to labels (e.g., 98% Happiness).

In my implementation, the model is initialized with ImageNet-1K weights to leverage pre-trained spatial features. The architecture is modified for a 7-class classification task by replacing the final fully connected layer with a linear layer consisting of 7 output units. To optimize the training process, a fine-tuning strategy is employed: the initial layers are frozen, while 'layer3', 'layer4', and the final fully connected (fc) layer remain trainable. This allows the model to adapt high-level semantic features specifically to facial expressions.

The preprocessing pipeline transforms the input images into a fixed resolution of 224 x 224 pixels with three color channels. To enhance the model's generalization capability and prevent overfitting on the CK+ dataset, we apply an extensive data augmentation suite, including Random Horizontal Flipping, Random Rotation (+/-15 degree), and Color Jitter (adjusting brightness and contrast by 0.3).

The model is optimized using the Adam optimizer with a learning rate of 1 x10^{-4} and Cross-Entropy Loss. Training is conducted over 20 epochs with a batch size of 16. Final evaluation is performed using a dedicated test partition, generating a confusion matrix and classification report to assess performance across all seven emotion categories. The used model is a ResNet-18 modified for emotion recognition using the CK+ dataset. Moreover, this model consists of three main "phases": the frozen feature extractor, the unfrozendeep layers (Layer 3 & 4), and the custom classification head. Tabel 2 illustrates the used mode architecture.

Tabel 2 illustrates the used model architecture.

Stage	Component	Output Shape	**Description**
Input	Imag**e**	224x224	image from CK+ dataset.
Initial Laye	Conv1 MaxPool	56x 56x 64	Frozen parameters (pre-trained)**.**
Layer 1 & 2	Residual Blocks	28x28x128	Frozen parameters; extracts basic features.
Layer 3 & 4	Residual Blocks	7x7x512	Unfrozen (Fine-tuning): requires_grad=True Extracts complex emotion features.
Global Pool	AvgPool	1x1x512	Flattens the spatial data.
Custom Head	Linear (FC)	1x1xNclasses	Modified: Maps 512 features to 7 emotions.

IV. Results and Discussion

We trained the model for 20 epochs. The training procedure was successful, as the accuracy and loss graphs demonstrate high performance and minimal loss, respectively. Figure 3 depicts the loss and accuracy results.

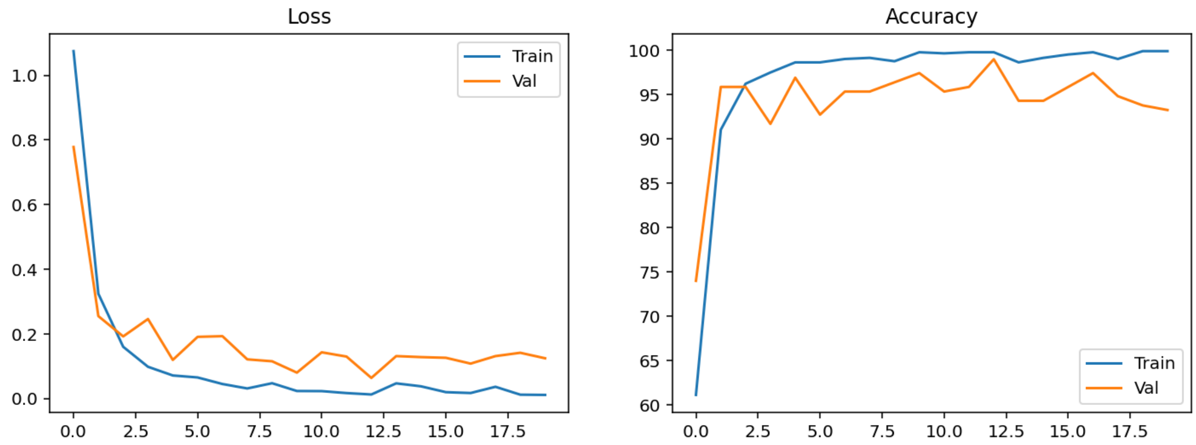

Figure 3. Loss and acuuuracy results.

Accuracy & Validation: The model achieved a validation accuracy of approximately 96%, while the training accuracy reached nearly 100%. A gap between training and validation accuracy was observed, suggesting a slight degree of overfitting. Nevertheless, the validation performance remained robust.

Loss Convergence: During the initial epochs, both training and validation loss declined significantly and remained low. This indicates that the model converged quickly and effectively learned to identify facial emotion expressions.

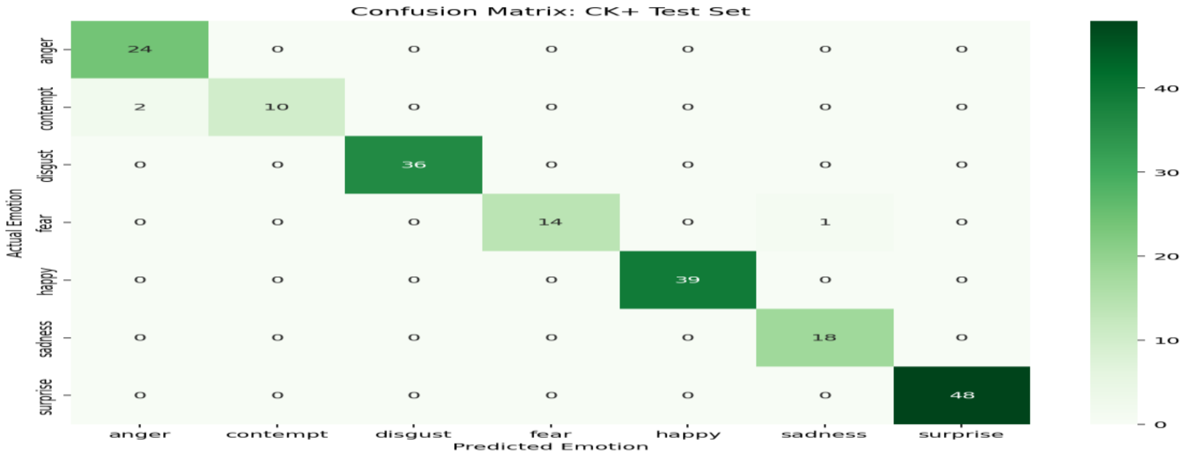

Figure 4. Confusion matrix.

- The confusion matrix revealed that my model scored perfect Classification in the most classes .the model achieved 100% accuracy for anger (24/24), disgust (36/36), fear (15/15), happy (39/39), sadness (18/18), and surprise (48/4). Therefore, Overall Test Accuracy: ~98.4% (189 correct out of 192 samples). However, the contempt class misclassified wrongly in some cases as follows: In one epoch, three contempt images were misclassified as anger, while in another epoch, two contempt images were falsely predicted as anger. Furthermore, one fear image was wrongly predicted as sadness. Consequently, the model encountered difficulties in discriminating between contempt and anger due to the feature similarities between these two expressions. Figure 4 shows Confusion matrix.

My model demonstrates exceptional performance, achieving 100% accuracy in 5 out of 7 categories. The table below explains the results: Tabel 3 illustrates the used model accuracies for each categories.

Tabel 3 the used model accuracies for each categories.

	Correct (TP)	Total Samples	Accuracy (%)
anger	24	24	**100.0%**
disgust	36	36	**100.0%**
happy	39	39	**100.0%**
sadness	18	18	**100.0%**
surprise	48	48	**100.0%**
contempt	9	12	**75.0%**
fear	15	15	**100.0%**

The model's primary challenge was accurately classifying the "contempt" category. Out of 12 actual contempt images, three were misclassified as "anger." This occurred because both classes share common facial features, such as tightened lips and furrowed brows. Additionally, the anger class contained twice as many images as the contempt class, leading to a class imbalance.

Figure 5 illustrates samples of predicted images. Table 4 shows the performance metrics of the model, while Figure 6 displays the performance metrics per emotion. Figure 6 reveals that five expressions—happy, sadness, surprise, disgust, and fear—all achieved 100% accuracy across all metrics. Anger has lower precision (blue bar) because it misclassified images originally belonging to contempt. This, in turn, led to a lower recall (orange bar) for contempt. The F1-score (green bar) provides the most balanced view of how the model handles the difficult "contempt" class (0.86). Lastly, live camera testing was conducted to evaluate the facial emotion recognition model (Figure 6). While the system effectively identified sadness, happiness, surprise, anger, and disgust, it struggled with fear and contempt due to the difficulty participants had in manifesting these emotions. Figure 7 displays the live testing environment.

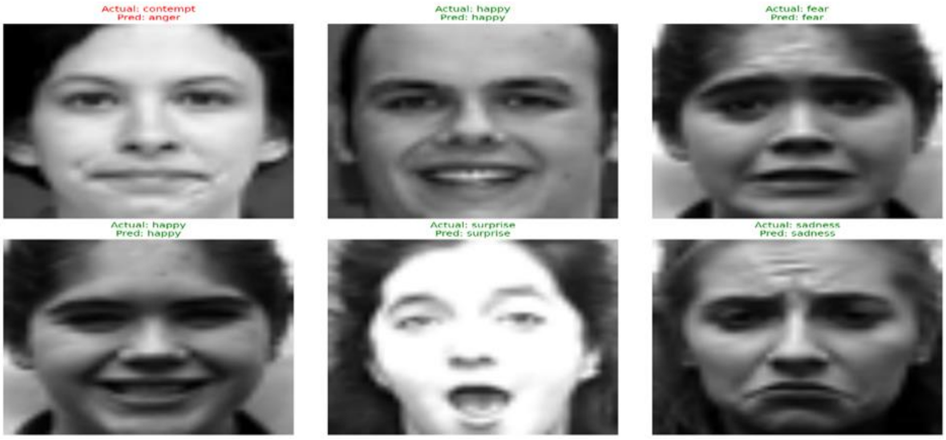

Figure 5. Samples of predicted image

Table 4. Some performance metrics of the used model.

Metric	Calculation	Result
Overall Accuracy	(24+9+36+15+39+18+48) / 192	**98.4%**
Precision (contempt)	9 / 9 (Model was always right when it said contempt)	**1.00**
Recall (contempt)	9 / 12 (Model missed 25% of actual contempt)	**0.75**
Precision (anger)	24 / 27 (Model was "confused" by contempt samples)	**0.89**

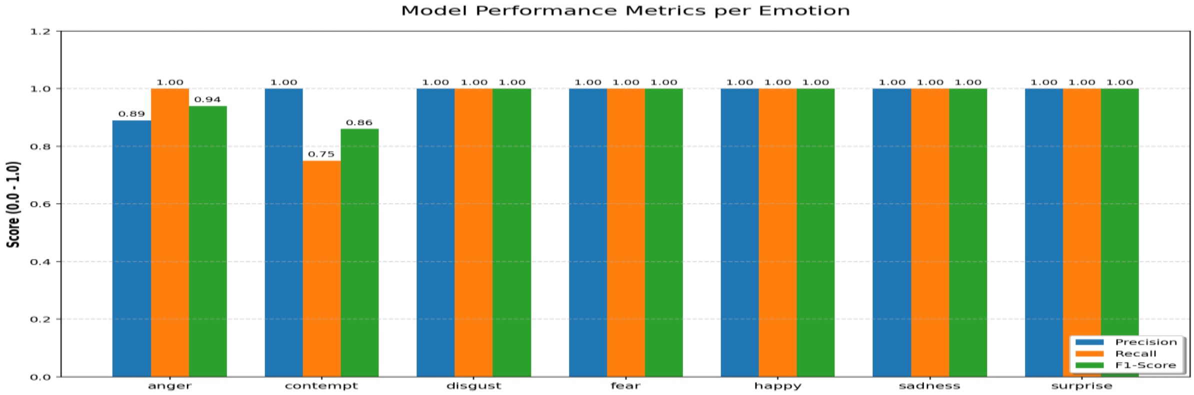

Figure 6. Model performance metrics per emotion.

V. Conclusion

In this paper, we utilize the ResNet-18 CNN architecture for facial emotion recognition to accurately extract distinct facial features. The proposed method extracts features directly from training images, enabling robust emotion classification. The model was trained to recognize seven categories: happiness, fear, disgust, sadness, surprise, anger, and contempt.

The ResNet-18 model achieved 100% accuracy for six of these classes; however, it scored only 75% for contempt, primarily due to samples being misclassified as anger. Despite this, the overall accuracy reached 98.4%, representing state-of-the-art performance for ResNet-18 on the CK+ dataset. To improve contempt prediction, we recommend employing oversampling, expanding the dataset for this specific class, or applying a higher loss weight to contempt during training.

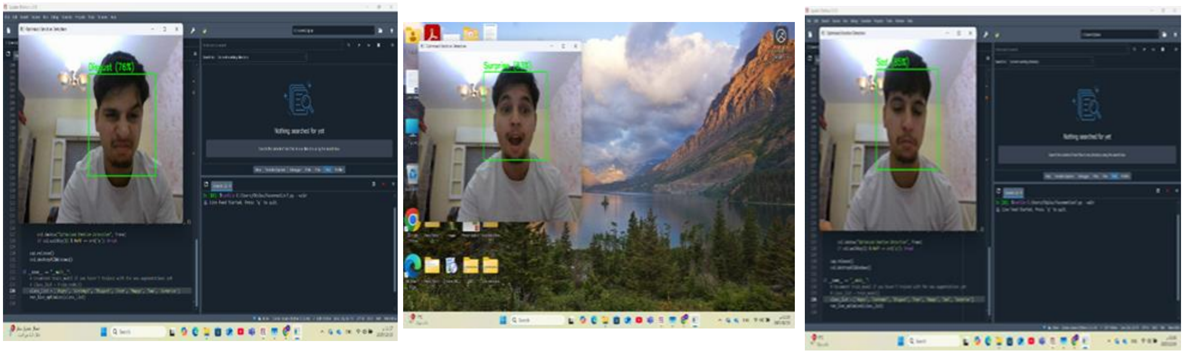

Figure 7 the live testing environment.

References

1-Tatikonda Lakshman, 2. S. (2022). FACIAL EMOTION DETECTION AND RECOGNITION. (pp. Issue 11 | ISSN: 2456-3315). IJRTI | Volume 7.

2-A.Saravanan, G. P. (2020). Facial Emotion Recognition using Convolutional Neural Networks. *Cornell university*.

3-Bin Li1, #. R. (2021). Facial expression recognition via ResNet-18. *International Conference on Multimedia Technology and Enhanced Learning.* Springer.

4-H. Zhang, M. A. (2019). A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing. (p. Volume 7). IEEE Access.

5-Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks. *Springer Nature Switzerland*.

6-Mohammad A. Haghpanah, E. S. (22 March 2022). Real-Time Facial Expression Recognition using. *International Conference on Machine Vision and Image.* Ahvaz, Iran, Islamic Republic of: ieeexplore.ieee.org.

7-Noha A. El-Hag1 · Walid El-Shafai2, 3. ·.-S. (9/9/2025). Enhanced Facial Emotion Recognition and Age. *Int J Comput Intell Syst*.

8-P. Tarnowski, M. A. (2017). Emotion recognition using facial expressions. *International Conference on Computational Science, ICCS* (pp. 12-14). Elsevier.

9-W. Mellouka, w. H. (2020). Facial emotion recognition using deep learning: review and insights. *The 2nd International Workshop on the future of Internet of Every thing (FIoE)* (pp. 689-694). Elsevier.

التعرف على عواطف الوجه

حمد م سليمان

استلمت الورقة بتاريخ 2025/12/13 وقبلت بتاريخ 2025/12/28، ونشرت بتاريخ 2025/12/28

الكلمات المفتاحية: التعرف علي عواطف الوجه ، الدقة، التدريب.

الملخص

تقدم هذه الورقة البحثية طريقة حسابية فعّالة قادرة على تحليل تعابير الوجه في الوقت الفعلي، واستخلاص السمات وتصنيف سبع حالات عاطفية متميزة من الصور الثابتة. ولسد الفجوة بين تقنيات التعرف التقليدية والتحليل الآلي القوي، نقترح إطار عمل مُحسَّن للتعلم العميق قائم على بنية **ResNet18**. يستخدم النموذج نهجًا هجينًا للتعلم بالنقل، ويعزز التعميم من خلال زيادة البيانات بشكل استراتيجي. علاوة على ذلك، تستفيد البنية من وصلات التخطي المتأصلة في **ResNet** للحفاظ على سلامة السمات. وقد حقق النموذج، الذي تم تدريبه على مجموعة بيانات **CK**+، دقة فائقة بلغت 98.4%.