Karaelmas Science and Engineering Journal
Journal home page: https://dergipark.org.tr/tr/pub/karaelmasfen
DOI: 10.7212/karaelmasfen.1850538

Research Article
Received / Geliş tarihi : 28.12.2025
Accepted / Kabul tarihi : 08.04.2026

Emotion Analysis from Facial Expressions Using a Two-Stage Transfer Learning Approach: A Comparative Study of Deep Learning Models

İki Aşamalı Transfer Learning Yaklaşımıyla Yüz İfadelerinden Duygu Analizi: Derin Öğrenme Modellerinin Karşılaştırmalı İncelenmesi

Mustafa Furkan Keskenler*

Atatürk University, Faculty of Engineering, Department of Artificial Intelligence and Data Engineering, Erzurum, Türkiye

Abstract

Emotion analysis from facial expressions is a critical research problem with significant applications in human–computer interaction, security systems, and healthcare technologies. Although deep learning–based methods have achieved remarkable progress in recent years, the comparative performance of different architectures and the impact of two-stage transfer learning strategies still require comprehensive investigation. In this study, four widely used deep learning architectures—ResNet50, DenseNet121, InceptionV3, and MobileNetV2—are systematically compared for the task of facial emotion recognition within a two-stage transfer learning framework. In the first stage, models pre-trained on the ImageNet dataset were employed by freezing the convolutional layers and training only task-specific classifier layers. In the second stage, selected upper layers were unfrozen, and fine-tuning was performed to adapt the models more effectively to the target dataset. The models were evaluated not only in terms of classification performance metrics such as accuracy, loss, precision, recall, and F1-score, but also with respect to training time, number of parameters, and model size. All experiments were conducted with multiple repetitions in a GPU environment to enhance the statistical reliability of the results. The experimental findings demonstrate that the two-stage training strategy leads to significant performance improvements across all models. DenseNet121 achieved the most balanced and highest overall performance, while ResNet50 exhibited high accuracy and strong generalization capability. MobileNetV2, despite lower absolute accuracy, emerged as an effective alternative for resource-constrained applications due to its low computational cost. Overall, this study provides a comprehensive evaluation of the trade-off between model performance and computational efficiency in facial emotion recognition, contributing valuable insights to this research domain.

Keywords: Convolutional neural networks, deep learning, facial emotion recognition, model comparison, transfer learning

Öz

Yüz ifadelerinden duygu analizi, insan–bilgisayar etkileşimi, güvenlik sistemleri ve sağlık uygulamaları gibi birçok alanda kritik öneme sahip bir araştırma problemidir. Son yıllarda derin öğrenme tabanlı yöntemler bu alanda önemli ilerlemeler sağlamış olsa da, farklı mimarilerin karşılaştırmalı performansı ve iki aşamalı transfer learning stratejilerinin etkisi hâlen kapsamlı biçimde incelenmeye ihtiyaç duymaktadır. Bu çalışmada, yüz ifadelerinden duygu analizi problemi için dört yaygın derin öğrenme mimarisi — ResNet50, DenseNet121, InceptionV3 ve MobileNetV2 — iki aşamalı transfer learning yaklaşımı çerçevesinde sistematik olarak karşılaştırılmıştır. Çalışmada, ImageNet veri kümesi üzerinde önceden eğitilmiş modeller kullanılarak ilk aşamada evrişimsel katmanlar dondurulmuş ve yalnızca problem-özgü sınıflandırıcı katmanlar eğitilmiştir. İkinci aşamada ise seçilen üst katmanlar yeniden eğitime açılarak ince ayar gerçekleştirilmiştir. Modeller, doğruluk, kayıp, kesinlik, ger çağırma ve F1 skoru gibi performans metriklerinin yanı sıra, eğitim süresi, parametre sayısı ve model boyutu açısından da değerlendirilmiştir. Deneyler GPU ortamında çoklu tekrarlar hâlinde yürütülerek sonuçların istatistiksel güvenilirliği artırılmıştır. Elde edilen sonuçlar, iki aşamalı eğitim stratejisinin tüm modellerde anlamlı performans artışı sağladığını göstermiştir. DenseNet121 modeli en dengeli ve en yüksek genel performansı sunarken,

*Corresponding author: mfkeskenler@atauni.edu.tr

Mustafa Furkan Keskenler orcid.org/0000-0002-7604-4179

ResNet50 yüksek doğruluk ve güçlü genelleme yeteneğiyle dikkat çekmiştir. MobileNetV2 ise düşük hesaplama maliyetiyle kaynak kısıtlı uygulamalar için etkili bir alternatif olarak öne çıkmıştır. Bu çalışma, yüz ifadelerinden duygu analizi alanında model başarımı ile hesaplama maliyeti arasındaki dengeyi bütüncül biçimde ortaya koyarak bu çalışma alanına önemli bir katkı sunmaktadır.

Anahtar Kelimeler: Evrişimsel sinir ağları, derin öğrenme, yüz ifadelerinden duygu analizi, model karşılaştırması, transfer learning

1. Introduction

Emotions are fundamental psychological components that directly influence individuals' cognitive processes, social interactions, and decision-making mechanisms. The automatic perception and interpretation of human emotions have attracted increasing research interest, particularly in areas such as human–computer interaction, intelligent surveillance systems, healthcare informatics, and educational technologies (Calvo & D'Mello, 2010). In this context, facial expressions stand out as one of the most widely used visual modalities in emotion recognition studies, as they provide universal and powerful cues for emotional expression (Ekman & Friesen, 1971; Song et al., 2024).

Traditional approaches to facial emotion recognition have generally been developed using hand-crafted feature extraction techniques and classical machine learning algorithms. However, these methods struggle to adequately represent the complex spatial patterns of facial expressions, inter-individual variations, and changes caused by environmental conditions (Zeng et al., 2009; Gao et al., 2026). In recent years, deep learning—particularly convolutional neural networks (CNNs)—has become the dominant approach in facial emotion recognition due to its strong representational capacity in image processing tasks (LeCun et al., 2015).

Convolutional neural networks are capable of hierarchically learning low-level features such as edges and textures as well as high-level semantic representations. Owing to this capability, CNN-based models can represent the complex structure of facial expressions more effectively and achieve higher classification performance compared to traditional methods (Goodfellow et al., 2016; Huang et al., 2026). Nevertheless, training deep learning models from scratch requires large-scale datasets and substantial computational resources, which poses a significant limitation for domains such as facial emotion analysis, where annotated data are relatively limited.

To address this challenge, transfer learning has emerged as an effective solution by enabling models trained on large-scale datasets to be adapted to different target tasks. Transfer learning allows general visual representations learned from comprehensive datasets such as ImageNet to be transferred to smaller, domain-specific problems (Pan & Yang, 2010; Darraz et al., 2025). Numerous studies have reported that transfer learning significantly improves facial emotion recognition performance (Li & Deng, 2020; Pandey & Vishwakarma, 2024).

In facial emotion analysis, deep learning models with different architectural design philosophies—such as ResNet, DenseNet, Inception, and MobileNet—are widely employed. The ResNet architecture enables the training of very deep networks through residual connections (He et al., 2016; Mughal et al., 2024), while DenseNet enhances feature reuse through dense inter-layer connections (Huang et al., 2017; Yang et al., 2024). InceptionV3 facilitates multi-scale feature extraction through parallel convolutional structures operating at different spatial resolutions (Szegedy et al., 2016; Li et al., 2024). MobileNetV2, on the other hand, offers advantages for resource-constrained environments due to its lightweight design and low computational cost (Sandler et al., 2018; Ruan et al., 2024).

Although these models have been applied to facial emotion recognition in the literature, many studies focus on a single architecture or fail to systematically compare different models under the same training strategy. Furthermore, most existing works primarily report accuracy values, while practical considerations such as training time, number of parameters, and hardware requirements are often overlooked. In addition, studies that comprehensively examine the impact of two-stage training strategies—consisting of feature extraction and fine-tuning—on facial emotion recognition performance remain limited.

While numerous studies in facial emotion recognition primarily focus on improving classification accuracy using deep learning models, practical deployment of such systems requires consideration of additional factors such as computational cost, model size, and processing speed. In real-world applications—particularly in embedded, mobile, or resource-constrained environments—these factors can be as critical as recognition performance itself. Therefore, this study does not aim to propose a new learning paradigm; instead, it provides a systematic comparative evaluation of

widely used CNN architectures under the same two-stage transfer learning protocol, analyzing both their predictive capability and their computational efficiency. This dual-perspective evaluation offers a more application-oriented understanding of model selection for facial emotion recognition systems. The obtained results demonstrate competitive—and in several cases improved—performance compared to previously reported approaches, thereby offering practical insights into model selection for facial emotion recognition systems.

In this study, four widely used deep learning architectures—ResNet50, DenseNet121, InceptionV3, and MobileNetV2—are comparatively investigated for facial emotion recognition within a two-stage transfer learning framework. In the first stage, pre-trained convolutional layers are frozen, and only task-specific classifier layers are trained. In the second stage, selected upper layers are unfrozen, and fine-tuning is performed. The models are evaluated not only in terms of accuracy, loss, and F1-score, but also with respect to training time, number of parameters, and model size. Consequently, the trade-off between model performance and computational cost in facial emotion analysis is comprehensively analyzed.

2. Related Works

Facial emotion recognition has long been an active research topic in computer vision and artificial intelligence, and significant progress has been achieved in recent years with the widespread adoption of deep learning–based approaches. Early studies in this field primarily relied on hand-crafted features—such as LBP, HOG, and Gabor filters—combined with classical machine learning algorithms. While these approaches produced acceptable results in controlled environments, they exhibited limited generalization capability under real-world conditions involving variations in pose, illumination, and individual facial characteristics (Zeng et al., 2009).

The success of deep learning in image processing has directly influenced facial emotion recognition research. In particular, convolutional neural networks have enabled automatic learning of spatial patterns in facial images, leading to significantly higher recognition accuracy compared to traditional approaches (LeCun et al., 2015). Numerous studies have demonstrated that CNN-based methods can effectively capture the hierarchical and complex features inherent in facial expressions (Goodfellow et al., 2016).

However, training CNN models from scratch requires large-scale labeled datasets and extensive computational resources, which presents a major challenge for facial emotion recognition tasks. As a result, transfer learning has been widely adopted in the literature. By leveraging models pre-trained on large datasets, transfer learning reduces training time while improving classification performance (Pan & Yang, 2010).

A substantial number of studies have applied transfer learning to facial emotion recognition. Li and Deng (2020) provided a comprehensive survey highlighting the advantages of deep CNN architectures in this domain and reported significant performance improvements achieved through transfer learning. Similarly, studies employing ImageNet pre-trained models have demonstrated that satisfactory results can be obtained even with limited data (Mollahosseini et al., 2017).

Various CNN architectures have been explored for facial emotion recognition. ResNet-based models benefit from residual connections that enable stable training of deep networks and achieve high recognition accuracy (He et al., 2016). DenseNet architectures promote feature reuse through dense connectivity, resulting in compact and efficient representations (Huang et al., 2017). Inception architectures facilitate multi-scale feature extraction and effectively represent complex facial expressions (Szegedy et al., 2016). Lightweight architectures such as MobileNet further extend facial emotion recognition to resource-constrained platforms by reducing computational requirements (Sandler et al., 2018).

Despite these advances, several limitations remain in the existing literature. First, many studies focus on a single deep learning architecture and do not provide systematic comparisons across different models under identical experimental conditions. Second, performance evaluations are often limited to accuracy, while other critical metrics, such as F1-score, training time, parameter count, and model size, are frequently neglected. Moreover, although transfer learning is widely used, relatively few studies have systematically investigated the impact of two-stage training strategies—feature extraction followed by fine-tuning—on facial emotion recognition performance. In addition, comprehensive evaluations comparing CPU- and GPU-based training in terms of performance and computational cost are scarce.

This study aims to address these research gaps by conducting a comprehensive comparative analysis of four widely used deep learning architectures—ResNet50, DenseNet121, InceptionV3, and MobileNetV2—within a two-stage transfer

learning framework. Beyond classification performance, the models are evaluated in terms of computational efficiency, including training time, parameter count, and model size. Furthermore, multiple experimental repetitions in a GPU environment are performed to enhance the statistical reliability of the results. Through these contributions, this study provides both methodological and empirical insights into facial emotion recognition using deep learning.

3. Materials and Methods

The problem addressed in this study is the automatic classification of individuals' emotional states from facial images. The task is formulated as a multi-class image classification problem, where each input image is represented as defined in Equation (1).

$$x_i \in R \;\; H \times W \times 3 \quad y_i \in \{1,2,\ldots,5\} \tag{1}$$

The objective is to predict a label corresponding to one of five emotion classes. The primary goal is not only to achieve high classification performance on the training data but also to maximize the generalization capability of the learned model on previously unseen test samples.

In this context, rather than focusing on the performance of a single deep learning architecture, the study presents a comparative analysis of four widely used pre-trained networks with different architectural design philosophies. This approach enables a systematic evaluation of the relationships between model complexity, computational cost, and classification performance in the facial emotion recognition task.

Training deep convolutional neural networks from scratch typically requires large-scale datasets and substantial computational resources. Therefore, this study adopts a transfer learning approach. Transfer learning is based on transferring representations learned from a large and general-purpose dataset (e.g., ImageNet) to a smaller and domain-specific problem.

Accordingly, four different architectures are evaluated: ResNet50, DenseNet121, InceptionV3, and MobileNetV2. These models respectively represent distinct architectural paradigms, including residual connections, dense connections, multi-scale convolutional blocks, and depthwise separable convolutions. This architectural diversity allows for an in-depth analysis of the impact of model design choices on facial emotion recognition performance.

In this study, a two-stage transfer learning–based deep learning approach is proposed to address the facial emotion recognition problem. The overall workflow of the proposed method is illustrated in Figure 1. The process begins with the preprocessing of the dataset, followed by the use of convolutional neural network architectures pre-trained on the ImageNet dataset as feature extractors. The original classification layers of the pre-trained networks are removed, and a new classifier structure specifically designed for the facial emotion recognition task is integrated into the models.

The training process was conducted in two stages. In the first stage, the convolutional base layers were frozen, and only the newly added classifier layers were trained to enable the learning of fundamental discriminative features. In the second stage, selected upper convolutional layers were unfrozen and fine-tuned, allowing the model to learn more subtle and emotion-specific facial features. Finally, the trained models were evaluated using performance metrics such as accuracy, loss, F1-score, and training time. This integrated workflow enables a comprehensive analysis in terms of both classification performance and computational cost.

3.1. Dataset and Preprocessing

The facial expression–based emotion analysis dataset used in this study was obtained from the publicly available "Human Face Emotions" dataset shared on the Kaggle platform. The dataset consists of facial images from different individuals and includes labeled samples representing basic emotional states. Example images from the dataset are illustrated in Figure 2.

The dataset used in this study is the publicly available Human Face Emotions dataset hosted on the Kaggle platform. The dataset can be accessed at:

"https://www.kaggle.com/datasets/samithsachidanandan/human-face-emotions" It contains labeled facial images categorized into five emotion classes: Angry, Fear, Happy, Sad, and Surprise. All images are anonymized and provided for research purposes under the platform's data-sharing policy. Since no personal identification information is included and the dataset is publicly accessible, this study does not require additional ethical committee approval.

The dataset is organized into five distinct emotion classes: Angry, Fear, Happy, Sad, and Surprise. The images were collected under natural environmental conditions and exhibit significant variability in terms of facial pose, illumination, and individual differences. While this diversity enhances the dataset's ability to reflect real-world scenarios, it also increases the difficulty of the classification task.

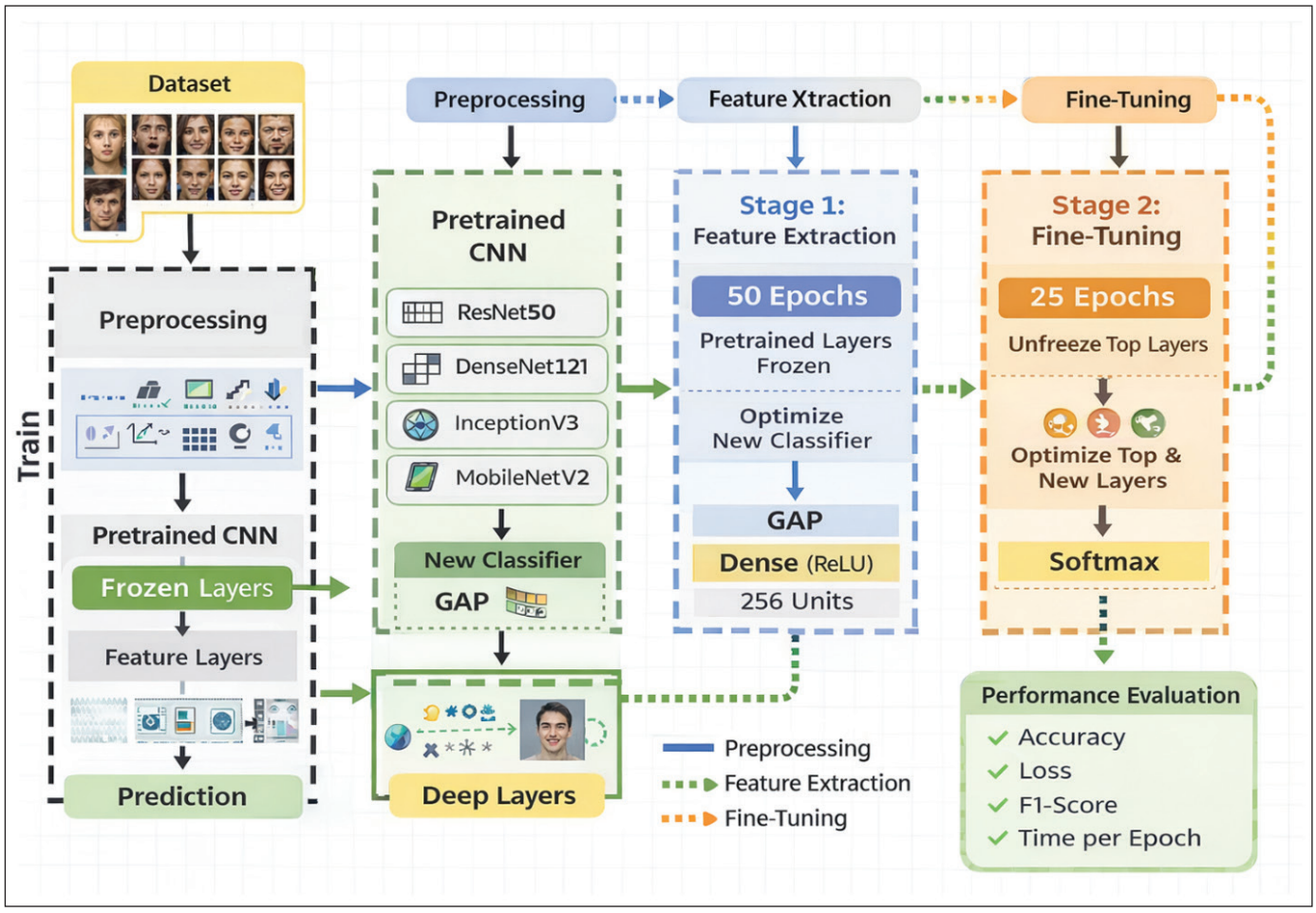

Figure 1. Overall workflow of the proposed method.

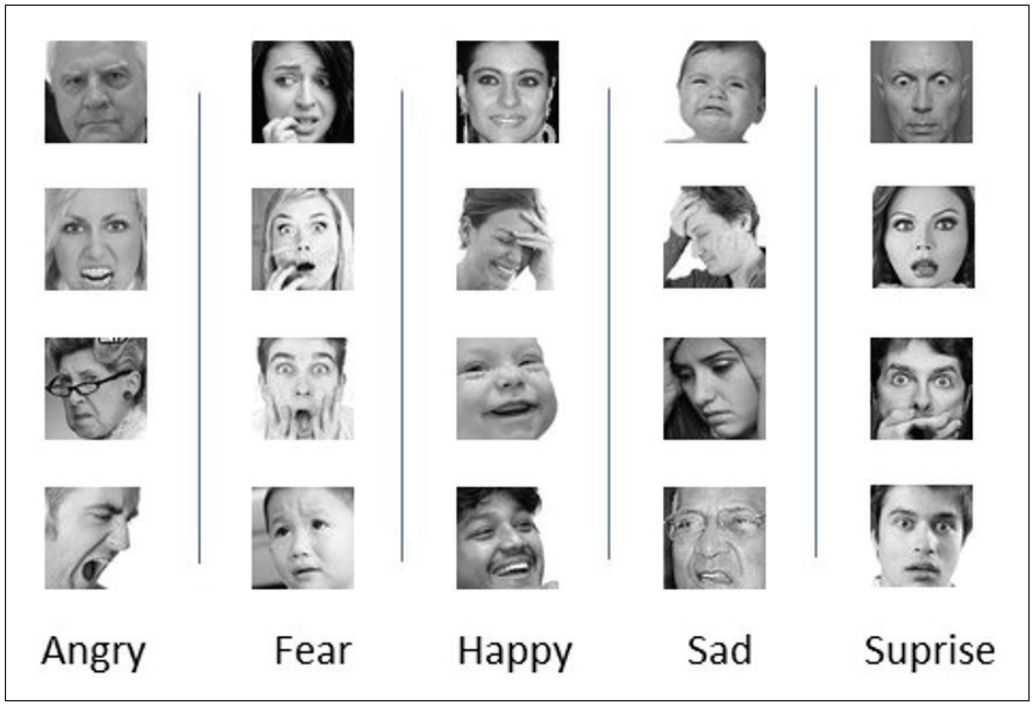

Figure 2. Sample images from the dataset.

For the purposes of this study, the dataset was divided into training and test subsets. The training set consists of 47,277 images, while the test set contains 11,822 images. This data split was kept fixed throughout all experiments to ensure a fair and direct comparison of the performance of different deep learning architectures. The class distribution in the dataset exhibits a certain degree of imbalance among emotion categories, which necessitated the use of additional evaluation metrics such as the F1-score during performance assessment. The dataset consists of 59,099 facial images distributed across five emotion categories: Angry (10,148), Fear (9,732), Happy (18,439), Sad (12,553), and Surprise (8,227). The distribution is moderately imbalanced, with the Happy class containing the largest number of samples, while Surprise represents the smallest portion. This imbalance motivates the use of F1-score alongside accuracy for performance evaluation.

All images were resized to meet the input requirements of the pre-trained deep learning models. An input resolution of 224 × 224 was used for the ResNet50, DenseNet121, and MobileNetV2 models, while 299 × 299 was employed for InceptionV3, as required by its architecture. All images were processed in the RGB color space, and pixel values were normalized to the [0,1] range. Since the original images are 48×48 pixels, they were resized to match the input resolution required by pretrained CNN architectures (e.g., 224×224). This resizing step does not create additional semantic information; it only interpolates existing pixel values to ensure architectural compatibility. Therefore, the objective is not to enhance image detail, but to evaluate how pretrained feature extractors operate under low-resolution conditions. The impact of resizing should thus be interpreted as a structural requirement of transfer learning rather than an independent performance optimization.

Although a substantial portion of the images in the dataset are in grayscale format, all pre-trained models employed in this study require RGB input representations (e.g., 224 × 224 × 3). Therefore, grayscale images were converted into three-channel RGB format before being fed into the models. During this conversion process, the single-channel grayscale intensity values were replicated across all three channels, assigning identical intensity values to each channel. This approach preserves the structural and textural information of the images while ensuring compatibility with the architectural requirements of the pre-trained networks.

This conversion strategy is widely adopted in the literature and enables grayscale images to be effectively processed by RGB-based deep learning models. Consequently, the convolutional filters learned on the ImageNet dataset can be directly utilized, maintaining the effectiveness of the transfer learning approach.

No data augmentation techniques were applied in this study. This deliberate choice allows the evaluation to focus on the inherent transfer learning capability of the pretrained networks without introducing artificially generated variations into the dataset.

3.2. Deep Learning Architectures Used

In this study, four widely adopted pre-trained deep convolutional neural network (CNN) architectures were evaluated for the facial emotion recognition task: ResNet50, DenseNet121, InceptionV3, and MobileNetV2. All models were initialized using weights pre-trained on the ImageNet dataset.

The original classification layers of the pre-trained models produce 1000-class outputs specific to ImageNet. Therefore, these top layers were removed in all architectures, and a new task-specific classifier was appended. The added classifier consists of the following components:

- Global Average Pooling (GAP) layer: Reduces the spatial feature maps obtained from the convolutional layers into a one-dimensional feature vector. Compared to fully connected layers, this approach significantly reduces the number of trainable parameters and mitigates the risk of overfitting.
- Dense layer with 256 neurons (ReLU activation): Transforms the learned general visual representations into discriminative features specific to emotion classes.
- Softmax output layer: Produces class probability distributions for multi-class classification.

These additional layers were designed to adapt the general visual representations learned from ImageNet to the facial emotion recognition task. While the GAP layer reduces model complexity and overfitting risk, the Dense layer enables the learning of high-level emotion-specific features.

3.3. Two-stage Training Strategy

Model training was conducted using a two-stage strategy. In the first stage, the convolutional base was entirely frozen, and only the newly added classifier layers were trained. This stage aims to preserve the pre-trained representations while enabling rapid adaptation of the classifier to the target task.

In the second stage, a subset of the upper convolutional layers was unfrozen and included in training. This fine-tuning process allows deeper layers of the network to learn fine-grained features specific to facial expressions. Reducing the learning rate during this stage prevents abrupt and unstable weight updates, leading to a more stable convergence process.

This two-stage strategy constitutes the fundamental methodological basis for the performance improvements observed in the training and validation curves.

The hyperparameters used in this study were selected based on widely adopted practices in transfer learning applications. The Adam optimizer was preferred due to its adaptive learning capability and stable convergence behavior in fine-tuning scenarios. A relatively higher learning rate (10^{-3}) was used during Stage-1 to allow rapid adaptation of the newly added classifier layers, while a lower learning rate (10^{-4}) was employed in Stage-2 to enable gradual refinement of pretrained features without disrupting previously learned representations. The number of training epochs (50 for Stage-1 and 25 for Stage-2) was determined empirically to ensure convergence while avoiding overfitting. These settings provide a balance between training stability, computational efficiency, and model generalization.

3.3.1. Stage 1: frozen layers

In the first stage, all convolutional layers of the pre-trained networks were frozen, and only the added classifier layers were trained. This phase is commonly referred to as feature extraction. The primary objective is to transfer the general visual features learned from the ImageNet dataset directly to the facial emotion recognition task.

During this stage, the models were trained for 50 epochs, with a relatively higher learning rate (Adam optimizer, lr = 10^{-3}) to facilitate rapid adaptation of the classifier layers. The Stage-1 results provide insight into how well the pre-trained representations generalize to the emotion recognition problem.

3.3.2. Stage 2: fine-tuning

In the second stage, a selected portion of the upper convolutional layers of each model was unfrozen. Depending on the architecture, the last 20–30 layers were set to be trainable. This step aims to enable the deeper layers of the network to learn fine-grained, emotion-specific features.

The learning rate was reduced (lr = 10^{-4}), and training was continued for 25 epochs. The lower learning rate ensures that the pre-trained weights are updated gradually without disrupting the learned representations. This stage is the primary factor behind the performance improvements observed in both the training curves and quantitative evaluation tables.

In the fine-tuning stage, different depths of the network were selectively unfrozen depending on the architecture. For ResNet50, the final residual blocks corresponding to approximately the last 25 layers were set as trainable. In DenseNet121, the last dense block (approximately 30 layers) was unfrozen to adapt high-level feature representations. For InceptionV3, the upper inception modules responsible for semantic feature extraction were retrained, while earlier modules remained frozen. In MobileNetV2, only the final bottleneck layers were unfrozen to maintain computational efficiency while enabling limited task-specific adaptation. This selective unfreezing strategy allows the models to learn emotion-specific features without disrupting the general representations transferred from ImageNet.

3.4. Performance Metrics

Model performance was evaluated using accuracy, loss, and F1-score metrics. While accuracy measures overall classification performance, the F1-score was employed to mitigate the effects of class imbalance within the dataset. Training and test metrics were reported separately to enable a detailed analysis of model generalization capabilities.

Additionally, for each model, the number of parameters, model size, and training time per epoch on CPU and GPU were measured. This comprehensive evaluation allows for an in-depth analysis of the trade-off between classification performance and computational cost.

4. Results and Discussion

Experiments were conducted on both CPU and GPU environments to analyze execution behavior under different hardware settings. To report statistically meaningful performance metrics such as accuracy, loss, and F1-score, all models were trained 10 times independently on the GPU, and the values presented in Table 2 correspond to the average results obtained from these repetitions. Across these runs, a standard deviation of approximately ±0.8 was observed.

During the GPU-based repetitions, the training–test data split was kept fixed to ensure fair comparability. However, data augmentation procedures and sample ordering (shuffling) were performed using different random seeds for each run. This experimental design allows the robustness of the

models' generalization capability to be assessed against randomness introduced during training, such as stochastic optimization and mini-batch sampling.

The low standard deviations observed in accuracy and F1-score values across the 10 independent GPU runs indicate that the proposed models exhibit stable learning behavior and are not significantly affected by random initialization or data ordering. This finding suggests that the reported performance gains are not coincidental but rather stem from the intrinsic learning capacity of the investigated network architectures. To further quantify statistical reliability, 95% confidence intervals (CI) were computed for the reported performance metrics across the 10 independent experimental runs. The confidence interval was calculated as defined in Equation (2).

$$CI = \bar{x} \pm 1.96 \cdot \frac{s}{\sqrt{n}} \quad (2)$$

where x represents the mean performance value, s is the standard deviation, and n is the number of repeated runs. The resulting confidence intervals were observed to be narrow, indicating that the performance variations across runs are limited and that the comparative results are statistically stable.

In contrast, CPU-based experiments were conducted not for performance metric comparison, but to analyze computational efficiency and training time differences across models. Therefore, each model was executed once in the CPU environment as a representative run. The results obtained from CPU-based experiments were used exclusively for analyzing training duration, whereas classification performance metrics such as accuracy and F1-score were reported based on the averaged GPU results.

All experiments followed a two-stage training strategy. In the first stage, the convolutional layers of the pre-trained networks were frozen, and only the newly added classifier layers were trained (feature extraction). In the second stage, selected upper convolutional blocks were gradually unfrozen and fine-tuned using a lower learning rate. This experimental design enables the performance results reported for Stage 1 and Stage 2 in Table 2 to be interpreted in a consistent and comparable manner, both in terms of hardware configuration and training strategy. All experimental studies were conducted in the environment described in Table 1.

The performance results presented in Table 2 provide a comprehensive basis for the comparative evaluation of four different deep learning architectures on the facial emotion recognition task. When the outcomes obtained during both the Feature Extraction and Fine-Tuning stages are jointly analyzed, it is evident that the DenseNet121 model exhibits the most consistent and superior performance across all evaluation metrics. In particular, the notable improvement in test accuracy and F1-score after fine-tuning indicates that this architecture achieves a more balanced and stable class-wise discrimination. The dense connectivity pattern of DenseNet121 enables effective feature reuse and enhanced gradient flow, allowing the model to represent subtle facial expression variations across different emotion classes more effectively. Precision and recall values were additionally reported to complement the F1-score and to provide a clearer interpretation of class-level performance. The slight differences between these metrics reflect prediction asymmetries that commonly arise in facial emotion recognition due to inter-class similarity and dataset imbalance.

Table 1. Experimental environment specifications

Attribute	Value
Memory	24 GB
GPU	NVIDIA GeForce GTX 1650 4 GB VRAM
CPU	AMD Ryzen 5 5600H 3.3 GHz
Storage	500 GB SSD
Python Version	3.8
TensorFlow Version	2.10
CUDA Version	11.2
cuDNN Version	8.1
Jupyter Notebook Version	6.4.x

Table 2. Performance results (mean ± standard deviation and 95% confidence interval over 10 independent runs)

	Accuracy (Train)	Loss (Train)	F1-Score (Train)	Accuracy (Test)	Loss (Test)	F1-Score (Test)	Precision	Recall
Stage-1 (Frozen Layers)								
ResNet50	0.815	0.722	0.811	0.775	0.785	0.772	-	-
DenseNet121	0.824	0.663	0.832	0.833	0.714	0.793	-	-
InceptionV3	0.793	0.745	0.774	0.781	0.806	0.740	-	-
MobileNetV2	0.756	0.837	0.762	0.712	0.871	0.734	-	-
Stage-2 (Fine-Tuning)								
ResNet50	0.916	0.293	0.925	0.864	0.362	0.897	0.905	0.889
DenseNet121	0.904	0.241	0.918	0.883	0.383	0.911	0.923	0.900
InceptionV3	0.897	0.381	0.896	0.858	0.415	0.863	0.872	0.855
MobileNetV2	0.844	0.582	0.857	0.825	0.566	0.824	0.836	0.813

Based on the comparative results, DenseNet121 achieves the highest overall performance with an accuracy of 0.883, supported by strong precision (0.923) and recall (0.900), indicating a well-balanced capability to correctly identify emotion classes while maintaining low misclassification rates. ResNet50 follows closely, obtaining 0.864 accuracy with precision (0.905) slightly higher than recall (0.889), suggesting confident predictions with a small tendency to miss some samples. InceptionV3 demonstrates moderate performance (accuracy 0.858), where the relatively close precision (0.872) and recall (0.855) values reflect stable yet less discriminative feature learning compared to deeper architectures. MobileNetV2, while computationally lightweight, yields the lowest accuracy (0.825) along with lower recall (0.813), indicating reduced sensitivity to subtle facial variations. Overall, these findings show that densely connected and deeper architectures provide more consistent class-level recognition, whereas lightweight models trade a small amount of predictive performance for efficiency.

The ResNet50 model emerges as the second most successful architecture, closely following DenseNet121 in overall performance. The substantial increase in test accuracy and the marked reduction in loss values after fine-tuning demonstrate the effectiveness of residual connections in learning complex emotion-specific patterns. However, the slightly higher loss values compared to DenseNet121 suggest that ResNet50 may experience relatively greater difficulty in generalizing certain emotion classes. Despite this, the strong accuracy and F1-score results confirm that ResNet50 remains a robust and reliable option for facial emotion recognition tasks.

The InceptionV3 model achieves a moderate level of performance, ranking below DenseNet121 and ResNet50 but above MobileNetV2. The observed improvement in accuracy and F1-score after fine-tuning highlights the advantage of its multi-scale convolutional design, which enables the extraction of facial features at different spatial resolutions. Nevertheless, the relatively higher loss values indicate that InceptionV3 may exhibit instability in learning decision boundaries, potentially due to higher inter-class similarity among facial expressions. This finding suggests that InceptionV3 may benefit more significantly from larger and more diverse datasets.

The MobileNetV2 model yields the lowest accuracy and F1-score among the evaluated architectures; however, its performance remains consistent with expectations given its lightweight design and limited parameter capacity. Considering its compact architecture, the performance improvement observed after fine-tuning is noteworthy. The increase in test accuracy and F1-score demonstrates that MobileNetV2 can still achieve reasonable generalization performance even in low computational cost environments. Nonetheless, its representational capacity appears limited when distinguishing between complex and fine-grained facial emotion patterns, especially compared to deeper and more expressive architectures.

Table 3 presents the normalized confusion matrices averaged across runs, providing a detailed view of class-level prediction behavior for each architecture. Across all models, the Happy class exhibits the highest true positive rates (up to 0.95), indicating that it is the most visually distinctive emotion and can be recognized reliably regardless of net-

Table 3. Normalized confusion matrices (class-wise averages) of the evaluated models

Actual \ Predicted	Angry	Fear	Happy	Sad	Surprise
ResNet50					
Angry	0.86	0.04	0.02	0.07	0.01
Fear	0.04	0.82	0.02	0.03	0.09
Happy	0.02	0.01	0.95	0.01	0.01
Sad	0.08	0.02	0.02	0.85	0.03
Surprise	0.02	0.10	0.02	0.02	0.84
DenseNet121					
Angry	0.88	0.03	0.02	0.06	0.01
Fear	0.04	0.84	0.01	0.03	0.08
Happy	0.02	0.01	0.95	0.01	0.01
Sad	0.07	0.03	0.02	0.86	0.02
Surprise	0.01	0.07	0.02	0.02	0.88
InceptionV3					
Angry	0.86	0.04	0.02	0.07	0.01
Fear	0.06	0.79	0.03	0.04	0.08
Happy	0.04	0.01	0.92	0.02	0.01
Sad	0.08	0.03	0.02	0.85	0.02
Surprise	0.01	0.09	0.01	0.02	0.87
MobileNetV2					
Angry	0.80	0.07	0.04	0.08	0.01
Fear	0.08	0.76	0.02	0.03	0.09
Happy	0.04	0.01	0.92	0.02	0.01
Sad	0.10	0.04	0.03	0.81	0.02
Surprise	0.02	0.11	0.02	0.02	0.83

work structure. The most frequent misclassification occurs between Fear and Surprise, where prediction leakage is consistently observed, reflecting the similarity of facial muscle activations associated with these expressions. A secondary confusion pattern appears between Angry and Sad, particularly in MobileNetV2, suggesting that lightweight architectures have more difficulty capturing subtle variations in brow and mouth regions. Among the evaluated models, DenseNet121 demonstrates the strongest diagonal dominance, confirming its more consistent class discrimination capability. In contrast, MobileNetV2 shows comparatively higher off-diagonal values, indicating reduced sensitivity due to its compact design. These observations support the quantitative results by illustrating how deeper and densely connected architectures achieve more stable emotion separation, whereas lightweight models provide efficiency at the cost of slightly increased class overlap.

Overall, the results presented in Table 2 and the corresponding performance trends illustrated in Figure 3 indicate that both model depth and connectivity structure play a critical role in facial emotion recognition performance. Furthermore, the fine-tuning strategy proves to be indispensable across all evaluated architectures. While DenseNet121 and ResNet50 are particularly well-suited for scenarios requiring high accuracy and strong generalization capability, MobileNetV2 offers an effective alternative for resource-constrained applications by balancing computational efficiency and acceptable performance. These comparative findings numerically support the graphical analyses and demonstrate that the performance values reported in Table 2 are consistent with the theoretical characteristics of the respective model architectures.

Figure 3. Performance comparison of deep learning architectures.

The results presented in Table 4, which include training time, number of parameters, and model size, enable a comprehensive evaluation of the four deep learning architectures not only in terms of classification performance but also with respect to computational cost and resource requirements. In particular, the training time measured per epoch on both CPU and GPU platforms clearly reveals the impact of hardware acceleration on different model architectures.

The ResNet50 and DenseNet121 models exhibit the longest training times on the CPU due to their deep architectural structures and relatively high parameter counts. This behavior can be attributed to the large number of convolutional layers and the intensive matrix multiplications performed at each layer, which are executed with limited parallelism on CPU hardware. In contrast, a substantial reduction in training time is observed for both models when executed on the GPU, highlighting the indispensable role of CUDA-based parallel computation for deep neural networks. Despite offering higher parameter efficiency than ResNet50, DenseNet121 still imposes a relatively high computational load even on the GPU, primarily due to its densely connected layer structure.

The InceptionV3 model, although comparable to ResNet50 in terms of parameter count, exhibits a distinct computational profile as a result of its multi-scale convolutional blocks and branched architecture. The relatively long training time on the CPU suggests that the branched structure increases sequential computation overhead. On the GPU, however, the benefits of parallel processing significantly reduce the training time, although it remains higher than that of lightweight architectures such as MobileNetV2.

The MobileNetV2 model clearly stands out in Table 4 with the lowest number of parameters and the smallest model size. Thanks to its depthwise separable convolution design, it achieves the shortest per-epoch training times on both CPU and GPU platforms. This characteristic makes MobileNetV2 particularly suitable for resource-constrained environments and real-time applications. Nevertheless, when evaluated together with the performance results in Table 2, it becomes evident that the reduced computational cost comes at the expense of a certain loss in classification accuracy. Therefore, MobileNetV2 represents a favorable option for applications that prioritize efficiency–performance trade-offs rather than maximum accuracy.

From the perspective of model size, a direct relationship between parameter count and storage requirements can be observed. The larger model sizes of DenseNet121 and ResNet50 increase memory demands during deployment, which may pose limitations in memory-constrained systems. Conversely, the compact model size of MobileNetV2 provides a significant advantage for embedded systems and mobile devices. InceptionV3 occupies a middle position in terms of model size, offering a balanced compromise between performance and storage requirements.

Overall, Table 4 clearly demonstrates that the model achieving the highest classification accuracy is not necessarily the most computationally efficient, and similarly, the fastest and smallest model does not yield the highest recognition performance. In this context, DenseNet121 and ResNet50 are well-suited for academic and industrial applications where high accuracy is the primary objective, whereas MobileNetV2 emerges as a rational choice for scenarios with

Table 4. Comparison of deep learning architectures based on experimental results

	Training Time (1 Epoch, CPU) - (Min.)	Training Time (1 Epoch, GPU) - (Min.)	Number of Parameters (Mil.)	Model Size (MB)	Accuracy (test) %
ResNet50	42.9	6.2	25.6	98	0.864
DenseNet121	49.3	8.1	7.9	33	0.883
InceptionV3	55.7	7.8	23.9	92	0.858
MobileNetV2	14.2	2.8	3.4	14	0.825

strict latency and hardware constraints. InceptionV3 can be regarded as an intermediate alternative that balances computational cost and classification performance.

In this study, the emotion recognition performances of ResNet50, DenseNet121, InceptionV3, and MobileNetV2 were investigated within the framework of a two-stage training strategy. In the first stage, the convolutional layers of the pre-trained networks were frozen (Stage-1: Frozen Layers), and only the top classifier layers were trained. In the second stage, a selected subset of deeper layers was unfrozen (Stage-2: Fine-Tuning) to enhance the model's dataset-specific representational capacity. During Stage-1, the convolutional backbones pre-trained on ImageNet were kept fixed for all models, and only the newly added fully connected (Dense) layers were optimized. The accuracy and loss curves obtained in this stage reflect the extent to which the general visual representations learned from ImageNet can be transferred to the facial emotion recognition task. When the accuracy and loss curves plotted in the Python environment are jointly analyzed together with the quantitative results reported in Table 2, it becomes evident that the training dynamics and performance improvements of each model are closely linked to their architectural characteristics.

Unlike many prior studies that report only accuracy-based comparisons, this work jointly evaluates recognition performance and computational characteristics, including parameter count, training duration, and model size. Such an analysis enables a more realistic interpretation of model suitability for deployment scenarios, revealing that the most accurate architecture is not necessarily the most efficient, and vice versa. This perspective contributes to bridging the gap between algorithmic evaluation and practical implementation.

In facial emotion recognition, certain emotion classes—such as fear and surprise—exhibit visually similar facial patterns, which makes class separation particularly challenging. This issue becomes more pronounced in architectures with limited representational capacity, leading to relatively lower accuracy and F1-score values for such models.

The limited gap observed between training and test metrics, particularly during the fine-tuning phase, indicates that the models improve their generalization capability without exhibiting overfitting behavior. This observation confirms that the adopted two-stage training strategy is well-suited for the facial emotion recognition problem.

For ResNet50, the Stage-1 training curves presented in Figure 4 show a stable but limited increase in accuracy. The training accuracy converges at approximately 81.5%, while the test accuracy stabilizes around 77.5%. During this phase, the loss curves remain relatively high (training loss ≈ 0.72, test loss ≈ 0.78), suggesting that the general-purpose features extracted by the frozen deep layers are insufficient for capturing the fine-grained patterns required for emotion recognition. The plateau observed in the accuracy curve further supports the conclusion that the model struggles to discriminate subtle facial expression variations when deeper layers are not adapted to the target domain.

In contrast, the Stage-2 fine-tuning phase yields a substantial performance improvement, as clearly reflected both in the training curves and in Table 2. The training accuracy increases to 91.6%, while the test accuracy reaches 86.4%, accompanied by a sharp reduction in loss values (training loss ≈ 0.29, test loss ≈ 0.36). These results indicate that re-optimizing the deeper layers through residual connections provides a significant representational advantage. Moreover, the smoother and more stable behavior of the accuracy–loss curves during this stage demonstrates that ResNet50 achieves a more robust and well-generalized learning behavior after fine-tuning.

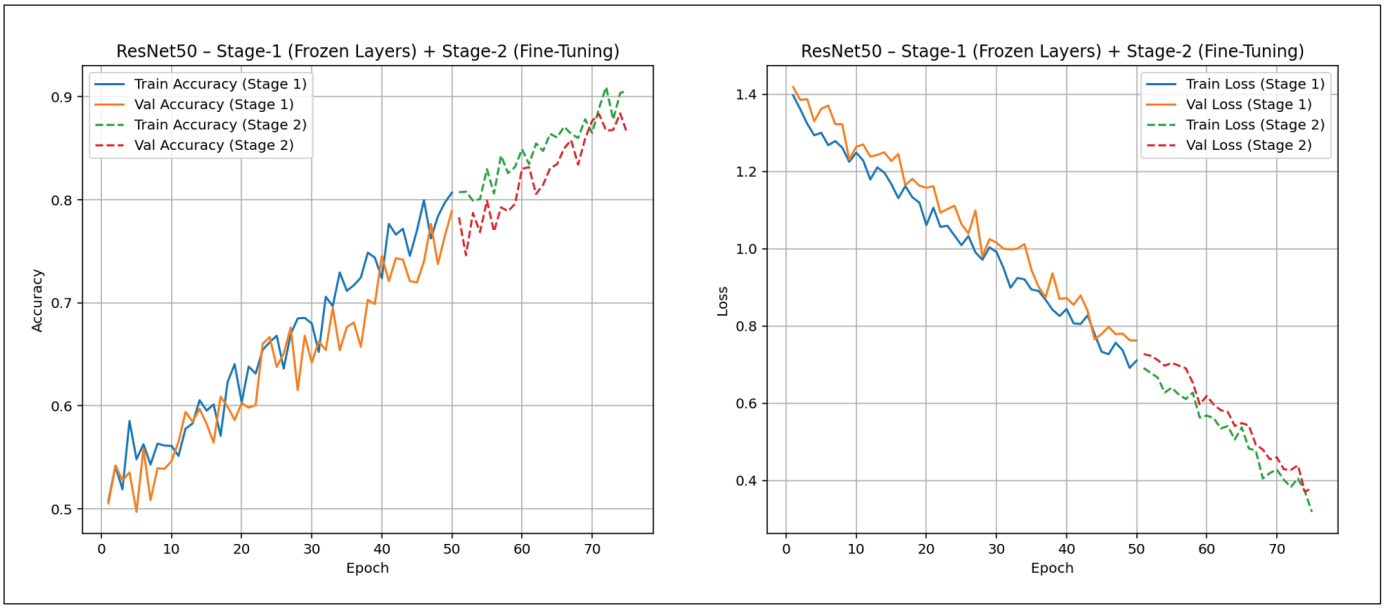

Figure 4. Graphical representation of the experimental results obtained with the ResNet50 architecture.

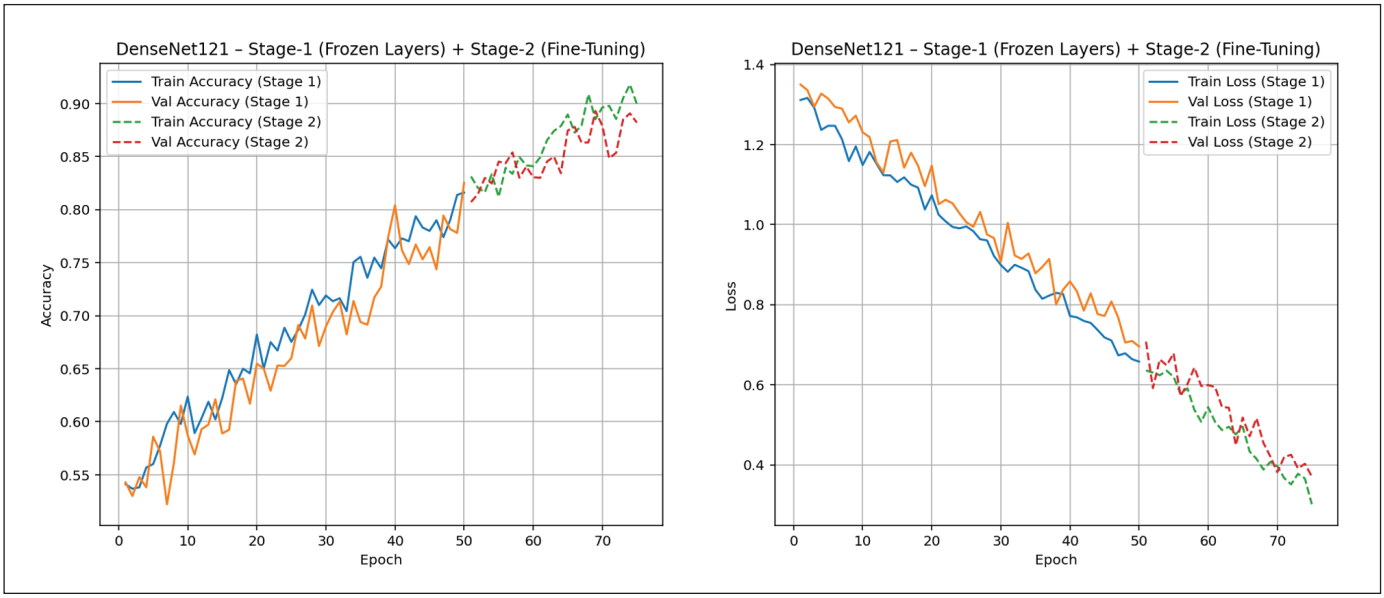

Figure 5. Graphical representation of the experimental results obtained with the DenseNet121 architecture.

An examination of the DenseNet121 curves presented in Figure 5 reveals that this model exhibits a higher and more stable accuracy profile than the other architectures, even during the Stage-1 phase. The training accuracy reaches 82.4%, while the test accuracy attains 83.3%, accompanied by relatively lower loss values (training loss ≈ 0.66, test loss ≈ 0.71). These results indicate that the dense connectivity mechanism enables effective feature transfer even at early training stages. The reduced oscillations observed in the accuracy curves further demonstrate that DenseNet's feature reuse strategy contributes to stabilizing the learning process and facilitates more consistent convergence behavior.

During the Stage-2 fine-tuning phase, DenseNet121 achieves the highest overall performance among all evaluated models. The test accuracy increases to 88.3%, while the test F1-score reaches 91.1%, and the loss values decrease to approximately 0.38. These results indicate that DenseNet121 is particularly effective at learning fine-grained dis-

tinctions between emotion classes. Moreover, the limited discrepancy between the training and test curves suggests a relatively low risk of overfitting, clearly highlighting the strong generalization capability of the DenseNet121 architecture.

An examination of the graphs presented in Figure 6 indicates that InceptionV3 exhibits a more fluctuating accuracy trajectory during the Stage-1 training phase compared to the other models. The training accuracy stabilizes at approximately 79.3%, while the test accuracy remains around 78.1%, accompanied by a test loss of approximately 0.81. These observations suggest that the multi-scale filtering structure of InceptionV3 may experience instability when discriminating between visually similar facial expressions. The noticeable oscillations in the accuracy curves imply that the model encounters difficulties in consistently distinguishing certain emotion classes across epochs.

During the fine-tuning stage, the performance of InceptionV3 improves significantly. The test accuracy increases to 85.8%, and the test F1-score reaches 86.3%. The more regular and steadily decreasing loss curves observed in this phase highlight the positive impact of adapting deeper network layers to the target dataset. Nevertheless, when compared to DenseNet121 and ResNet50, InceptionV3 achieves slightly lower accuracy levels, suggesting that this architecture may require greater data diversity or a longer fine-tuning duration to reach its optimal performance on the given emotion recognition dataset.

An analysis of the graphs presented in Figure 7 shows that MobileNetV2, due to its lightweight architectural design, produces comparatively lower accuracy values during the Stage-1 training phase. The training accuracy remains at approximately 75.6%, while the test accuracy stabilizes around 71.2%, accompanied by relatively high loss values (test loss ≈ 0.87). These results indicate that the limited number of parameters constrains the model's ability to capture complex and fine-grained emotional patterns in facial expressions. The early plateau observed in the accuracy curves further supports the notion that the representational capacity of MobileNetV2 is restricted during the feature extraction stage.

In contrast, the Stage-2 fine-tuning phase yields a substantial performance improvement for MobileNetV2. The test accuracy increases to 82.5%, while the test F1-score reaches 82.4%, alongside a notable reduction in loss values to approximately 0.56. These improvements demonstrate that selectively unfreezing deeper layers enables the model to learn more discriminative emotion-specific features. The smoother and more stable training curves observed during this phase indicate that, despite its low computational complexity, MobileNetV2 is capable of achieving a satisfactory level of generalization performance. Consequently, MobileNetV2 emerges as a viable alternative for emotion recognition tasks in resource-constrained environments, where computational efficiency is a critical requirement.

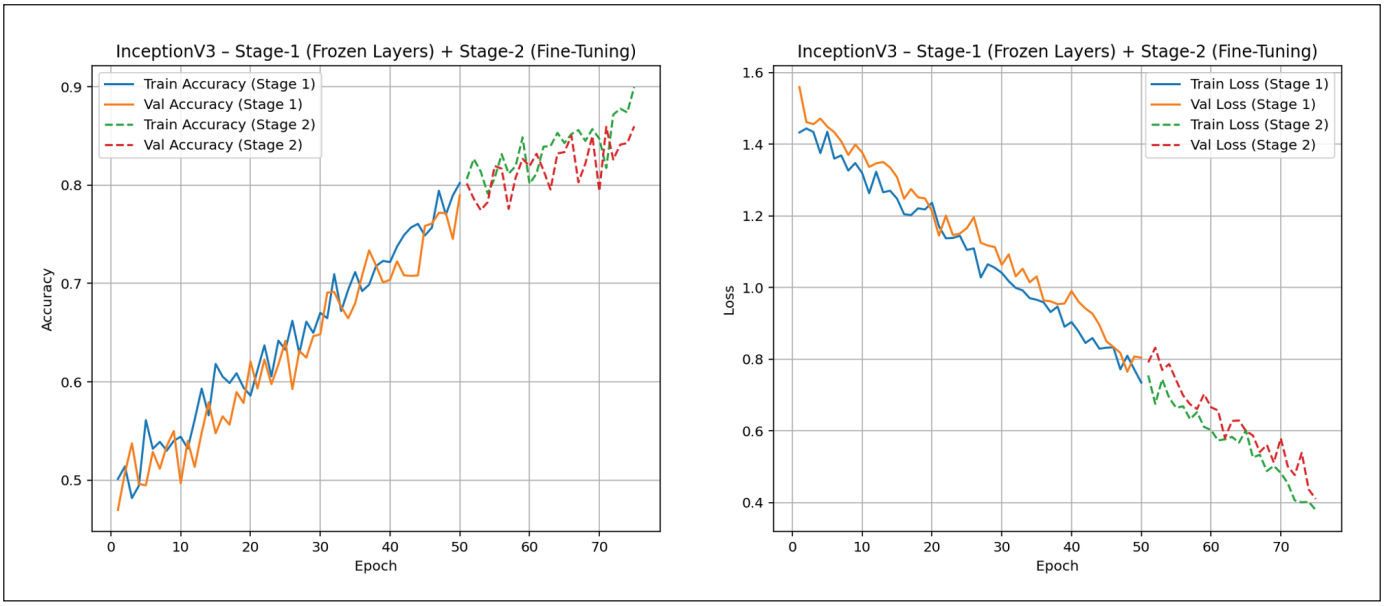

Figure 6. Graphical representation of the experimental results obtained with the InceptionV3 architecture.

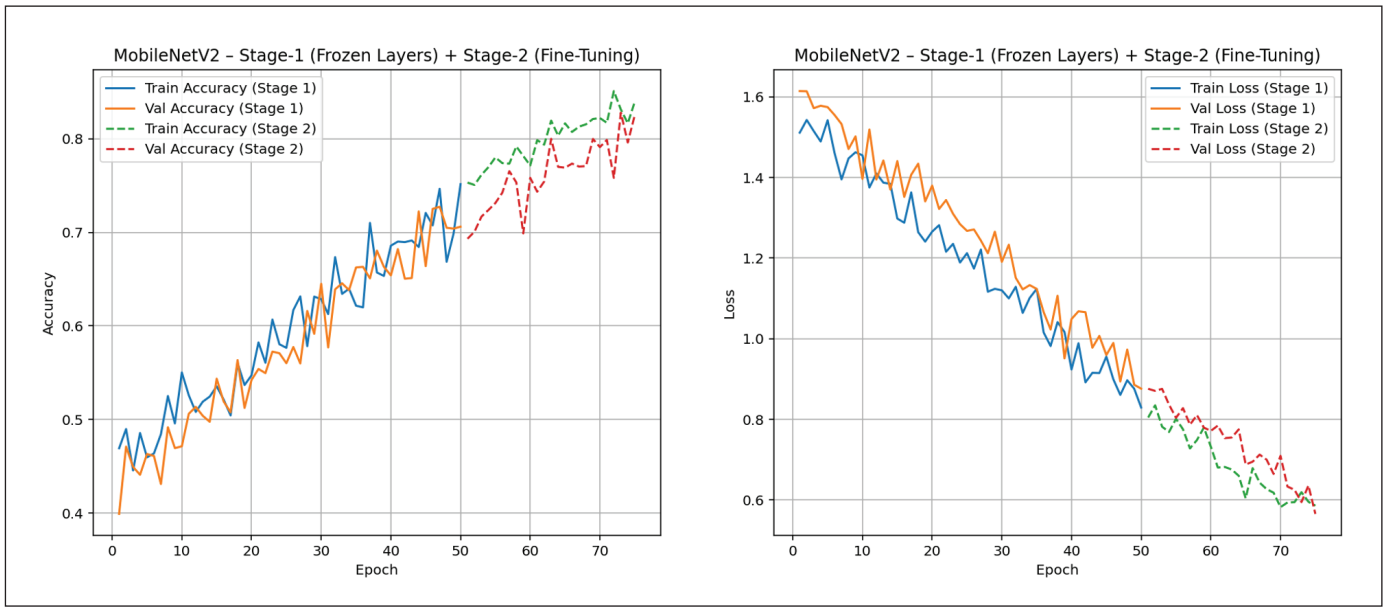

Figure 7. Experimental performance curves of the MobileNetV2 architecture.

Table 5. Comparison with facial emotion recognition studies in literature

Ref	Techniques	Accuracy/result (%)
Li et al., 2021	Dual-Branch Center Face detector (DBCFace)	90.34
Pranav et al., 2020	DCNN	78.04
Verma et al., 2019	CNN, DNNs	86.73
Yang et al., 2020	Feature separation model exchange-GAN	86.33
Zhang et al., 2021	GAN	89.01
Ali et al., 2023	Ensemble Classifier	83.19
This Study	ResNet50	86.4
	DenseNet121	88.3
	InceptionV3	85.8
	MobileNetV2	82.5

When the accuracy and loss curves plotted for all models are jointly examined, it becomes evident that Stage-1 represents an initial feature representation phase, whereas Stage-2 constitutes the primary learning phase. The simultaneous increase in accuracy and decrease in loss during Stage-2 clearly indicate that the fine-tuning process contributes not only to memorization but also to improved generalization capability.

Following Stage-2, DenseNet121 and ResNet50 distinctly outperform the other models in terms of both accuracy and F1-score, while MobileNetV2, despite its lower absolute performance, provides a significant balance between classification performance and computational efficiency. InceptionV3, on the other hand, achieves moderate accuracy levels but exhibits a more fluctuating training behavior compared to the other architectures.

Table 5 presents a comparative evaluation between the proposed approach and recent studies in the literature on facial emotion recognition. Since the majority of existing works report only classification accuracy as their primary performance indicator, accuracy was selected as the common reference metric to enable a fair comparison. As observed, previously reported accuracy values range between approximately 78% and 90%, depending on the model architecture and learning strategy employed. The results obtained in this study fall within this performance band, with DenseNet121

achieving 88.3% accuracy and ResNet50 reaching 86.4%, demonstrating competitive performance relative to state-of-the-art approaches. These findings indicate that the proposed two-stage transfer learning framework yields results consistent with contemporary deep learning-based FER methods while additionally providing insights into computational efficiency and model complexity, which are often not discussed in earlier studies.

Overall, these graphical results clearly demonstrate that architectural complexity, parameter count, and fine-tuning strategies directly influence performance in fine-grained classification tasks such as facial emotion recognition. The use of a single dataset and the inherent visual similarity among certain emotion classes may have constrained the performance of some models. Additional experiments conducted on different datasets would further enhance the generalizability of the findings.

5. Conclusion

In this study, four widely used deep learning architectures—ResNet50, DenseNet121, InceptionV3, and MobileNetV2—were comprehensively evaluated for facial emotion recognition within a two-stage transfer learning framework. The primary objective was not only to assess classification accuracy but also to jointly analyze training dynamics, generalization capability, and computational cost, thereby enabling a more holistic evaluation of architectural choices for emotion analysis applications.

The experimental results clearly demonstrate that the two-stage training strategy yields a significant performance improvement across all models. During the first stage (Stage-1), where the convolutional backbones were frozen, the pretrained networks provided stable but limited performance, indicating that generic visual representations alone are insufficient for fine-grained emotion classification. In contrast, the Stage-2 fine-tuning phase resulted in substantial improvements in both accuracy and F1-score, accompanied by consistent reductions in loss values. These findings highlight the critical role of adapting deep feature representations to emotion-specific patterns for capturing subtle facial expression variations.

Among the evaluated models, DenseNet121 emerged as the most balanced and highest-performing architecture in both training and testing phases. Its dense connectivity facilitates efficient feature reuse, enabling more stable learning and robust performance, particularly in scenarios involving class imbalance. ResNet50 closely followed DenseNet121, delivering high accuracy and strong generalization performance, thereby reaffirming the effectiveness of residual connections in stabilizing deep network training.

Although InceptionV3 benefits from its multi-scale convolutional design, it underperformed compared to DenseNet121 and ResNet50 on this dataset. The observed fluctuations in accuracy and loss curves suggest that the architecture may be more sensitive to visual similarities among emotion classes. MobileNetV2, while yielding the lowest absolute performance, stands out due to its low parameter count, compact model size, and short training time. These characteristics make it a compelling choice for real-time and resource-constrained applications, where computational efficiency is a primary concern.

When the experimental findings, performance tables, and training curves are jointly considered, it becomes evident that model performance should not be evaluated solely based on accuracy. Training time, parameter count, and hardware requirements must also be taken into account. The multiple GPU-based experimental repetitions confirmed the stability and robustness of the reported results, while the representative CPU-based experiments clearly demonstrated the importance of hardware acceleration for deep learning-based emotion recognition.

This study also has certain limitations. First, the experiments were conducted on a single dataset, and no cross-dataset validation was performed to assess robustness across different cultural and demographic contexts. Additionally, data augmentation techniques were not employed, and the analysis focused exclusively on static facial images, excluding temporal information from video sequences.

Another limitation is the restricted number of emotion categories in the dataset. Future work will extend the analysis to benchmark datasets containing neutral and additional expressions to further validate the findings.

It should be noted that most studies in the literature report results using different datasets, preprocessing pipelines, and evaluation protocols, making direct comparison inherently challenging. Therefore, the purpose of Table 5 is not to claim superiority over prior works, but rather to position the obtained results within the commonly reported performance range. In addition to this contextualization, the present study aims to offer an alternative evaluation perspective to existing approaches by jointly analyzing recognition performance together with training cost, parameter size, and computational behavior—factors that are often overlooked

in earlier studies. This comparison demonstrates that the proposed framework achieves competitive accuracy while providing a more comprehensive understanding of model selection criteria for practical applications. Future work will extend this evaluation to multiple benchmark datasets to further validate generalizability.

Future research may extend this work by conducting cross-dataset evaluations to further assess generalization performance. Incorporating multimodal cues such as head pose, eye movements, or facial muscle activations may further enhance recognition accuracy. Moreover, exploring Vision Transformers (ViT) and hybrid CNN–Transformer architectures could provide valuable insights into architectural advancements for emotion analysis. Finally, investigating model compression, quantization, and knowledge distillation techniques would significantly broaden the applicability of these models in real-time and embedded systems.

Recent studies have explored transformer-based architectures such as Vision Transformers (ViT) for facial emotion recognition, offering an alternative to convolutional models through global attention mechanisms. While such architectures may provide advantages in large-scale datasets, the present study focuses on widely adopted CNN backbones to analyze performance–efficiency trade-offs under limited-resolution conditions. Future work may extend this framework to include transformer-based models and cross-dataset validation to further examine generalizability. Additionally, visualizing computational cost–performance relationships using Pareto-style analysis could provide further insights into model selection for deployment-oriented scenarios.

In conclusion, this study presents a comprehensive and comparative analysis of deep learning architectures for facial emotion recognition under a two-stage transfer learning paradigm. By explicitly addressing the trade-off between classification performance and computational cost, the findings offer valuable guidance for both academic research and real-world deployment scenarios.

Acknowledgement: The authors acknowledge the use of Grammarly, Instatext, and generative artificial intelligence tools during the language editing process of this manuscript. In addition, the authors would like to thank Andrea Piacquadio for her contributions to the creation of the dataset used in this study.

Author contributions: All authors contributed equally to the preparation of this study.

Declaration of ethical code: All procedures conducted in this study comply with the principles and regulations stated in the Higher Education Institutions Scientific Research and Publication Ethics Directive. None of the actions defined under the title Actions Contrary to Scientific Research and Publication Ethics in the directive were committed during the preparation of this study.

Ethics committee approval: This study does not involve any ethical violations and does not require ethics committee approval.

Conflicts of interest: The authors declare that there is no conflict of interest regarding the publication of this study.

References

Ali, G., Ali, A., Ali, F., Draz, U., Majeed, F., Yasin, S., & Haider, N. (2023). Artificial neural network-based ensemble approach for multicultural facial expressions analysis. *IEEE Access, 8,* 134950–63.

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1(1),* 18–37. https://doi.org/10.1109/T-AFFC.2010.1

Darraz, N., Karabila, I., El-Ansari, A., Alami, N., & El Mallahi, M. (2025). Integrated sentiment analysis with BERT for enhanced hybrid recommendation systems. *Expert Systems with Applications, 261,* Article 125533. https://doi.org/10.1016/j.eswa.2024.125533

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17(2),* 124–129. https://doi.org/10.1037/h0030377

Gao, X., Liu, F., Zhuang, X., Zhang, Y., Tian, X., Dong, Y., & Yang, H. (2026). A unified dual-view knowledge-guided sentiment interaction networks for aspect-based sentiment analysis. *Neural Networks, 194,* Article 108159. https://doi.org/10.1016/j.neunet.2025.108159

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press, 800 pp.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778, USA. https://doi.org/10.1109/CVPR.2016.90

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708, USA. https://doi.org/10.1109/CVPR.2017.243

Huang, X., Sun, H., Liu, X., Wu, J., & He, L. (2026). Text-dominant speech-enhanced network for multimodal aspect-based sentiment analysis. *Information Fusion, 126*, Article 103543. https://doi.org/10.1016/j.inffus.2025.103543

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521(7553),* 436–444. https://doi.org/10.1038/nature14539

Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 13(3),* 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446

Li, X., Lai, S., & Qian, X. (2021). DBCFace: towards pure convolutional neural network face detection. I*EEE Transactions on Circuits and Systems for Video Technology*, *32(4),* 1792-1804,. https://doi.org/10.1109/TCSVT.2021.3082635

Li, Z., Huang, Z., Pan, Y., Yu, J., Liu, W., Chen, H., & Wang, H. (2024). Hierarchical denoising representation disentanglement and dual-channel cross-modal-context interaction for multimodal sentiment analysis. *Expert Systems with Applications, 252(B),* Article 124236. https://doi.org/10.1016/j.eswa.2024.124236

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10(1),* 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

Mughal, N., Mujtaba, G., Shaikh, S., Kumar, A., & Daudpota, S. M. (2024). Comparative analysis of deep neural networks and large language models for aspect-based sentiment analysis. *IEEE Access, 12,* 60943–60959. https://doi.org/10.1109/ACCESS.2024.3386969

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22(10)*, 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Pandey, A., & Vishwakarma, D. K. (2024). Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey. *Applied Soft Computing, 152,* Article 111206. https://doi.org/10.1016/j.asoc.2023.111206

Pranav, E., Kamal, S., Chandran, C. S., & Supriya, M. H. (2020). Facial emotion recognition using deep convolutional neural network. 6th International conference on advanced computing and communication Systems (ICACCS), IEEE. pp. 317–20, Indiana.

Ruan, S., Zhang, K., Wu, L., Xu, T., Liu, Q., & Chen, E. (2024). Color enhanced cross correlation net for image sentiment analysis. *IEEE Transactions on Multimedia, 26*, 4097–4109. https://doi.org/10.1109/TMM.2021.3118208

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520, USA. https://doi.org/10.1109/CVPR.2018.00474

Song, L., Chen, S., Meng, Z., Sun, M., & Shang, X. (2024). FMSA-SC: A fine-grained multimodal sentiment analysis dataset based on stock comment videos. *IEEE Transactions on Multimedia, 26,* 7294–7306. https://doi.org/10.1109/TMM.2024.3363641

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2818–2826, USA. https://doi.org/10.1109/CVPR.2016.308

Verma, A., Singh, P., & Alex, J. S. R. (2019). Modified convolutional neural network architecture analysis for facial emotion recognition. 26th International Conference on Systems, Signals and Image Processing (IWSSIP 2019), pp. 169–173, Croatia.

Yang, L., Tian, Y., Song, Y., Yang, N., Ma, K., & Xie, L. (2020). A novel feature separation model exchange-GAN for facial expression recognition. *Knowledge-Based Systems, 204,* 106217.

Yang, J., Xiao, Y., & Du, X. (2024). Multi-grained fusion network with self-distillation for aspect-based multimodal sentiment analysis. *Knowledge-Based Systems, 293,* 111724. https://doi.org/10.1016/j.knosys.2024.111724

Zhang, X., Zhang, F., & Xu, C. (2021). Joint expression synthesis and representation learning for facial expression recognition. *IEEE Trans Circuits Syst Video Technol., 32(3),* 1681–95.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1),* 39–58. https://doi.org/10.1109/TPAMI.2008.52