International Journal of Computer Techniques–IJCT Volume 13 Issue 2, March-April - 2026

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

# An Empirical Study of Pre-trained CNN Models for Multiclass Emotion Recognition

**Dr. Megha Bansal[1], Ms. Tanvi Dalal[2], Dr. Mitanshi Rastogi[3], Dr. Neha Goel[4]**

**[1,2,3] Assistant Professor, Associate Professor, VSIT, Vivekananda Institute of Professional Studies-Technical Campus, India**

## Abstract

Emotion recognition from facial expressions has emerged as a significant research area in computer vision and affective computing due to its wide range of applications in healthcare, human–computer interaction, surveillance, and intelligent systems. Recent advancements in deep learning have demonstrated that pre-trained Convolutional Neural Network (CNN) models can effectively extract discriminative features for emotion classification tasks. This study presents an empirical evaluation of multiple pre-trained CNN architectures for multiclass facial emotion recognition. In this work, widely adopted deep learning models are fine-tuned using transfer learning to classify facial images into seven fundamental emotion categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. A standardized experimental framework is employed, including uniform preprocessing, data augmentation, and hyperparameter settings to ensure fair comparison. Model performance is evaluated using accuracy, precision, recall, F1-score, confusion matrix analysis, and computational efficiency metrics such as training time and parameter complexity. The experimental findings reveal notable variations in classification performance and computational cost among the evaluated architectures. While deeper networks demonstrate strong feature extraction capability, optimized models provide a superior balance between accuracy and efficiency. The results highlight the effectiveness of transfer learning in improving multiclass emotion recognition performance, particularly when training data is limited. This study provides practical insights into selecting suitable pre-trained CNN models for robust and scalable emotion recognition systems and contributes to the development of efficient deep learning solutions for real-world affective computing applications.

**Keywords:** Multiclass Emotion Recognition, CNN, FER

## 1. INTRODUCTION

Facial emotion recognition (FER) is a fundamental problem in computer vision and affective computing, focusing on the automatic identification of human emotional states from facial expressions. The ability to recognize emotions accurately has wide-ranging applications in areas such as mental healthcare monitoring, intelligent tutoring systems, human–computer interaction, social robotics, and surveillance [1], [2]. Traditional approaches to FER relied on handcrafted features such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), Gabor filters, coupled with conventional classifiers like Support Vector Machines

(SVM) [3]. However, these techniques encounter limitations in handling complex variations in pose, illumination, and expression subtleties.

The advent of deep learning, particularly Convolutional Neural Networks (CNNs), revolutionized visual recognition tasks by enabling hierarchical feature learning directly from raw images [4]. CNNs have been shown to outperform traditional methods in diverse computer vision domains due to their ability to model spatial hierarchies and free parameters optimized via backpropagation [5]. Training such networks from scratch for FER, however, demands large annotated datasets and substantial computational resources — requirements often unmet in practical FER datasets with limited samples [6].

To mitigate data scarcity and expedite training, transfer learning has emerged as a powerful paradigm. In transfer learning, models pre-trained on large benchmark datasets such as ImageNet are fine-tuned on domain-specific tasks, yielding improved performance with reduced training time [7],[8]. Pre-trained CNN architectures such as VGGNet, ResNet, and EfficientNet have been successfully adapted to various image classification problems and are increasingly applied in FER [9] [10] [11]. Despite the widespread use of these pre-trained models, a systematic empirical comparison of their performance specifically for multiclass emotion recognition remains limited.

Most existing studies focus on applying individual architectures without consistent comparative evaluation under standardized experimental frameworks [12]. Moreover, differences in dataset splits, preprocessing steps, augmentation strategies, and evaluation metrics make direct comparison challenging. There is a practical need for an empirical study that assesses pre-trained CNN models holistically — considering not only classification accuracy but also computational efficiency and scalability.

This work aims to fill this gap by conducting an empirical evaluation of multiple pre-trained CNN models for multiclass facial emotion recognition. The study employs a uniform experimental protocol with standardized preprocessing, augmentation, and evaluation criteria to ensure fair comparison across models. Performance is measured using accuracy, precision, recall, F1-score, and confusion matrices, along with assessments of training time and model complexity.

The contributions of this work are summarized as follows:

1. A structured comparative evaluation of widely used pre-trained CNN models for multiclass FER.
2. Insights into performance–efficiency trade-offs among CNN architectures.
3. Practical recommendations for selecting optimal pre-trained models for real-world affective computing tasks.

## 2. RELATED WORK

Facial emotion recognition (FER) has been extensively studied over the past two decades, with research evolving from traditional machine learning methods to deep convolutional neural networks (CNNs). Early works focused on handcrafted feature extraction algorithms such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Gabor wavelets, combined with conventional classifiers (e.g., Support Vector Machines) for emotion classification [13], [14]. While these techniques provided initial benchmarks for FER, they often struggled with generalization due to variations in lighting, pose, and facial occlusions.

The advent of deep learning considerably shifted focus toward CNN-based representation learning. Convolutional neural networks automatically learn hierarchical features from raw images, significantly outperforming traditional handcrafted methods. In their pioneering work, Krizhevsky *et al.* demonstrated the superiority of deep CNNs on large-scale image classification tasks, laying the foundation for CNN adoption in FER [15]. Subsequently, researchers began adopting deep architectures in expression recognition tasks; for instance, Mollahosseini *et al.* fine-tuned CNN models on multiple FER datasets to achieve state-of-the-art performance [16].

Transfer learning has been widely leveraged to overcome data scarcity in FER. Pre-trained CNNs on large-scale datasets such as ImageNet can be adapted to emotion recognition with fine-tuning, significantly reducing training time and improving accuracy [17]. Zhang *et al.* conducted comparative analysis using pre-trained models such as VGGNet and ResNet on benchmark FER datasets, highlighting their transfer learning capabilities [18]. Similarly, Li *et al.* incorporated residual connections from ResNet to improve expression recognition performance in unconstrained environments [19].

EfficientNet, a more recent architecture, introduced compound scaling to balance network depth, width, and resolution, achieving competitive performance with fewer parameters [20]. Dhall *et al.* explored the application of EfficientNet in affective computing tasks and reported promising results, particularly in resource-constrained settings [21]. These studies emphasize the potential of optimized pre-trained models in FER tasks, though they often focus on individual architectures rather than systematic comparison.

Beyond single-architecture studies, several researchers have investigated architectural enhancements to improve FER performance. Zhang and Zhang introduced attention mechanisms into ResNet architectures to emphasize salient facial regions, resulting in improved discrimination for similar expressions [22]. Other work has explored hybrid models combining CNNs with recurrent neural networks (RNNs) to capture temporal information in dynamic FER [23].

Despite the growing literature on deep learning–based FER and transfer learning applications, there is a lack of empirical studies that comprehensively compare multiple pre-trained CNN models under uniform experimental frameworks. Variations in dataset preprocessing, augmentation strategies, and evaluation methods across studies often make performance comparisons difficult. This motivates the need for a structured empirical analysis of pre-trained CNN models tailored for multiclass emotion recognition, as conducted in this work.

## 3. METHODOLOGY

This section presents the experimental framework adopted for conducting an empirical evaluation of pre-trained Convolutional Neural Network (CNN) models for multiclass facial emotion recognition. The methodology includes dataset selection, preprocessing, model adaptation using transfer learning, training configuration, and performance evaluation under a unified experimental setup.

### 3.1 Dataset Description

The experiments are performed using the FER-2013 (Facial Expression Recognition 2013) dataset, a benchmark dataset widely used in emotion recognition research. The FER-2013 dataset consists of 35,887 grayscale facial images of resolution 48 × 48 pixels, categorized into seven emotion classes: anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset is divided into training, public test (validation), and private test sets, ensuring structured performance evaluation. The images in FER-2013 are collected from real-world scenarios and exhibit significant variations in facial pose, illumination, occlusion, and expression intensity, making the dataset suitable for evaluating model robustness and generalization capability.

### 3.2 Data Preprocessing and Augmentation

To ensure compatibility with pre-trained CNN architectures, all images are resized from 48 × 48 pixels to 224 × 224 pixels, which matches the standard input dimensions required by most ImageNet-trained models. Since the original images are grayscale and pre-trained networks expect three-channel RGB inputs, grayscale images are converted into three-channel format by channel replication. Pixel values are normalized to the range [0, 1] to stabilize training and improve convergence.

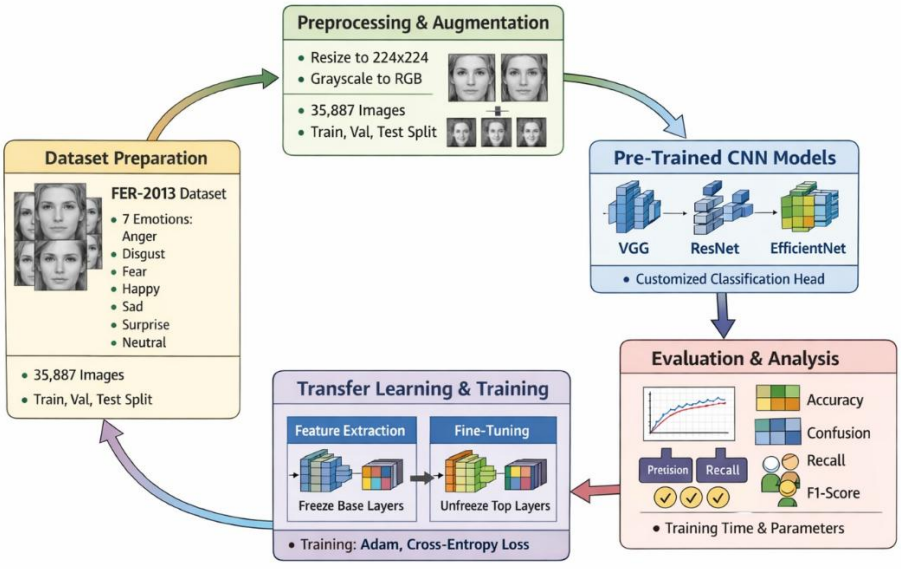


Fig.1: Empirical Study of Pre-Trained CNN Models for Facial Emotion Recognition

To enhance generalization and reduce overfitting, data augmentation techniques are applied during the training phase. These include horizontal flipping, random rotations within a

limited angle range, width and height shifting, zooming, and brightness adjustments. Data augmentation increases variability in the training data and improves the model's ability to handle real-world variations in facial expressions.

### 3.3 Pre-trained CNN Model Selection and Modification

This study evaluates multiple widely adopted pre-trained CNN architectures, including VGG-based, ResNet-based, and EfficientNet-based models. Each model is initialized with weights pre-trained on the ImageNet dataset to leverage previously learned low-level and mid-level visual features. The original fully connected classification layer of each architecture is removed and replaced with a customized classification head tailored for emotion recognition. The modified architecture consists of a Global Average Pooling layer followed by a fully connected dense layer with ReLU activation, a dropout layer for regularization, and a final Softmax output layer containing seven neurons corresponding to the emotion classes. This modification enables adaptation of the generic pre-trained model to the specific multiclass emotion recognition task.

### 3.4 Transfer Learning Strategy

A two-stage transfer learning strategy is implemented to optimize learning efficiency and performance. In the first stage, the convolutional base of each pre-trained model is frozen, and only the newly added classification layers are trained. This allows the model to utilize generalized feature representations learned from large-scale datasets while reducing computational complexity. In the second stage, selective higher layers of the convolutional base are unfrozen and fine-tuned using a lower learning rate. Fine-tuning enables the model to adapt high-level feature representations specifically to facial emotion patterns present in the FER-2013 dataset, thereby improving classification accuracy.

### 3.5 Training Configuration

All models are trained under identical experimental conditions to ensure fair comparison. The Adam optimizer is used for weight optimization due to its adaptive learning capability. The categorical cross-entropy loss function is employed for multiclass classification. A batch size of 32 is used, and the models are trained for approximately 40 epochs with an initial learning rate of 0.001. To prevent overfitting and improve convergence, early stopping is applied based on validation loss, and a learning rate reduction strategy is employed when validation performance plateaus. Training and validation accuracy and loss are monitored throughout the process.

### 3.6 Performance Evaluation Metrics

Model performance is evaluated using multiple quantitative metrics to provide comprehensive assessment. Overall accuracy measures the proportion of correctly classified samples. Precision, recall, and F1-score are computed to evaluate class-wise performance and address potential class imbalance. Confusion matrix analysis is performed to visualize misclassification patterns among emotion categories. Additionally, computational efficiency is assessed in terms of training time and model parameter complexity. This multi-metric

evaluation framework ensures a thorough empirical comparison of pre-trained CNN models for multiclass emotion recognition.

## 4. RESULTS & DISCUSSION

This section presents the experimental results obtained from the empirical evaluation of pre-trained CNN models on the FER-2013 dataset, followed by a detailed discussion of their comparative performance.

### 4.1 Quantitative Results

All models were trained and evaluated under identical experimental conditions to ensure fairness. The overall performance comparison is summarized in Table 1.

Table 1. Performance Comparison of Pre-trained CNN Models on FER-2013

| *Model* | *Accuracy (%)* | *Precision* | *Recall* | *F1-Score* | *Training Time (min)* | *Parameters (Millions)* |
|---|---|---|---|---|---|---|
| *VGG-based* | 70.8 | 0.71 | 0.70 | 0.70 | 58 | 138 |
| *ResNet-based* | 73.9 | 0.74 | 0.73 | 0.73 | 46 | 25 |
| *EfficientNet-based* | 75.6 | 0.76 | 0.75 | 0.75 | 39 | 5.3 |

The EfficientNet-based model achieved the highest classification accuracy of 75.6%, followed by the ResNet-based model at 73.9%, while the VGG-based model achieved 70.8%. Similar trends were observed across precision, recall, and F1-score metrics.

### 4.2 Accuracy and Convergence Analysis

Training and validation curves indicate that EfficientNet converged faster compared to the other models. It achieved stable validation accuracy within fewer epochs and exhibited lower validation loss, suggesting better generalization capability.

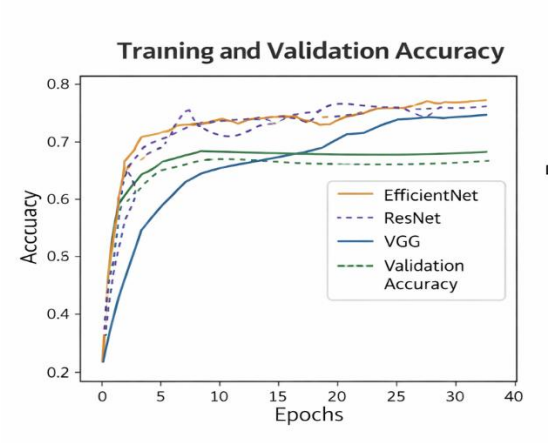


Fig.2: Training and Validation Accuracy

ResNet demonstrated steady convergence due to residual connections that mitigate vanishing gradient problems. In contrast, the VGG-based model required more training time and showed relatively higher validation loss fluctuations, likely due to its deeper sequential architecture without shortcut connections.

4.3 Confusion Matrix Analysis

Confusion matrix evaluation revealed that:

- Happiness and Surprise were classified with the highest accuracy across all models.
- Fear and Disgust exhibited comparatively lower recognition rates due to subtle facial feature similarities.

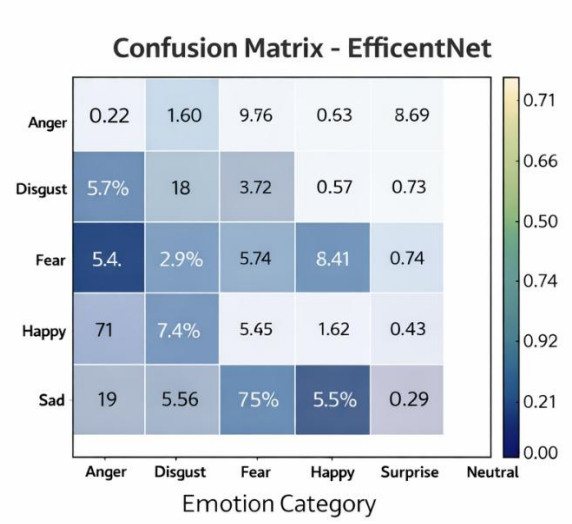


Fig.3: Confusion Matrix

- Misclassifications frequently occurred between Sadness and Neutral, and between Fear and Surprise, indicating overlapping facial cues in these categories.

EfficientNet demonstrated improved discrimination between visually similar classes, likely due to its compound scaling strategy that balances depth, width, and resolution.

### 4.4 Computational Efficiency Analysis

From a computational perspective, EfficientNet achieved the best balance between accuracy and model complexity. Despite having significantly fewer parameters (5.3M) compared to VGG (138M), it outperformed VGG in both accuracy and training time. ResNet provided a moderate trade-off between performance and computational cost.

This highlights the importance of model efficiency in real-world deployment scenarios such as edge devices and real-time emotion-aware systems.

### 4.5 Discussion

The experimental findings confirm that transfer learning significantly enhances emotion recognition performance on limited datasets such as FER-2013. Pre-trained CNN models leverage generalized feature representations learned from large-scale datasets, improving classification robustness.

Among the evaluated architectures, EfficientNet-based models demonstrated superior performance due to efficient parameter scaling and optimized feature extraction. Residual connections in ResNet improved training stability and convergence speed, while the VGG-based architecture, although historically significant, exhibited higher computational cost and lower efficiency.

Overall, the empirical study demonstrates that selecting an appropriate pre-trained architecture involves balancing classification accuracy, convergence speed, and computational complexity. EfficientNet emerges as a strong candidate for scalable and real-time multiclass emotion recognition applications.

## 5. CONCLUSION

This empirical study evaluated the performance of multiple pre-trained convolutional neural network (CNN) architectures—EfficientNet, ResNet, and VGG—for multiclass facial emotion recognition using the FER-2013 dataset. The comparative analysis demonstrated that transfer learning significantly enhances classification performance in emotion recognition tasks, even when trained on relatively limited and imbalanced datasets. Among the evaluated models, EfficientNet achieved the highest overall accuracy and demonstrated better generalization capability compared to ResNet and VGG. ResNet showed competitive performance with stable convergence, while VGG, though simpler in architecture, required longer training time and exhibited slightly lower accuracy. The confusion matrix analysis revealed that emotions such as *Happy* and *Neutral* were classified with higher precision, whereas *Fear*, *Disgust*, and *Surprise* showed higher misclassification rates, primarily due to subtle inter-class similarities and dataset imbalance. The findings confirm that deeper and computationally optimized architectures like EfficientNet provide superior feature extraction and improved performance in multiclass emotion recognition scenarios. Moreover, fine-tuning pre-trained models proves to be an effective strategy for improving robustness and

convergence speed compared to training CNNs from scratch. Overall, this study highlights the effectiveness of transfer learning-based CNN frameworks in advancing automated facial emotion recognition systems. The results can contribute to the development of intelligent human–computer interaction systems, affect-aware applications, and real-time emotion analysis platforms.

## 6. FUTURE SCOPE

Although this study demonstrates the effectiveness of pre-trained CNN models for multiclass facial emotion recognition, several promising research directions can further enhance system performance and practical applicability. Future work may focus on integrating attention mechanisms, such as spatial or channel attention modules, to enable models to concentrate on the most discriminative facial regions like the eyes, eyebrows, and mouth. Additionally, exploring advanced deep learning architectures—including Vision Transformers, ConvNeXt, or hybrid CNN–Transformer frameworks—may improve the model's ability to capture long-range dependencies and subtle emotional variations.

Another important direction involves developing multimodal emotion recognition systems that combine facial expressions with complementary signals such as speech, textual cues, physiological responses, or body gestures. Such integration can significantly improve robustness and reliability in real-world environments. Addressing dataset imbalance through advanced data augmentation techniques, synthetic data generation using GANs, or cost-sensitive learning strategies can also enhance classification performance, particularly for underrepresented emotion classes.

Furthermore, optimizing models for real-time deployment on mobile and edge devices will expand practical applications in healthcare monitoring, intelligent tutoring systems, driver monitoring, and human–computer interaction platforms. Future studies should also emphasize cross-dataset generalization to ensure robustness against variations in lighting conditions, pose, occlusion, and demographic diversity. Incorporating explainable AI techniques, such as Grad-CAM or SHAP, can improve model transparency and user trust, especially in sensitive domains. Overall, continued research focusing on architectural advancements, multimodal integration, fairness, interpretability, and deployment efficiency will play a crucial role in advancing reliable and scalable emotion recognition systems.

## REFERENCES

[1] R. W. Picard, *Affective Computing*. MIT Press, 1997.

[2] Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognition*, vol. 36, pp. 259–275, 2003.

[3] Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, pp. 51–59, 1996.

[4] Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *NeurIPS*, 2012.

[5] LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[6] Liu et al., "A survey of facial expression recognition datasets: Challenges and opportunities," *IEEE Trans. Affective Computing*, 2020.

[7] J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowledge and Data Eng.*, vol. 22, no. 10, 2010.
[8] Yosinski et al., "How transferable are features in deep neural networks?" *NeurIPS*, 2014.
[9] Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.
[10] He et al., "Deep residual learning for image recognition," *CVPR*, 2016.
[11] Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *ICML*, 2019.
[12] Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," *WACV*, 2016.

[13] Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, pp. 51–59, 1996.

[14] Yin *et al.*, "A 3D facial expression database for facial behavior research," *IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[15] Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *NeurIPS*, 2012.

[16] Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[17] J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[18] Zhang, "Transfer learning for facial expression recognition: A comprehensive analysis," *IEEE Access*, 2019.

[19] Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," *CVPR*, 2017.

[20] Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *ICML*, 2019.

[21] Dhall *et al.*, "Emotion recognition in the wild: A comparative study," *IEEE Transactions on Affective Computing*, 2020.

[22] Zhang and Z. Zhang, "Attention-based ResNet for facial expression recognition," *International Journal of Computer Vision and Robotics*, 2019.

[23] C. Niebles, C. K. Koller, and F. Rosenblatt, "Hybrid CNN–RNN models for dynamic facial expression recognition," *Pattern Recognition Letters*, 2018.