EMOTION RECOGNITION FROM STATIC FACIAL IMAGES USING TRANSFER LEARNING AND CNN ARCHITECTURES

*Abdul Ahmed Abdul[1,2,3,4], Samual Gali[2,4], Abdulhafiz Umar Dabo[2,4], Zarau Baidu[2,4], Muhammad Saleh Ibrahim[2,4], Lukman Aliyu Jibril[2,4]

[1]Infomation Technology Department, Shamrock Innovations.

[2]DeepLearning Fellowship Pytorch, Arewa Data Science Academy.

[3]Computer Science Department, Ahmadu Bello University Zaria

[4]Computer Science Department, Bayero University Kano.

*Corresponding author email: ahmadabdul592@gmail.com, Tel: +2348032642267

ABSTRACT

Facial Emotion Recognition (FER) is a crucial task in affective computing, with wide-ranging applications in healthcare, human-computer interaction, and smart systems. This study presents a deep learning-based FER system using transfer learning on a pre-trained ResNet-18 model, fine-tuned to recognize seven universal emotions: angry, disgust, fear, happy, neutral, sad, and surprise. The system is trained on a curated Kaggle dataset containing static facial images, employing robust preprocessing techniques such as face cropping, image cleaning, and data augmentation. A custom classification head with multiple dense layers and dropout is designed to enhance emotion-specific feature extraction. The model achieved a peak validation accuracy of 69.87% and a final test accuracy of 70%, outperforming several baseline methods and prior CNN-based FER systems. Precision and F1-score metrics demonstrate strong performance on expressive classes like "happy" and "surprise," while highlighting challenges in less distinguishable classes such as "fear" and "sad." The experimental results affirm the effectiveness of transfer learning and customized architecture in improving FER performance. Future work will focus on mitigating class imbalance, exploring advanced architectures like EfficientNet and Swin Transformer, and integrating ensemble strategies for enhanced generalization across facial orientations.

INTRODUCTION

Emotions are integral aspects of human nature, serving critical roles in social interaction and communication (Ekman, 2006). People express emotions through various channels, including facial expressions (Avila et al., 2021), speech (Soleymani et al., 2012), and body posture or gestures (Noroozi et al., 2019). Among these, facial expressions are the most commonly explored and well-established medium for emotion recognition, as initially categorized by Ekman and Friesen (1971). Ekman (2006) identified a set of universal facial expressions—happiness, sadness, anger, fear, surprise, disgust, and neutrality—that are widely accepted across cultures. In recent years, recognizing emotions through facial expression analysis has emerged as a compelling research area across diverse fields such as psychology, mental health, and human-computer interaction (Suchitra & Tripathi, 2016).

The automated detection of emotions from facial cues holds great potential in numerous domains, including smart environments (Yaddaden et al., 2016), healthcare systems (Fernandez-Caballero et al., 2016), and the diagnosis of emotional disorders in conditions such as autism spectrum disorder (Wingate, 2014) and schizophrenia (Thonse et al., 2018). Furthermore, facial emotion recognition (FER) is increasingly critical in enhancing human-computer interaction (Pantic et al., 2005), human-robot interaction (Gross et al., 2010), and social robotics aimed at welfare applications (O'Toole et al., 2005). Consequently, FER has garnered significant interest due to its broad applicability in real-world scenarios.

Recent advancements in deep learning, particularly convolutional neural networks (CNNs), have revolutionized FER by enabling automatic extraction of relevant features from facial images (Alom et al., 2019). Several studies have utilized CNNs to address FER challenges (Pranav et al., 2020). However, many existing approaches rely on shallow CNN architectures, despite evidence that deeper models often yield better performance in complex image analysis tasks (Khan et al., 2020). This limitation may stem from several challenges intrinsic to FER. Firstly, recognizing emotions requires moderately high-resolution images, resulting in high-dimensional input data. Secondly, the subtle variations in facial features corresponding to different emotions make classification particularly difficult. Additionally, training very deep CNNs is computationally intensive and often hampered by issues such as vanishing gradients, which degrade performance beyond a certain depth (Kolen & Kremer, 2010).

To mitigate these challenges, enhancements to CNN architectures and training methods have

been introduced. Pre-trained deep convolutional networks, such as VGG-16 (Simonyan & Zisserman, 2014), Inception-v3 (Szegedy et al., 2015), and DenseNet-161 (Huang et al., 2017), have been widely adopted. However, these models demand substantial training data and high computational resources. Another practical concern in FER is robustness across facial poses. Most studies focus solely on frontal face images, while real-world conditions often involve non-frontal or profile views. Although some datasets include profile views, these are frequently omitted during experiments for simplicity (Sahu & Dash, 2020). Therefore, a more adaptable FER system should be capable of recognizing emotions from both frontal and angled facial images. In this context, the present study introduces a facial emotion recognition system that leverages deep CNNs and transfer learning (TL) to enhance efficiency and performance. Transfer learning enables models to leverage previously acquired knowledge, thereby significantly reducing the need for extensive data and computational resources (Oquab et al., 2014). By utilizing pre-trained models, our approach addresses the challenges of training deep architectures from scratch and provides a viable solution for real-world FER applications.

RELATED WORKS

Deep learning has emerged as a relatively recent yet promising approach in facial emotion recognition (FER), with numerous CNN-based methods now reported in the literature. Zhao and Zhang (2015) proposed a hybrid model that combined a Deep Belief Network (DBN) for unsupervised feature extraction with a neural network (NN) for emotion classification. Similarly, Pranav et al. (2020) employed a basic CNN architecture consisting of two convolutional and pooling layers to analyze a self-curated dataset of emotional facial expressions. Expanding upon this, Mollahosseini et al. (2016) incorporated four inception modules along with convolutional layers to enhance the network's depth and performance. Pons and Masip (2018) adopted an ensemble strategy, training 72 CNN models with varying convolutional filter sizes and fully connected layer configurations. Wen et al. (2017) also explored ensemble learning, training 100 CNNs and selecting the best-performing subset for final predictions. Ashamshi et al. (2017) demonstrated the advantages of initializing CNN weights using a stacked convolutional autoencoder rather than random initialization, yielding improved learning efficiency and accuracy.

In another study, Ding et al. (2017) adapted a deep face recognition framework to FER and introduced the FaceNet2ExpNet architecture, which was later extended by Li et al. (2020) using transfer learning techniques. Hybrid deep learning architectures have also been explored, such as the combination of CNN and RNN in Jain et al. (2020), and the integration of pre-trained AlexNet with an SVM classifier in the work of Shaees et al. (2020). Bendjillali et al. (2019) utilized CNNs trained on features extracted through Discrete Wavelet Transform (DWT), while Liliana (2019) implemented a deep CNN model with 18 convolutional layers and four subsampling layers.

Innovative approaches continue to emerge, including Shi et al. (2020), who integrated clustering with CNNs for FER, and Ngoc et al. (2020), who proposed a graph-based CNN utilizing facial landmark features. Jin et al. (2019) introduced semi-supervised learning by incorporating unlabeled data alongside labeled samples in their CNN-based framework. Furthermore, Porcu et al. (2020) investigated the effectiveness of various data augmentation strategies—particularly synthetic image generation—which enhanced CNN training and performance.

Despite these advancements, most existing deep learning-based FER methods are limited by their reliance on frontal face images. In many cases, profile view images present in datasets are deliberately excluded to simplify experimentation (Porcu et al., 2020). This highlights the need for more robust FER systems capable of handling both frontal and profile facial views for practical, real-world deployment

METHODOLOGY

Our methodology follows a structured pipeline for training and testing an Emotion Recognition model, as illustrated in Figure 1. The process involves two main stages: (1) training a pre-trained ResNet-18 model on a facial expression dataset, and (2) testing the fine-tuned model on new images to predict emotion probabilities.

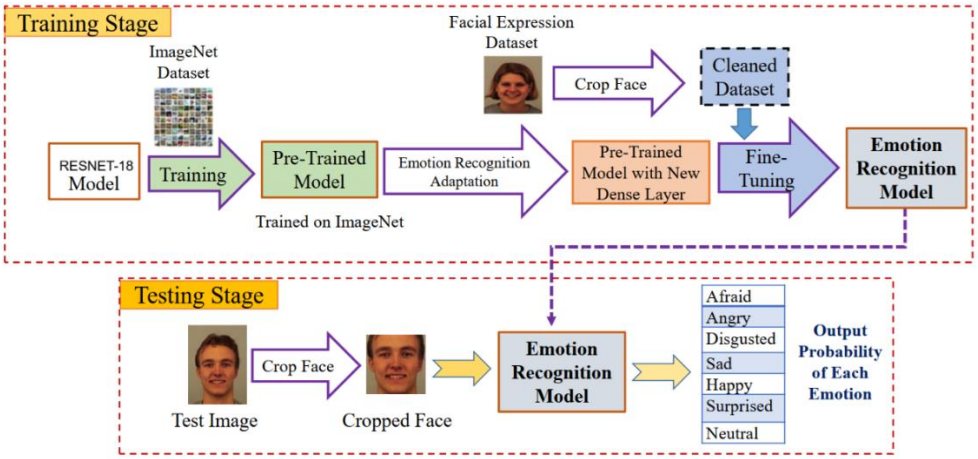

Figure 1: Pipeline for Training and Testing the Emotion Recognition Model

- Training Stage: Start with a pre-trained ResNet-18 model (trained on ImageNet), process a facial expression dataset by cropping faces, clean the dataset, adapt the model with a new dense layer, and fine-tune to produce an Emotion Recognition Model.
- Testing Stage: Crop the face from a test image, input it into the fine-tuned model, and output probabilities for each emotion (afraid, angry, disgusted, sad, happy, surprised, neutral).

Dataset and Preprocessing

We used a Kaggle dataset with static facial images labeled across seven emotions, split into 80% training and 20% validation sets using a custom train_val function to ensure balanced class distribution (random seed: 42). As shown in the training stage of Figure 1, the dataset undergoes face cropping to isolate facial regions, followed by cleaning to remove misaligned or low-quality images. Transformations include:

- Training: Convert to 3-channel grayscale, resize to 256x256, random crop to 224x224, apply horizontal flips, ±10-degree rotations, brightness/contrast jitter, and ImageNet normalization (mean: [0.485, 0.456, 0.406], std: [0.229, 0.224, 0.225]).
- Validation: Convert to 3-channel grayscale, resize to 256x256, center crop to 224x224, and normalize.

Model Architecture

We fine-tuned ResNet-18 (pre-trained on ImageNet) by adapting it for emotion recognition, as depicted in Figure 1. The adaptation involves replacing the original fully connected layer with a new dense layer, forming a custom classification head: three linear layers (512→512, 512→256, 256→7) with ReLU activations and 50% dropout to reduce overfitting. This aligns with the "Emotion Recognition Adaptation" and "Pre-Trained Model with New Dense Layer" steps. our complex head enhances emotion-specific feature extraction. We kept all layers trainable, as freezing reduced accuracy in preliminary experiments. DataLoaders are created with a batch size of 32, utilising 2 workers and shuffling for training to enhance stochasticity during optimization.

Training Procedure

Training is conducted using PyTorch with the following configuration:

- **Optimizer**: Stochastic Gradient Descent (SGD) with a learning rate of 0.0229 (determined via FastAI's learning rate finder), momentum of 0.9, and weight decay of 1e-4.
- **Loss Function**: Cross-Entropy Loss.
- **Learning Rate Scheduler**: OneCycleLR, adjusting the learning rate per batch over 50 epochs.

The training loop includes a train_step function for forward pass, loss calculation, and optimization, and a test_step function for validation, computing average loss and accuracy per epoch. Training ran for up to 50 epochs or until early stopping was triggered.

EXPERIMENTAL SETUP

Hardware and Software

Experiments were conducted on a system with GPU support (CUDA-enabled where available, otherwise CPU). The software stack includes PyTorch, torchvision, FastAI, and scikit-learn for evaluation metrics. The dataset is hosted on Kaggle, accessed via the Kaggle datasets database.

Data Splitting and Transformations

The training dataset is split into 80% training and 20% validation sets using the train_val function, ensuring balanced class distribution. The test dataset is processed separately with validation transformations. Transformations are tailored to ResNet-18's input requirements, converting grayscale images to 3-channel format and normalising with ImageNet statistics. The face cropping and cleaning steps enhance dataset quality.

Testing Procedure

As shown in the testing stage of Figure 1, a test image undergoes face cropping to isolate the facial region, matching the preprocessing applied during training. The cropped face is input into the fine-tuned Emotion Recognition Model, which outputs probabilities for each of the seven emotions (afraid, angry, disgusted, sad, happy, surprised, neutral). This process ensures consistency between training and inference, enhancing model reliability.

RESULT DISCUSSION

The model achieved a peak validation accuracy of 69.87% at epoch 48, with early stopping not triggered as improvements continued within the patience window. The final test accuracy was 70%, evaluated on a separate test set. The classification report provides detailed metrics:

Emotion	Precision	Recall	F1-Score	Support
Angry	**0.63**	**0.63**	**0.63**	**958**
Disgust	**0.81**	**0.71**	**0.76**	**111**
Fear	**0.56**	**0.52**	**0.54**	**1024**
Happy	**0.88**	**0.88**	**0.88**	**1774**
Neutral	**0.65**	**0.66**	**0.66**	**1233**
Sad	**0.56**	**0.58**	**0.57**	**1247**
Surprise	**0.80**	**0.82**	**0.81**	**831**
Macro Avg	**0.70**	**0.69**	**0.69**	**7178**
Weighted Avg	**0.70**	**0.70**	**0.70**	**7178**

Table 1: FER Performance Result

The model performs best on “happy” (F1-score: 0.88) and “surprise” (F1-score: 0.81), likely due to distinct facial features. “Fear” and “sad” show lower performance (F1-scores: 0.54 and 0.57), possibly due to class imbalance or feature similarity. Compared to the first model experiment (67% accuracy), the increased complexity of the classification head and higher dropout rates improved accuracy by 3%. Also, Ng et al., (2015) in their paper stated that after conducting a similar experiment with CNN got 56% test accuracy, which is significantly lower than our test accuracy. As such there is a 14% increase in the model while using Transfer Learning. The results indicate that the custom classification head and higher dropout rates effectively reduced overfitting compared to similar experiment done on the same dataset, as evidenced by the sustained improvement in validation accuracy. The decision to keep base layers trainable was critical, as freezing them in preliminary tests led to poor performance (33–42%). The OneCycleLR scheduler and FastAI’s learning rate finder optimized training dynamics, contributing to the 70% test accuracy.

Furthermore, to further enhance performance, we propose:

- Addressing Class Imbalance: Implement techniques like oversampling (e.g., SMOTE) or class-weighted loss functions to improve performance on underrepresented classes like “disgust.”
- Advanced Data Augmentation: Incorporate Test-Time Augmentation (TTA) to enhance robustness during inference.
- Model Enhancements: Experiment with deeper architectures (e.g., ResNet-256, SwimTransformers, EfficientNet) or ensemble methods to capture more complex patterns

REFERENCE

Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.S.; Asari, V.K. (2019). A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics*, *8*, 292

Alshamsi, H.; Kepuska, V.; Meng, H. (2017). Stacked deep convolutional auto-encoders for emotion recognition from facial expressions. *Proc. Int. Jt. Conf. Neural Netw.* 1586–1593

Avila, A.R.; Akhtar, Z.; Santos, J.F.; O’Shaughnessy, D.; Falk, T.H (2021). Feature Pooling of Modulation Spectrum Features for Improved Speech Emotion Recognition in the Wild. IEEE Trans. Affect. Comput. 12, 177–188.

Bendjillali, R.I.; Beladgham, M.; Merit, K.; Taleb-Ahmed, A. (2019). Improved Facial Expression Recognition Based on DWT Feature for Deep CNN. *Electronics 8*, 324.

Ding, H.; Zhou, S.K.; Chellappa, R. (2017). FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington.

Ekman, P. (2006) *Cross-Cultural Studies of Facial Expression. Darwin and Facial Expression*; Malor Books: Los Altos, CA, USA, pp. 169–220.

Ekman, P.; Friesen, W.V. (1971) Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol. 17*, 124–129

Fernández-Caballero, A.; Martínez-Rodrigo, A.; Pastor, J.M.; Castillo, J.C.; Lozano-Monasor, E.; López, M.T.; Zangróniz, R.; Latorre, J.M.; Fernández-Sotos, (2016) A. Smart

environment architecture for emotion detection and regulation. *J. Biomed. Inf*, *64*, 55–73

Gross, R.; Matthews, I.; Cohn, J.; Kanade, T.; Baker, S. (2010). Multi-PIE. *Image Vis. Comput. 28*, 807–813.

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), Honolulu, HI, USA. pp. 2261–2269

Jin, X.; Sun, W.; Jin, Z. (2019). A discriminative deep association learning for facial expression recognition. Int. J. Mach. Learn. Cybern.11, 779–793

Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A.S. (2020) A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev. 53*, 5455–5516

Kolen, J.F.; Kremer, S.C. (2010). Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies. In *A Field Guide to Dynamical Recurrent Networks*; Wiley-IEEE Press: Hoboken, NJ, USA, pp. 237–243.

Li, J.; Huang, S.; Zhang, X.; Fu, X.; Chang, C.-C.; Tang, Z.; Luo, Z. (2020). Facial Expression Recognition by Transfer Learning for Small Datasets. In *Advances in Intelligent Systems and Computing*; Springer: Berlin/Heidelberg, Germany, Volume 895, pp. 756–770.

Liliana, D.Y. (2019). Emotion recognition from facial expression using deep convolutional neural network. *J. Phys. Conf. Ser..*, *1193*, 012004

Mollahosseini, A.; Chan, D.; Mahoor, M.H. (2016) Going deeper in facial expression recognition using deep neural networks. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, pp. 1–10.

Ng, Hong-Wei., Nguyen, V. D., Vonikakis, V., & Winkler, S. (2015). Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning. In *Proceedings of the 17th International Conference on Multimodal Interaction (ICMI '15)*. ACM.

Ngoc, Q.T.; Lee, S.; Song, B.C. (2020). Facial Landmark-Based Emotion Recognition via Directed Graph Neural Network. *Electronics 9*, 764

Noroozi, F.; Marjanovic, M.; Njegus, A.; Escalera, S.; Anbarjafari, G. (2019) Audio-Visual Emotion Recognition in Video Clips. *IEEE Trans. Affect. Comput. 10*, 60–75

O'Toole, A.J.; Harms, J.; Snow, S.L.; Hurst, D.R.; Pappas, M.R.; Ayyad, J.H.; Abdi, H. (2005). A video database of moving faces and people. *IEEE Trans. Pattern Anal. Mach. Intell. 27*, 812–816

Oquab, M.; Bottou, L.; Laptev, I.; Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 pp. 1717–1724.

Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L. (2005) Web-Based Database for Facial Expression Analysis. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, pp. 317–321

Pons, G.; Masip, D. (2018). Supervised Committee of Convolutional Neural Networks in Automated Facial Expression Analysis. *IEEE Trans. Affect. Comput. 9*, 343–350

Porcu, S.; Floris, A.; Atzori, L. (2020). Evaluation of Data Augmentation Techniques for Facial Expression Recognition Systems. Electronics 9, 1892

Pranav, E.; Kamal, S.; Chandran, C.S.; Supriya, (2020) M. Facial emotion recognition using deep convolutional neural network. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 ; pp. 317–320

Sahu, M.; Dash, R. (2021). A Survey on Deep Learning: Convolution Neural Network (CNN). In *Smart Innovation, Systems and Technologies*; Springer: Singapore, 2021; Volume 153, pp. 317–325

Shaees, S.; Naeem, H.; Arslan, M.; Naeem, M.R.; Ali, S.H.; Aldabbas, H. (2020). Facial Emotion Recognition Using Transfer Learning. In Proceedings of the 2020 International Conference on Computing and Information Technology (ICCIT-1441), Tabuk, Saudi Arabia

Shi, M.; Xu, L.; Chen, X. (2020). A Novel Facial Expression Intelligent Recognition Method Using Improved Convolutional Neural Network. IEEE Access 8, 57606–57614

Simonyan, K.; Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv* , arXiv:1409.1556

Soleymani, M.; Pantic, M.; Pun, T. (2012) Multimodal Emotion Recognition in Response to Videos. *IEEE Trans. Affect. Comput. 3*, 211–223.

Suchitra, P.S.; Tripathi, S. (2016). Real-time emotion recognition from facial images using Raspberry Pi II. In Proceedings of the 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 11–12 February; pp. 666–670

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, (2015). A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 1–9.

Thonse, U.; Behere, R.V.; Praharaj, S.K.; Sharma, P.S.V.N. (2018). Facial emotion recognition,

socio-occupational functioning and expressed emotions in schizophrenia versus bipolar disorder. *Psychiatry Res. 264*, 354–360.

Wen, G.; Hou, Z.; Li, H.; Li, D.; Jiang, L.; Xun, E. (2017). Ensemble of Deep Neural Networks with Probability-Based Fusion for Facial Expression Recognition. *Cogn. Comput. 9*, 597–610.

Wingate, M. (2014) Prevalence of Autism Spectrum Disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, United States, 2010. *MMWR Surveill. Summ. 63*, 1–21

Zhao, X.; Shi, X.; Zhang, S. (2015). Facial Expression Recognition via Deep Learning. *IETE Tech. Rev. 32*, 347–355