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

*Abdul Ahmed Abdul[1,2,3,4], Samual Gali[2,5], 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.
[5]Computer Science Department, University of Lagos.
*Corresponding author email: ahmadabdul592@gmail.com, Tel: +2348032642267

ABSTRACT

Facial Emotion Recognition (FER) plays a key role in affective computing, with applications in healthcare, smart systems, and human-computer interaction. This study proposes a FER model based on transfer learning using a pre-trained ResNet-18, fine-tuned to classify seven universal emotions: angry, disgust, fear, happy, neutral, sad, and surprise. Trained on a Kaggle facial image dataset, the approach incorporates face cropping, cleaning, and augmentation to enhance data quality. A customized classification head improves feature extraction, leading to a peak validation accuracy of 69.87% and test accuracy of 70%. The model shows strong performance on expressive emotions like “happy” and “surprise,” but struggles with subtle ones such as “fear” and “sad.” Results highlight the benefits of transfer learning and architectural tuning. Future work will address class imbalance, evaluate advanced models like EfficientNet and Swin Transformer, and explore ensemble methods for better generalization.

INTRODUCTION

Emotions are fundamental to human interaction and communication, expressed through facial expressions, speech, and body language (Ekman, 2006; Avila et al., 2021; Soleymani et al., 2012; Noroozi et al., 2019). Among these, facial expressions are the most widely studied for emotion recognition. Ekman and Friesen (1971) identified seven universal emotions—happiness, sadness, anger, fear, surprise, disgust, and neutrality—recognized across cultures (Ekman, 2006). Consequently, facial emotion recognition (FER) has gained prominence in domains like psychology, mental health, and human-computer interaction (Suchitra & Tripathi, 2016). Automated FER has broad applications, including smart environments (Yaddaden et al., 2016), healthcare (Fernandez-Caballero et al., 2016), and diagnosis of conditions such as autism and schizophrenia (Wingate, 2014; Thonse et al., 2018). It also plays a crucial role in enhancing human-computer and human-robot interaction (Pantic et al., 2005; Gross et al., 2010; O’Toole et al., 2005).

Recent advances in deep learning, especially convolutional neural networks (CNNs), have transformed FER by enabling automatic feature extraction (Alom et al., 2019; Pranav et al., 2020). However, many approaches still rely on shallow models, limiting accuracy in recognizing subtle facial variations (Khan et al., 2020). Challenges such as high-dimensional inputs, computational demands, and issues like vanishing gradients further complicate deep model training (Kolen & Kremer, 2010).

To address these issues, researchers have adopted pre-trained CNNs like VGG-16, Inception-v3, and DenseNet-161 (Simonyan & Zisserman, 2014; Szegedy et al., 2015; Huang et al., 2017). While effective, these models require significant data and computing power. Additionally, FER systems often lack robustness to non-frontal facial poses, despite their prevalence in real-world settings (Sahu & Dash, 2020).

This study proposes a FER approach that leverages deep CNNs and transfer learning (TL) to improve efficiency and generalization. TL enables the reuse of learned features, reducing data and computation requirements (Oquab et al., 2014) and enhancing the model’s applicability across diverse facial angles and real-world conditions.

RELATED WORKS

Recent deep learning approaches to FER predominantly leverage CNNs with various architectural and training enhancements. For instance, Zhao and Zhang (2015) combined a Deep Belief Network for unsupervised feature extraction with a neural network for classification, while Pranav et al. (2020) used a simple two-layer CNN on a self-curated dataset. Mollahosseini et al. (2016) enhanced network depth and performance by adding four Inception modules. Ensemble strategies have been explored by Pons and Masip (2018), who trained 72 CNNs with different filter sizes and fully connected layers, and Wen et al. (2017), who trained 100 CNNs and

selected the top-performing models. Ashamshi et al. (2017) improved learning efficiency by initializing CNN weights with a stacked convolutional autoencoder instead of random weights.

Transfer learning and hybrid architectures also show promise: Ding et al. (2017) adapted a face recognition framework into FaceNet2ExpNet for FER, later extended by Li et al. (2020) using transfer learning. Jain et al. (2020) combined CNNs with RNNs to capture sequential features, and Shaees et al. (2020) integrated a pre-trained AlexNet with an SVM for classification. Bendjillali et al. (2019) trained CNNs on Discrete Wavelet Transform features, while Liliana (2019) designed a deep CNN with 18 convolutional and four subsampling layers.

Emerging innovations include Shi et al. (2020), who fused clustering techniques with CNNs, and Ngoc et al. (2020), who proposed a graph-based CNN using facial landmarks. Jin et al. (2019) introduced semi-supervised learning by incorporating unlabeled data, and Porcu et al. (2020) demonstrated that synthetic image augmentation can substantially boost CNN performance. Despite these advances, most methods focus on frontal faces, often excluding profile views to simplify experiments (Porcu et al., 2020). This gap underscores the need for FER systems robust to both frontal and angled facial images.

METHODOLOGY

This study employs a structured pipeline for training and testing a facial emotion recognition (FER) model, as shown in Figure 1. The process involves two main stages: training a fine-tuned ResNet-18 model and evaluating it on unseen images.

1. Training Stage:

We start with a ResNet-18 model pre-trained on ImageNet. A facial expression dataset is processed by cropping facial regions and removing low-quality images. The model's final layer is replaced with a custom classification head consisting of three fully connected layers (512→512→256→7) with ReLU activations and dropout (50%) to mitigate overfitting. All layers remain trainable to retain performance gains observed in preliminary tests.

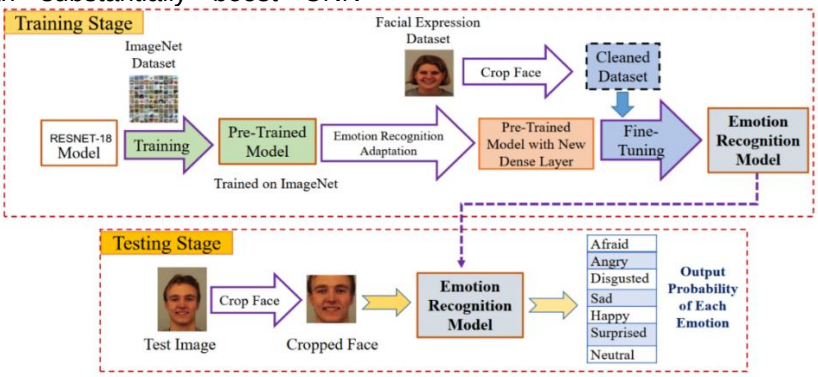

Figure 1: Pipeline for Training and Testing the Emotion Recognition Model

2. Testing Stage:

Each test image undergoes the same preprocessing—face cropping and resizing—before being passed to the trained model, which predicts the probability distribution across seven emotions: afraid, angry, disgusted, sad, happy, surprised, and neutral.

Dataset and Preprocessing

We use a static image dataset from Kaggle labelled with seven emotions as shown in figure 2. It is split 80/20 into training and validation sets using a custom function that ensures balanced classes. Key preprocessing steps include:

- **Training:** Convert images to 3-channel grayscale, resize to 256×256, apply random crops, flips, ±10° rotations, and brightness/contrast adjustments, followed by ImageNet normalization.
- **Validation:** Convert and resize images, then apply center cropping and normalization.

Face cropping and dataset cleaning are applied to both sets to improve data quality.

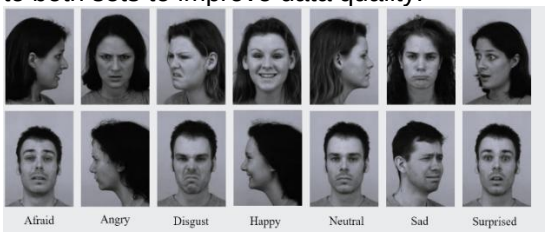

Figure 2: Sample Images from FER 2013

Model Architecture

The ResNet-18 model is fine-tuned by replacing its classifier with a three-layer dense head tailored for emotion classification. The model is

trained with a batch size of 32 and uses PyTorch DataLoaders with shuffling and two workers to optimise performance.

Training Procedure

Training is performed using PyTorch with the following setup:

- **Optimizer:** SGD (learning rate = 0.0229, momentum = 0.9, weight decay = 1e-4).
- **Loss Function:** Cross-Entropy Loss.
- **Scheduler:** OneCycleLR over 50 epochs with early stopping enabled.

The training loop includes functions for batch-wise forward pass, loss computation, and validation.

EXPERIMENTAL SETUP

Experiments run on a GPU-enabled system using PyTorch, FastAI, torchvision, and scikit-learn. Data transformations match ResNet-18's input requirements, including grayscale conversion, resizing, and ImageNet normalisation. During testing, the same preprocessing is applied, ensuring consistency with training and improving model reliability in emotion classification.

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%, and it was evaluated on a separate test set. 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.

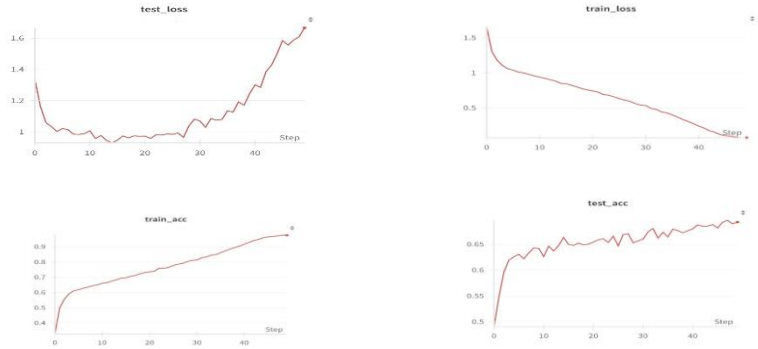

Figure 3: Loss and Accuracy Result

Additionally, Figure 3a-d illustrates the model's performance trends during training and testing. The 'train_acc' and 'test_acc' plots show the accuracy progression for the 'Resnet-18' model, with 'train_acc' reaching approximately 0.9 and 'test_acc' stabilising around 0.7, reflecting the reported 70% test accuracy. The 'train_loss' plot indicates a consistent decrease to below 0.5, demonstrating effective learning, while the 'test_loss' plot shows an initial decline followed by an increase. These visualisations support the effectiveness of the custom classification head and the training strategy, aligning with the observed improvements over baseline models.

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