JOURNAL
COMPUTER SCIENCES INSTITUTE

JCSI 37 (2025) 443-450
Received: 30 June 2025
Accepted: 19 August 2025

Evaluating the effectiveness of selected tools in recognizing emotions from facial photos

Ocena skuteczności wybranych narzędzi w rozpoznawaniu emocji na podstawie zdjęć twarzy

Klaudiusz Wierzbowski*

Department of Computer Science, Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland

Abstract
Emotion recognition from facial images has become a key area in computer vision and affective computing. Deep learning models such as convolutional neural networks and vision transformers have shown high potential in this domain. In this study, the performance of two representative architectures, ResNet-50, a convolutional neural networks based model, and ViT-B/16, a transformer-based model, is evaluated on the widely used Facial Expression Recognition 2013 dataset. Both models are trained using data augmentation and regularization techniques to enhance generalization. Their effectiveness is assessed using metrics including accuracy, precision, recall, and F1-score, alongside a detailed examination of confusion matrices. The observed differences in classification performance across emotion categories highlight the influence of architectural design on model behavior. The obtained results serve as a reference point for selecting appropriate deep learning architectures.

Keywords: convolutional neural networks; vision transformers; emotion recognition

Streszczenie
Rozpoznawanie emocji na podstawie zdjęć twarzy stanowi istotny obszar badań w dziedzinie wizji komputerowej oraz obliczeniowej analizy emocji. Modele głębokiego uczenia, takie jak konwolucyjne sieci neuronowe oraz transformatory wizyjne, wykazują duży potencjał w tym zakresie. W niniejszym badaniu oceniono skuteczność dwóch reprezentatywnych architektur, ResNet-50, opartej na konwolucyjnych sieciach neuronowych, oraz ViT-B/16, opartej na mechanizmie transformatora na szeroko stosowanym zbiorze danych Facial Expression Recognition 2013. Oba modele trenowano z zastosowaniem technik augmentacji danych i regularyzacji w celu poprawy generalizacji. Ocena skuteczności została przeprowadzona z wykorzystaniem metryk takich jak dokładność, precyzja, czułość oraz miara F1, a także poprzez analizę macierzy pomyłek. Zaobserwowane różnice w jakości klasyfikacji poszczególnych emocji ukazują wpływ architektury modelu na jego zachowanie. Uzyskane rezultaty stanowią źródło odniesienia przy wyborze odpowiednich architektur głębokiego uczenia.

Słowa kluczowe: konwolucyjne sieci neuronowe; transformatory wizyjne; rozpoznawanie emocji

*Corresponding author
Email address: s95605@pollub.edu.pl (K. Wierzbowski)

1. Introduction

Emotion recognition from facial images has become a significant area of interest in the intersection of computer vision, psychology, and artificial intelligence. Understanding human emotions through automated systems enables a wide range of applications, including human-computer interaction, mental health assessment, and surveillance [1-2].

Deep learning techniques, particularly convolutional neural networks (CNNs), have demonstrated high effectiveness in visual emotion classification [3]. These architectures are specifically designed to automatically extract relevant visual features from image data with minimal preprocessing. The core principle involves the use of convolutional layers, which apply learnable filters across the input image to detect distinctive patterns such as edges, textures, and progressively more complex structures [4]. CNNs are particularly well-suited for tasks such as image classification, object detection, and semantic segmentation. Typical CNN architectures consist of a series of convolutional and pooling layers that hierarchically capture spatial dependencies, followed by fully connected layers for final classification. The ability of CNNs to learn spatial hierarchies and generalize from raw pixel data has made them a dominant approach in visual recognition problems across a wide range of domains [5].

More recently, Vision Transformers (ViTs) – transformer-based architectures originally developed for natural language processing – have emerged as a compelling alternative, offering an enhanced ability to capture long-range dependencies in image data [6-7]. Unlike CNNs, which rely on local receptive fields and hierarchical feature extraction [8], ViTs treat an image as a sequence of fixed-size patches and apply self-attention mechanisms to model global dependencies between them [9]. ViTs have demonstrated competitive performance in image classification, especially when trained on large datasets [10]. Their ability to capture long-range relationships and model contextual information across the entire image

makes them well-suited for complex visual tasks. However, they tend to require more data and computational resources compared to CNNs during training [11].

These two paradigms represent distinct architectural approaches, each with its own strengths and limitations in terms of classification accuracy, model complexity, and generalization capacity.

The Facial Expression Recognition 2013 (FER2013) dataset, introduced as part of the ICML 2013 Challenges in Representation Learning, remains one of the most widely used benchmarks for facial emotion recognition [12]. Despite its relatively modest image resolution, it provides a valuable testing ground for evaluating the robustness and generalization capabilities of emotion recognition models.

The aim is to identify performance differences between two deep learning architectures – ResNet-50, a well-established convolutional neural network based on residual learning [13], and ViT-B/16, a transformer-based model adapted for image classification – using key evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis. Both models were independently instantiated and fine-tuned to ensure a consistent evaluation. Particular attention was given to maintaining equivalent preprocessing steps and data augmentation strategies across architectures. As a result, any observed differences in performance can be attributed primarily to the model design itself rather than external factors. The impact of architectural design on emotion recognition effectiveness is assessed, offering insights into the selection of suitable models for real-world emotion classification applications.

2. Related works

Harnessing the expressive power of deep neural networks, contemporary emotion-classification systems are beginning to decode the subtle patterns of human feeling embedded in speech, text, and facial cues with unprecedented accuracy [14-15].

Facial expression recognition has been extensively studied with deep learning models achieving notable success, particularly on benchmark datasets such as FER2013 [16]. One of the most commonly applied architectures is ResNet-50 due to its robust feature extraction and transfer learning capabilities.

Altaha et al. [17] proposed a ResNet-50-based pipeline incorporating ArcFace features and a Tiny-Siamese network for classification. Designed to reduce memory load and training time, the approach achieved 60.43% accuracy on FER2013, illustrating a trade-off between computational efficiency and recognition performance.

Sheng and Lau [18] compared multiple ResNet variants, including ResNet18, ResNet34, and ResNet-50. The ResNet-50 model achieved the highest accuracy in their study, attaining 65.40% after fine-tuning.

Soni et al. [19] evaluated two deep CNN architectures, VGG and ResNet-50 on the FER2013 dataset, obtaining accuracies of 50.12% and 52.40%, respectively. After adding two dense layers to each architecture, the results improved to 55.90% for VGG and 57.20% for ResNet-50. By combining both models, they achieved an accuracy of 66.15%, demonstrating that ensemble methods can improve recognition in complex environments.

Li and Li [20] proposed an architecture integrating spatial and frequency domain transformations, using ResNet-50 pretrained on VGGFace2 for appearance features and combining it with geometric features from dense SIFT. Fine-tuned on FER2013 and RAF Basic, the ensemble model achieved 66.97% accuracy on the RAF Compound set.

More recently, ViTs have emerged as a strong alternative to CNNs in facial expression recognition tasks. Bobojanov et al. [21] performed a comparative analysis of multiple ViT architectures, applying dataset cleaning and augmentation to reduce class imbalance. Mobile ViT emerged as the most effective, achieving 62.73% accuracy on FER2013.

Soni et al. [22] applied a fine-tuned ViT model to the FER2013 dataset with extensive preprocessing and augmentation. Their approach achieved an accuracy of 70.00%, underlining the transformer's ability to generalize across emotional classes.

Song [23] introduced novel ViT variants (ViTTL and ViTEH) that process self-attention outputs through global average pooling, improving the detection of local patterns. The best variant reached 70.37% accuracy.

Bie et al. [24] presented Swin-FER, a Swin Transformer that fuses middle and deep-layer features while controlling parameter growth through mean, split and group convolution modules. The model reached 71.11% accuracy on FER2013.

While ResNet-50 remains a strong and widely adopted baseline for facial expression recognition, ViT-based architectures that incorporate hybrid mechanisms or architectural refinements have demonstrated consistently competitive and, in some cases, superior performance. Reported accuracies for ResNet-50-based models on the FER2013 dataset range from 60.43% to 66.97%, reflecting the impact of design choices and feature integration strategies. In comparison, ViT-based approaches achieve accuracies from 62.73% to 71.11% (Table 1).

Table 1: Comparison of the classification accuracies achieved by ResNet-50- and ViT-based models across datasets reported in selected studies

Study	Model	Dataset	Accuracy
[17]	ResNet-50	FER2013	60.43%
[18]	ResNet-50	FER2013	65.40%
[19]	ResNet-50 and VGG	FER2013	66.15%
[20]	ResNet-50	FER2013 and RAF-DB	66.97%
[21]	Mobile ViT	FER2013	62.73%
[22]	ViT	FER2013	70.00%
[23]	ViT	FER2013	70.37%
[24]	Swin Transformer	FER2013	71.11%

3. Material and methods

3.1. ResNet-50

ResNet-50 (Residual Network, 50 layers) is a convolutional neural network architecture introduced by He et al.

in 2016 [25]. It addresses the problem of vanishing gradients in deep networks by introducing residual connections, which allow the network to learn identity mappings and thus preserve gradient flow through many layers.

The architecture consists of 49 convolutional layers and one fully connected layer at the end. The network is structured into residual blocks, each containing convolutional layers followed by batch normalization and ReLU activation. A key feature of these blocks is the shortcut connection, which bypasses one or more layers, allowing the model to train deeper networks more effectively. ResNet-50 is widely used in image classification tasks due to its balance between depth, accuracy, and computational efficiency [26].

3.2. ViT-B/16

ViT-B/16 (Vision Transformer Base with a 16×16 input patch size) is a Vision Transformer model introduced by Dosovitskiy et al. in 2021 [6], which relies on a transformer encoder. In ViT-B/16, the input images are first resized to 224×224 pixels and partitioned into 196 non-overlapping 16×16 patches, which are then linearly embedded and combined with positional encodings to retain spatial information. A learnable classification token ([CLS] token) is prepended to serve as a representation of an image.

The resulting 197-token sequence passes through a stack of 12 Transformer encoder layers, each comprising 12-head self-attention and a feedforward network that expands the 768-dimensional representation to 3072 dimensions and then reduces it back to 768, using Gaussian Error Linear Unit (GELU) activations. Ultimately, [CLS] token is fed into a single linear layer that produces the class logits for image classification [9-10].

3.3. Facial Expression Recognition 2013 dataset

Facial Expression Recognition 2013 dataset is one of the most widely used benchmarks for training and evaluating models in the field of facial emotion recognition. It was introduced during the ICML 2013 Challenges in Representation Learning and contains a total of 35,887 grayscale images, each with a resolution of 48×48 pixels [16].

This dataset is divided into three subsets: 28,709 images for training, 3,589 for validation (public test), and 3,589 for testing (private test). Each image is labeled with one of seven emotion categories: angry, disgust, fear, happy, sad, surprise, and neutral (Figure 1).

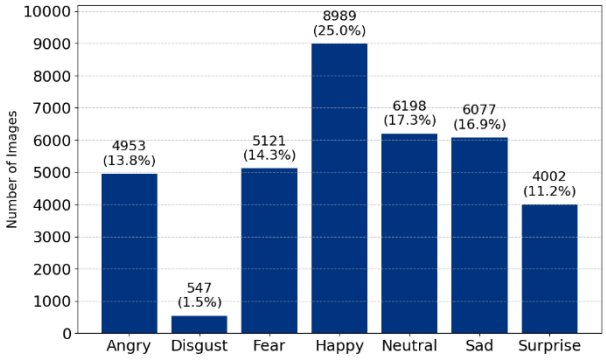

Figure 1: FER2013 class distribution.

Despite its relatively low resolution, FER2013 has proven effective for benchmarking deep learning models due to its size and diversity.

However, the dataset presents several challenges. The class distribution is imbalanced, with certain emotions like disgust being significantly underrepresented. Moreover, facial expressions in the images vary in pose, lighting, and occlusion (Figure 2), which introduces additional complexity and requires robust generalization capabilities from the models [27].

Figure 2: Example images from the FER2013 dataset [15].

3.4. Image augmentation

To address data imbalance and enhance generalization, data augmentation techniques and class-balanced weighting were applied [28-29]. The augmentation pipeline included random resized cropping, horizontal and vertical flips, affine transformations (rotation, translation, shear), and random erasing. These operations diversify pose, lighting, and partial-occlusion patterns while preserving the underlying facial content, providing the model with a richer and more balanced training distribution.

3.5. Evaluation Metrics

In order to compare the performance of the evaluated emotion recognition models, a set of well-established evaluation metrics is utilized. These include accuracy, precision, recall, F1-score, and the confusion matrix, each providing complementary insights into model behavior and effectiveness [30].

In statistical analysis of classification performance, results are summarized in a confusion matrix that records how many test samples the model assigns to each outcome:

- True Positives (TP) – positive instances classified correctly,
- True Negatives (TN) – negative instances classified correctly,

- False Positives (FP) – negative instances misclassified as positive,
- False Negatives (FN) – positive instances misclassified as negative.

An examination of the confusion matrix provides insight into the model's generalization ability and highlights misclassification patterns [31]. The matrix components form the basis of evaluation metrics.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where *accuracy* expresses the proportion of all images that were classified correctly.

$$precision = \frac{TP}{TP + FP} \quad (2)$$

where *precision* measures how many of the model's positive predictions are actually correct.

$$recall = \frac{TP}{TP + FN} \quad (3)$$

where *recall* quantifies the model's ability to identify all positive instances.

$$F_1 = 2 * \frac{precision * recall}{precision + recall} = \frac{2TP}{2TP + FP + FN} \quad (4)$$

where F_1 *score* is the harmonic mean of precision and recall, balancing the trade-off between these two quantities and proving especially informative under class-imbalance conditions such as those present in the FER2013 dataset [32].

4. Results

The ResNet-50 model, pre-trained on ImageNet [33], was fine-tuned [34] on the FER2013 dataset. This model was trained for a maximum of 100 epochs with early stopping activated; training concluded at epoch 40, selecting the checkpoint with the highest validation score [35].

During the first ten epochs both training and validation loss declined rapidly, intersecting at roughly 1.7, after which the validation curve stabilized (Figure 3).

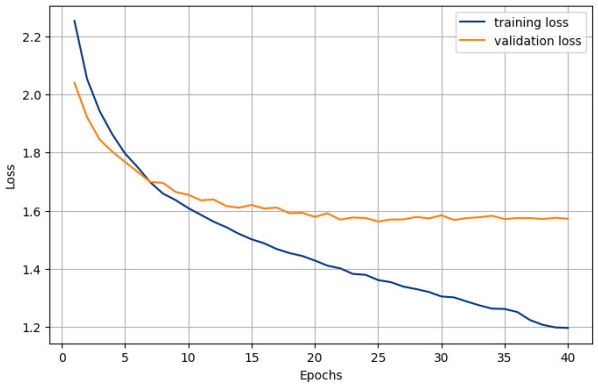

Figure 3: Training and validation loss over epochs for ResNet-50.

Validation accuracy increased during the initial training phase and subsequently plateaued at approximately 69% (Figure 4).

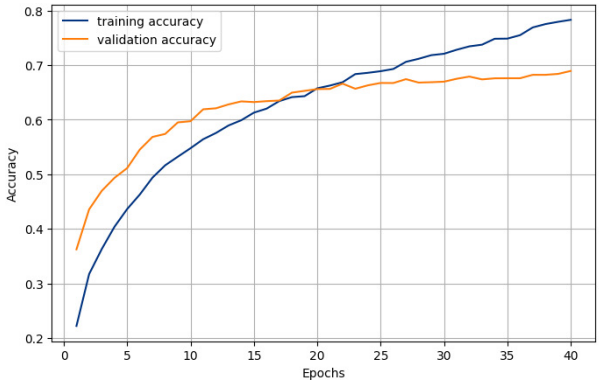

Figure 4: Training and validation accuracy over epochs for ResNet-50.

The normalized confusion matrix indicates the extent to which the model can distinguish individual emotions from one another (Figure 5). The diagonal values represent the recall for each class, whereas the off-diagonal entries in a given row show the percentage of that class that was misassigned to other labels. Happiness and surprise are the most distinctly recognized emotions, with 86.70% of happiness images and 83.27% of surprise images classified correctly, and most of their remaining misclassifications are evenly spread across the other five classes.

Conversely, fear emerges as the most challenging class, being misclassified most frequently as sadness or anger. Some overlap is visible between neutral and sadness as well, reflecting the subtle difference in facial cues between these two moods. Disgust, despite having the fewest samples in the dataset, achieves a recall of 65.77%, indicating that the class-balanced loss succeeded in preventing systematic neglect of this minority category.

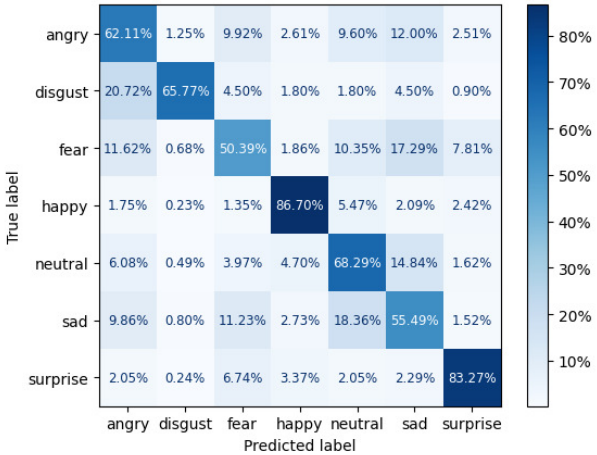

Figure 5: Normalized confusion matrix obtained by the ResNet-50 model on the FER2013 validation set.

The weighted-averaged F1-score reaches 68.89%, confirming that the network retains balanced performance despite the pronounced class imbalance. The macro-averaged F1-score is 67.13%, indicating that no single category dominates the overall performance and that the model maintains a comparable level of sensitivity across all seven emotions (Table 2).

Table 2: Evaluation metrics for the ResNet-50 model on the FER2013 dataset

ResNet-50	Precision (%)	Recall (%)	F1-Score (%)
angry	60.53	62.11	61.31
disgust	64.04	65.77	64.89
fear	58.31	50.39	54.06
happy	90.26	86.70	88.44
neutral	60.79	68.29	64.32
sad	56.35	55.49	55.92
surprise	78.73	83.27	80.94
accuracy			68.93
macro average			67.13
weighted average			68.89

The per-class and aggregate metrics confirm that the ResNet-50 performs consistently across all classes and attains an overall accuracy of 68.93% on FER2013.

The ViT-B/16 model, initialized with ImageNet weights [33], was fine-tuned [34] on FER2013 dataset. This model was trained with early-stopping monitor with a patience window of twenty epochs, which identified the optimal checkpoint at epoch 20, indicating that the transformer reached its peak validation performance sooner than the convolutional model [35]. Training and validation loss exhibited a rapid initial decline followed by a gradual, monotonic descent (Figure 6).

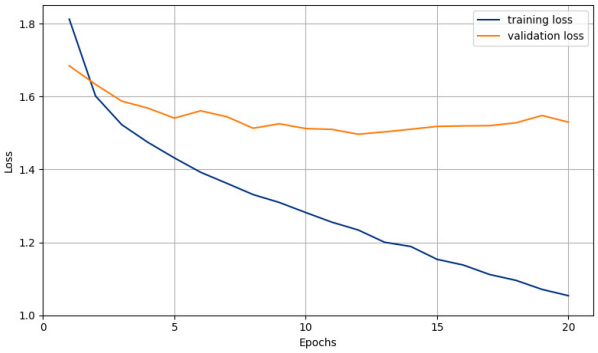

Figure 6: Training and validation loss over epochs for ViT-B/16.

Validation accuracy increased progressively to a maximum of 71.30%, indicating stable generalization throughout training (Figure 7).

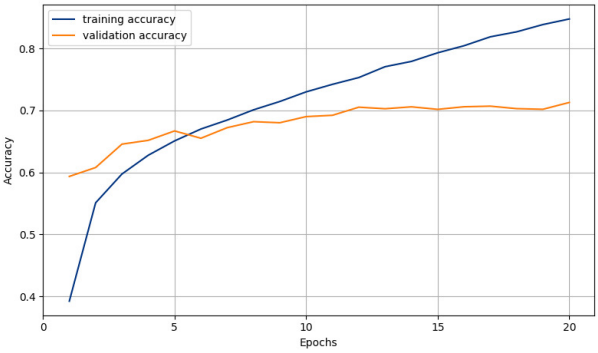

Figure 7: Training and validation loss over epochs for ViT-B/16.

The normalized confusion matrix (Figure 8) provides a detailed view of class-level behavior and highlights the model's relative strengths and weaknesses across categories. The model achieves its highest recall for happiness at 88.44% and for surprise at 85.32%, underscoring its capacity to capture the distinctive facial patterns associated with these high-expression emotions. Recall for disgust reaches 72.07%, demonstrating effective learning for this minority class despite limited sample availability.

In contrast, fear remains the most challenging emotion; the majority of its misclassifications are funnelled into the anger and sadness classes, indicating that the decision boundary for fear still overlaps most strongly with these two negative emotions. Neutral is most frequently confused with sadness; 15% of neutral images are predicted as sadness, underscoring the subtle and subtle and ambiguous visual distinction between these two low-intensity expressions.

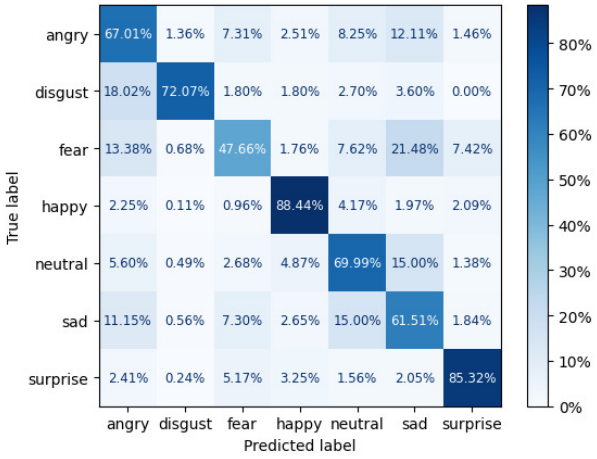

Figure 8: Normalized confusion matrix obtained by the ViT-B/16 model on the FER2013 validation set.

The quantitative evaluation presented in the classification report (Table 3) yields a weighted-averaged F1-score of 71.16% and a macro-averaged F1-score of 69.82%, confirming that performance is distributed across all seven emotions rather than being dominated by a subset of categories.

Table 3: Evaluation metrics for the ViT-B/16 model on the FER2013 dataset

ViT-B/16	Precision (%)	Recall (%)	F1-Score (%)
angry	60.17	67.01	63.41
disgust	68.38	72.07	70.18
fear	65.59	47.66	55.20
happy	90.54	88.44	89.48
neutral	66.54	69.99	68.22
sad	57.07	61.51	59.20
surprise	80.94	85.32	83.07
accuracy			71.30
macro average			69.82
weighted average			71.16

ViT-B/16 model fine-tuned on FER2013 achieved an overall classification accuracy of 71.30%, supported by consistently high scores across precision, recall and F1-measure. These results confirm that the vision transformers can reliably identify each of the seven basic emotions while maintaining stable generalization throughout training.

To better understand the internal mechanisms of both models, Grad-CAM visualizations (Figure 9) were employed to identify which regions of the input images contributed most to the final classification decisions [36].

The ResNet-50 model consistently focuses on localized facial regions, particularly the eyes, mouth, and eyebrows, suggesting its reliance on well-defined facial landmarks. In contrast, the ViT-B/16 model exhibits more distributed attention patterns, often capturing broader contextual features across the entire face. While this broader focus may offer resilience to noise or occlusion, it also occasionally results in less concentrated activation, especially in ambiguous expressions.

Figure 9: Grad-CAM visualizations for selected images from the FER2013 dataset, one per emotion category.

These observations underscore the architectural distinctions between convolutional and transformer-based approaches in facial expression recognition. The contrast in activation patterns highlights differing strategies in feature prioritization, with ResNet-50 leveraging spatial hierarchies and ViT-B/16 capitalizing on global context. Such insights are valuable for selecting models in applications where interpretability, robustness, and sensitivity to specific facial cues are critical.

5. Conclusions

ResNet-50 and Vision Transformer, both pre-trained on ImageNet and fine-tuned on the FER2013 dataset, demonstrated strong capabilities in facial emotion recognition. ViT-B/16 achieved the highest validation accuracy at 71.30%, slightly outperforming ResNet-50, which reached 68.93%, particularly in recognizing emotions such as disgust and sadness. In contrast, ResNet-50 showed more stable behavior across training and generated clearer, spatially focused Grad-CAM visualizations. Despite occasional confusion between ambiguous classes like fear and sad, both models showed reliable generalization.

The complementary strengths of convolutional and transformer-based approaches suggest promising directions for further refinement, particularly in improving class-level precision and leveraging hybrid or ensemble strategies for enhanced interpretability and robustness in real-world emotion recognition tasks.

It should be noted that both models were assessed only on the FER2013 dataset, which is limited to low-resolution, grayscale images. While FER2013 is a widely adopted reference set, its characteristics may not capture the full variability encountered in higher-resolution or in-the-wild scenarios. Extending the evaluation to additional, more diverse datasets in future work would provide a broader view of the models' generalizability.

Future work may also involve the use of other neural network models [37], methods of aggregating classification results or ensemble learning [38]. Other methods of visualizing results such as SHAP Values [39] are also worth considering.

References

[1] M. Dhuheir, A. Albaseer, E. Baccour, A. Erbad, M. Abdallah, M. Hamdi, Emotion Recognition for Healthcare Surveillance Systems Using Neural Networks: A Survey, In 2021 International Wireless Communications and Mobile Computing Conference (IWCMC) (2021) 681–687, https://doi.org/10.1109/IWCMC51323.2021.9498861.

[2] T. Olajumoke, B. Al-Bander, Emotion-aware psychological first aid: Integrating BERT-based emotional distress detection with Psychological First Aid-Generative Pre-Trained Transformer chatbot for mental health support, Cognitive Computation and Systems 7(1) (2025) e12116, https://doi.org/10.1049/ccs2.12116.

[3] B. C. Ko, A Brief Review of Facial Emotion Recognition Based on Visual Information, Sensors 18(2) (2018) 401, https://doi.org/10.3390/s18020401.

[4] K. O'Shea, R. Nash, An Introduction to Convolutional Neural Networks, arXiv preprint arXiv:1511.08458 (2015), https://doi.org/10.48550/arXiv.1511.08458.

[5] M. Krichen, Convolutional Neural Networks: A Survey, Computers 12(8) (2023) 151, https://doi.org/10.3390/computers12080151.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, In International Conference on Learning Representations (2021) 2010.11929, https://doi.org/10.48550/arXiv.2010.11929.

[7] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, Z. He, A Survey of Visual Transformers, IEEE Transactions on Neural Networks and Learning Systems 35(6) (2024) 7478–7498, https://doi.org/10.1109/TNNLS.2022.3227717.

[8] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, Communications of the ACM 60(6) (2017) 84–90, https://doi.org/10.1145/3065386.

[9] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, In International conference on machine learning 139 (2021) 10347–10357, https://doi.org/10.48550/arXiv.2012.12877.

[10] K. Islam, Recent Advances in Vision Transformer: A Survey and Outlook of Recent Work, arXiv preprint arXiv:2203.01536 (2023), https://doi.org/10.48550/arXiv.2203.01536.

[11] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, L. Wang, Vision Transformers for Image Classification: A Comparative Survey, Technologies 13(1) (2025) 32, https://doi.org/10.3390/technologies13010032.

[12] P. Giannopoulos, I. Perikos, I. Hatzilygeroudis, Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013, Advances in Hybridization of Intelligent Methods, Springer (2018) 1–16, https://doi.org/10.1007/978-3-319-66790-4_1.

[13] H. Mikami, H. Suganuma, P. U-chupala, Y. Tanaka, Y. Kageyama, Massively distributed SGD: ImageNet/ResNet-50 Training in a Flash, arXiv preprint arXiv:1811.05233 (2019), https://doi.org/10.48550/arXiv.1811.05233.

[14] P. Powroźnik, Polish emotional speech recognition using artifical neural network, Advances in Science and Technology Research Journal 8(24) (2014) 24–27, https://doi.org/10.12913/22998624/562.

[15] P. Powroźnik, D. Czerwiński, Spectral methods in Polish emotional speech recognition, Advances in Science and Technology Research Journal 10(32) (2016) 73–81, https://doi.org/10.12913/22998624/65138.

[16] I.J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, Y. Bengio, Challenges in Representation Learning: A report on three machine learning contests, In International Conference on Neural Information Processing (ICONIP) (2013) 117–124, https://doi.org/10.48550/arXiv.1307.0414.

[17] M. A. Altaha, I. Jarraya, T. M. Hamdani, A. M. Alimi, Facial Expression Recognition based on ArcFace Features and TinySiamese Network, In 2023 International Conference on Cyberworlds (CW) (2023) 24–31, https://doi.org/10.1109/CW58918.2023.00014.

[18] H. Sheng, M. Lau, Optimising Facial Expression Recognition: Comparing ResNet Architectures for Enhanced Performance, Proceedings of the 11th International Conference of Control, Dynamic Systems, and Robotics (CDSR) (2024) 123, https://doi.org/10.11159/cdsr24.123.

[19] P. Soni, H. Jain, P. Bharti, A.K. Dubey, Identification of Facial Expressions using Deep Neural Networks, Fusion: Practice and Applications 2(1) (2020) 22–30, https://doi.org/10.54216/FPA.020101.

[20] H. Li, Q. Li, End-to-End Training for Compound Expression Recognition, Sensors 20(17) (2020) 1–25, https://doi.org/10.3390/s20174727.

[21] S. Bobojanov, B.M. Kim, M. Arabboev, S. Begmatov, Comparative Analysis of Vision Transformer Models for Facial Emotion Recognition Using Augmented Balanced Datasets, Applied Sciences 13(22) (2023) 12271, https://doi.org/10.3390/app132212271.

[22] J. Soni, N. Prabakar, H. Upadhyay, Vision Transformer-Based Emotion Detection in HCI for Enhanced Interaction, In 15th International Conference on Intelligent Human Computer Interaction (IHCI) (2023) 76–86, https://doi.org/10.1007/978-3-031-53827-8_8.

[23] H. Song, Facial Expression Recognition with ViT Considering All Tokens towards More Informative Self-attention Outputs, Highlights in Science, Engineering and Technology 41 (2023) 72–79, https://doi.org/10.54097/hset.v41i.6745.

[24] M. Bie, H. Xu, Y. Gao, K. Song, X. Che, Swin-FER: Swin Transformer for Facial Expression Recognition, Applied Sciences 14(14) (2024) 6125, https://doi.org/10.3390/app14146125.

[25] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 770–778, https://doi.org/10.1109/CVPR.2016.90.

[26] K. He, X. Zhang, S. Ren, J. Sun, Identity Mappings in Deep Residual Networks, In Computer Vision–ECCV 2016: 14th European Conference (2016) 630–645, https://doi.org/10.1007/978-3-319-46493-0_38.

[27] F. X. Gaya-Morey, C. Manresa-Yee, C. Martinie, J. M. Buades-Rubio, Evaluating Facial Expression Recognition Datasets for Deep Learning: A Benchmark Study with Novel Similarity Metrics, arXiv preprint arXiv:2503.20428 (2025), https://doi.org/10.48550/arXiv.2503.20428.

[28] M. Buda, A. Maki, M. A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural Networks 106 (2018) 249–259, https://doi.org/10.1016/j.neunet.2018.07.011.

[29] C. Shorten, T. M. Khoshgoftaar, A survey on Image Data Augmentation for Deep Learning, Journal of Big Data 6(1) (2019) 1–48, https://doi.org/10.1186/s40537-019-0197-0.

[30] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management 45(4) (2009) 427–437, https://doi.org/10.1016/j.ipm.2009.03.002.

[31] P. Machart, L. Ralaivola, Confusion Matrix Stability Bounds for Multiclass Classification, arXiv preprint arXiv:1202.6221 (2012), https://doi.org/10.48550/arXiv.1202.6221.

[32] S. Sathyanarayanan, B. R. Tantri, Confusion Matrix-Based Performance Evaluation Metrics, African Journal of Biomedical Research 27(4S) (2024) 4023–4031, https://doi.org/10.53555/AJBR.v27i4S.4345.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, In 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009) 248–255, https://doi.org/10.1109/CVPR.2009.5206848.

[34] Z. Mai, A. Chowdhury, P. Zhang, C.-H. Tu, H.-Y. Chen, V. Pahuja, T. Berger-Wolf, S. Gao, C. Stewart, Y. Su, W.-L. Chao, Fine-Tuning is Fine, if Calibrated, Advances in Neural Information Processing Systems 37 (2024) 136084–136119, https://doi.org/10.48550/arXiv.2409.16223.

[35] L. Prechelt, Early Stopping - But When?, Neural Networks: Tricks of the trade (2002) 55–69, https://doi.org/10.1007/3-540-49430-8_3.

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, In 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 618–626, https://doi.org/10.1109/ICCV.2017.74.

[37] P. Powroznik, P. Wojcicki, S. W. Przylucki, Scalogram as a representation of emotional speech, IEEE Access 9 (2021) 154044–154057, https://doi.org/10.1109/ACCESS.2021.3127581.

[38] M. Skublewska-Paszkowska, P. Powroznik, P. Karczmarek, E. Lukasik, Aggregation of tennis groundstrokes on the basis of the choquet integral and its generalizations, In 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) IEEE (2022) 1–8, https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882592.

[39] L. Qin, Y. Zhu, S. Liu, X. Zhang, Y. Zhao, The Shapley Value in Data Science: Advances in Computation, Extensions, and Applications, Mathematics 13(10) (2025) 1581, https://doi.org/10.3390/math13101581.