Proceedings of the 11th International Conference of Control Systems, and Robotics (CDSR 2024)
Chestnut Conference Centre - University of Toronto, Toronto, Canada – 10 -12 June, 2024
Paper No. 123
DOI: 10.11159/cdsr24.123

Optimising Facial Expression Recognition: Comparing ResNet Architectures for Enhanced Performance

Haoliang Sheng, MengCheng Lau
Laurentian University
School of Engineering and Computer Science
935 Ramsey Lake Rd, Sudbury, Ontario, Canada
hsheng@laurentian.ca; mclau@laurentian.ca

Abstract **-** This study investigates how ResNet architectures (ResNet18, ResNet34, ResNet50) perform in recognising facial expressions using the FER-2013 dataset. It applies transfer learning, fine-tunes the models, and conducts real-time tests to evaluate their performances. The results show that ResNet18 achieves the best balance between accuracy and efficiency, despite its simplicity. This research highlights the essential role that comprehensive datasets play in improving model generalisation, particularly for underrepresented expressions. It points out the need for a nuanced balance between model complexity, computational efficiency, and suitability for real-world applications, making a significant contribution to the field of real-time facial expression recognition.

Keywords**:** Facial Expression Recognition; ResNet Architectures; Transfer Learning; Real-time Webcam

1. Introduction

Facial expression recognition is crucial in fields such as human-computer interaction, security, and health assessment. The use of deep learning, particularly Convolutional Neural Networks (CNNs), has greatly enhanced its capabilities, providing major improvements over older methods [1][2][3][4]. Among these advancements, Deep Residual Networks (ResNets) stand out for their ability to tackle deep network training challenges, offering more accurate and efficient recognition solutions [5][6]. Their use has particularly been beneficial in recognizing subtle changes in facial expressions, essential for seamless human-computer interactions.

The need for real-time and dynamic facial expression recognition has grown with advancements in technology, especially in surveillance and real-time communications, where high accuracy and low latency are paramount [7]. ResNets have been instrumental in addressing these needs due to their superior performance and computational efficiency.

The success of facial expression recognition also depends on advanced feature extraction methods like higher-order statistical features and Local Binary Patterns, although integrating these into ResNet architectures yielded minimal performance gains in our trials [8][9][10][11][12][13]. This exploration, though not central to our study due to its limited impact, indicates our attempts to refine recognition models, which we don't detail further. However, the application of these technologies in areas such as education and social cognition highlights their broad utility [3][14].

Our study begins with an analysis of the FER-2013 dataset [15], emphasizing its diversity and the challenges it presents, which are pivotal for evaluating advanced deep learning models in facial expression recognition. The methodology section outlines our approach using ResNet architectures and transfer learning, detailing the fine-tuning necessary for emotion recognition tasks. Experimental results compare model performances, assessing accuracy, efficiency, and real-time applicability through webcam tests. We conclude by highlighting the balance between model complexity and training data comprehensiveness, proposing future research avenues for improving model sensitivity and real-time performance.

2. Dataset

The FER-2013 dataset [15] provides a comprehensive collection specifically designed for facial expression recognition tasks. It originally includes seven distinct emotion categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The dataset comprises a total of 35,886 images, with 28,708 designated for training and 3,589 for testing. Each image within the dataset is a grayscale picture of 48x48 pixels, offering a uniform standard for processing and analysis. The images' relatively low resolution, in contrast to the high-resolution input from webcams, poses a significant challenge in ensuring the model's effectiveness in real-world scenarios.

To bridge this gap and boost the model's robustness, we have opted to use grayscale images in our training process. This strategy achieves two objectives: first, it simplifies the model by concentrating on structural features rather than colour information, which holds less significance for recognising facial expressions. Second, it replicates the conditions of real-time webcam input, where lighting conditions and background variability can drastically change colour information. By adopting this grayscale method and combining it with our other data augmentation techniques, such as random flips, rotations, and adjustments in brightness, contrast, and affine transformations, we ensure the development of a robust model capable of accurately classifying facial expressions in a variety of dynamic real-world settings.

3. Methodology

3.1. Transfer Learning Approach

In our research, we have adopted the concept of transfer learning to exploit the advanced capabilities of pre-trained models for feature extraction. Specifically, we chose the ResNet18, ResNet34, and ResNet50 architectures—variants of the Residual Network—guided by our aim to find a careful balance between model depth and computational efficiency, essential for real-time application deployment.

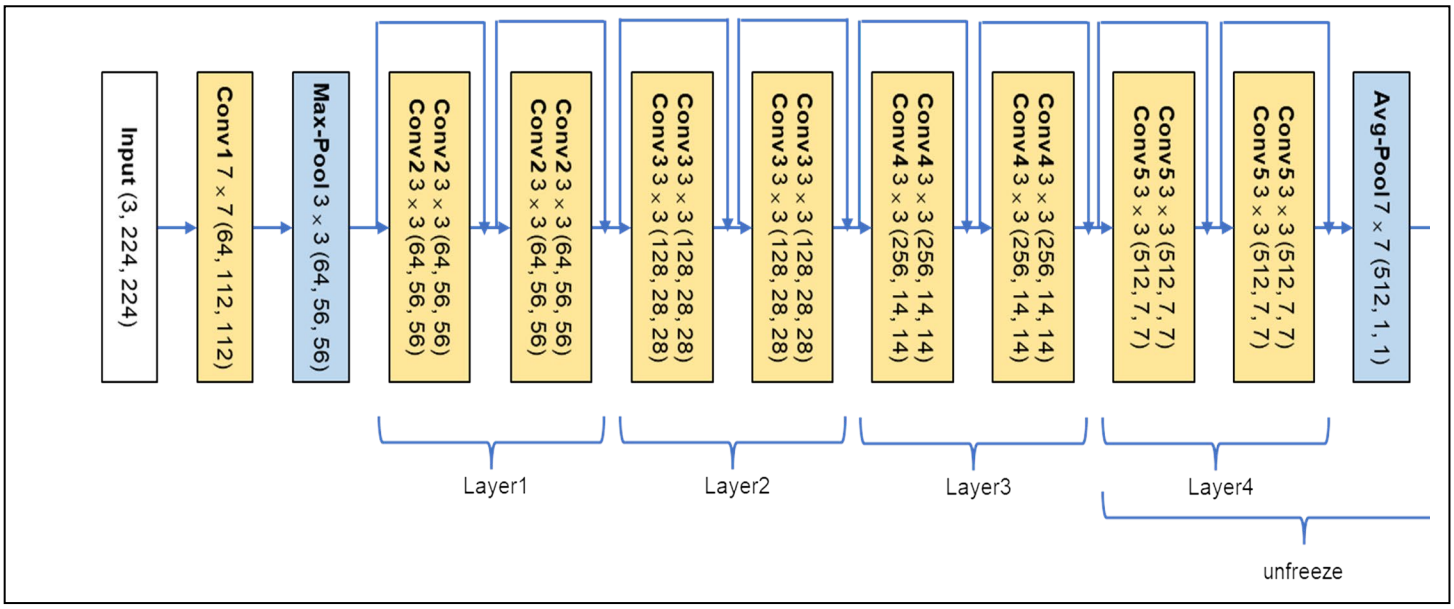

Fig. 1: Architecture of ResNet18 model.

The illustration in Fig. 1 outlines the architecture configurations of the ResNet18 model. The text in each box indicates the module name in bold, the kernel size, and the output dimensions in brackets. We customised the architecture of these chosen ResNet models by adjusting their final fully connected, *fc* layer to classify images into seven specific classes: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. This modification was crucial to customise the models for the unique demands of our study, especially for the classification task using the FER-2013 dataset—a dataset noted for its facial expression recognition challenges. We aimed to create a framework that demonstrates how we have modified the network.

To optimally fine-tune the models for our task, we performed a fine-tuning process. This involved freezing the initial layers of the models to retain the features learned from the ImageNet dataset, while training the latter layers to adapt to our specific classification requirements. This strategy enabled us to significantly reduce training time and resource use, maintaining high accuracy and efficiency for real-time facial expression classification. A novel aspect of our methodology was our decision to not only modify the last fully connected layer but also to unfreeze and train the penultimate layer, *layer4*. We based this decision on our finding that allowing *layer4* to learn from our specific dataset could enhance the model's

classification accuracy. This approach indeed led to a marked improvement in training accuracy, which jumped from 41.6% to an impressive 91.2%. This strategic use of transfer learning, alongside selective fine-tuning of network layers, highlights our dedication to developing an efficient and accurate system for real-time facial expression recognition. Through this method, we have effectively harnessed existing deep learning architectures, adapting them to confront the particular challenges posed by our research field.

3.2. Real-Time Processing

The core of our methodology for real-time prediction involves deploying a pre-trained model to detect and classify facial expressions in real time, utilising webcam input. This section describes the technical implementation designed to enable live classification of facial expressions, fostering interactive user engagement. We have constructed the system within a Python environment, utilising the OpenCV library for video capture and face detection, alongside PyTorch for deploying the trained model. The process initiates with activating the user's webcam, which continuously streams live images into the system. We then employ OpenCV's `*CascadeClassifier*`, specifically the `*haarcascade_frontalface_default.xml*`, to process these images and detect the presence of a face. Upon detecting a face within the video frame, the system extracts the region of interest (ROI), the detected face. We subsequently preprocess and transform this face ROI to meet the input specifications of the pre-trained model. This transformation typically includes resizing the image, normalising pixel values, and converting the image into a tensor format apt for model input. After preprocessing, we feed the face ROI tensor into the trained ResNet model, set in evaluation mode to ensure batch normalisation and dropout layers function in inference mode. The model then predicts the facial expression by outputting a class label corresponding to the detected expression.

The real-time prediction functionality resides within a user-friendly interface that exhibits the live video feed from the webcam. When the system identifies and classifies a facial expression, it outlines the detected face with a bounding box and displays the predicted expression label on the screen. This enables users to witness the classification results in real time as they exhibit various facial expressions to the webcam. The real-time prediction system's implementation is encapsulated in the `*realtime_prediction*` function, which accepts several parameters including the model, transformation procedures, display duration for the classification label, and the preferred computational device (CPU or GPU). This function controls the entire procedure from video capture, face detection, and expression classification to result display, ensuring a smooth and interactive user experience.

4. Experimental Results and Discussion

This section provides a comprehensive analysis of the performance and effectiveness of various ResNet architectures—ResNet18, ResNet34, and ResNet50—in tasks related to facial expression recognition. We carefully record and examine the training performance, covering both training accuracy and time, for later analysis on the computational efficiency and learning capabilities of each model. Additionally, we evaluate the testing performance across different models and within specific classes, illustrating how each architecture performs overall and in identifying the seven distinct facial expressions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. We broaden our analysis to include real-world relevance by assessing the models in a live setting using a webcam, thereby offering a complete picture of each model's practical value.

4.2. Training Performance

Fig. 2 displays the training performance for each ResNet model, covering training loss, training accuracy, and training time. Following is our observations:

- Performance Over Time: Initially, ResNet50 led with the best performance, yet by the 9th epoch, ResNet18, the simplest model, overtook it. This indicates that although deeper networks might initially excel, simpler architectures can catch up and even surpass them as training advances.
- Consistency in Performance: ResNet34 consistently delivered the least effective performance throughout the training. Despite its medium complexity, it never reached the higher accuracy rates of ResNet18 or ResNet50 at any measured epoch.
- Training Efficiency: By the 9th epoch, ResNet18 not only matched ResNet50 in accuracy but also required significantly less training time, showcasing greater training efficiency.

- Resource Utilisation: ResNet50 took the longest to train, indicating a higher computational cost, which might not be justifiable since its performance was equalled and then exceeded by the simpler ResNet18.
- Model Complexity vs. Performance Trade-Off: There seems to be a diminishing return on added model complexity, as illustrated by ResNet50, where the extra complexity does not equate to a proportionate improvement in performance in later epochs.
- Practical Implications for Model Selection: Selecting a model for practical applications necessitates a consideration of the trade-off between accuracy, training time, and computational resources. ResNet18 might present the best balance in scenarios where resource efficiency is crucial.
- Potential for Further Improvement: None of the models showed a plateau in accuracy after 20 epochs, suggesting room for further enhancement in model performance with more training. This hints that continued training could yield better accuracy, provided that overfitting is managed.

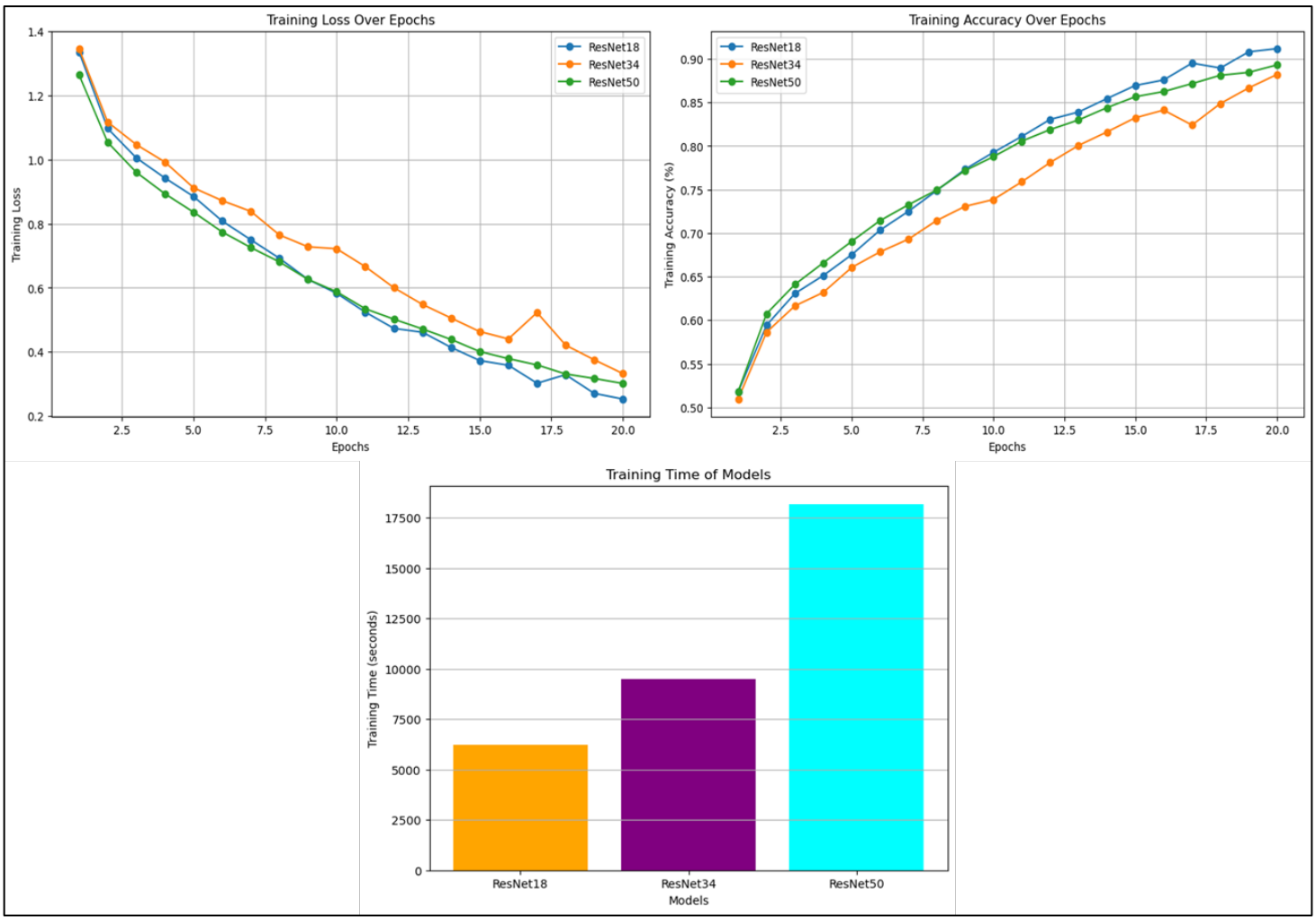

Fig. 2: Training performances of ResNet models.

4.2. Testing Performance

Table 1: Test performance for each ResNet model.

Model	Precision	Recall	F1-Score	Support	Accuracy	AUC
resnet18	0.648	0.650	0.646	7178	0.650	0.898
resnet34	0.652	0.651	0.650	7178	0.651	0.893
resnet50	0.664	0.654	0.658	7178	0.654	0.885

Table 1 provides a comprehensive overview of the testing performance for three distinct ResNet models: ResNet18, ResNet34, and ResNet50. Six key metrics are reported for each model, which include Precision, Recall, F1-Score, Support, Accuracy, and AUC (Area Under the Curve). Notably, the Precision, Recall, and F1-Score values correspond to the '*weighted avg*' figures obtained from the `*classification_report*` function of `*sklearn.metrics*`, which accounts for the imbalance in the support of each class. The *Support* is consistent across all models, signifying the total count of test dataset instances used to evaluate each model. The *Accuracy* metric reported is the overall accuracy across all classes. The *AUC* values represent the arithmetic mean of the *AUC* for the seven different facial expressions, providing a consolidated view of the model's discriminative power.

Table 2: Test performance for each Facial Expression.

Expression	Precision	Recall	F1-Score	Support	AUC
angry	0.560	0.590	0.571	958	0.872
disgust	0.727	0.583	0.647	111	0.958
fear	0.522	0.483	0.500	1024	0.805
happy	0.849	0.857	0.851	1774	0.965
neutral	0.589	0.619	0.601	1233	0.856
sad	0.516	0.507	0.511	1247	0.832
surprise	0.811	0.768	0.789	831	0.956

Table 2 offers a detailed analysis of classification performance metrics for seven different facial expressions: angry, disgust, fear, happy, neutral, sad, and surprise. These metrics encompass Precision, Recall, F1-Score, Support, and AUC, representing the arithmetic mean values from the performance of three ResNet models (ResNet18, ResNet34, and ResNet50). The following are our observations:

- Model Complexity vs. Performance: Increased complexity through deeper networks, such as ResNet50, does not always lead to improved performance metrics. Often, ResNet18 either surpasses or equals ResNet50 in accuracy and AUC, underscoring the importance of adopting a balanced approach in model selection.
- Precision-Recall Trade-off: The models exhibit a nuanced trade-off between precision and recall across different expressions, with no single model consistently outperforming others in all categories. This stresses the significance of evaluating multiple metrics for a comprehensive performance assessment.
- Variability in Expression Recognition: The models demonstrate significant variability in recognising different facial expressions. Expressions like 'happy' and 'surprise' are classified with high precision and recall, whereas recognising 'fear' and 'sad' poses greater challenges, pointing towards the need for model adjustments or targeted data augmentation.

- AUC: AUC scores differ among expressions, with 'happy' and 'disgust' achieving the highest scores. This indicates that the models more easily distinguish these expressions, likely due to more distinct or consistent features.
- Impact of Support on Performance: The support number, or the count of instances, does not directly correlate with performance, as illustrated by the 'fear' and 'sad' expressions. This highlights the importance of both the quantity and quality of training data.
- Opportunity for Model Optimisation: The close range of metrics such as precision, recall, and F1-scores suggests room for improvement through hyperparameter tuning, further training, or experimenting with alternative architectures to boost performance.
- Importance of a Balanced Dataset: Differences in performance across expressions stress the need for a balanced and representative training dataset, crucial for enhancing model generalisation and improving accuracy on less precisely classified expressions.

4.3. Real-Time Webcam Performance

In this section on Real-Time Webcam, we explore the practical application of our trained models in a live environment. Using a standard webcam, we carried out real-time facial expression recognition to assess the models' capacity to interpret and classify expressions as they naturally occur.

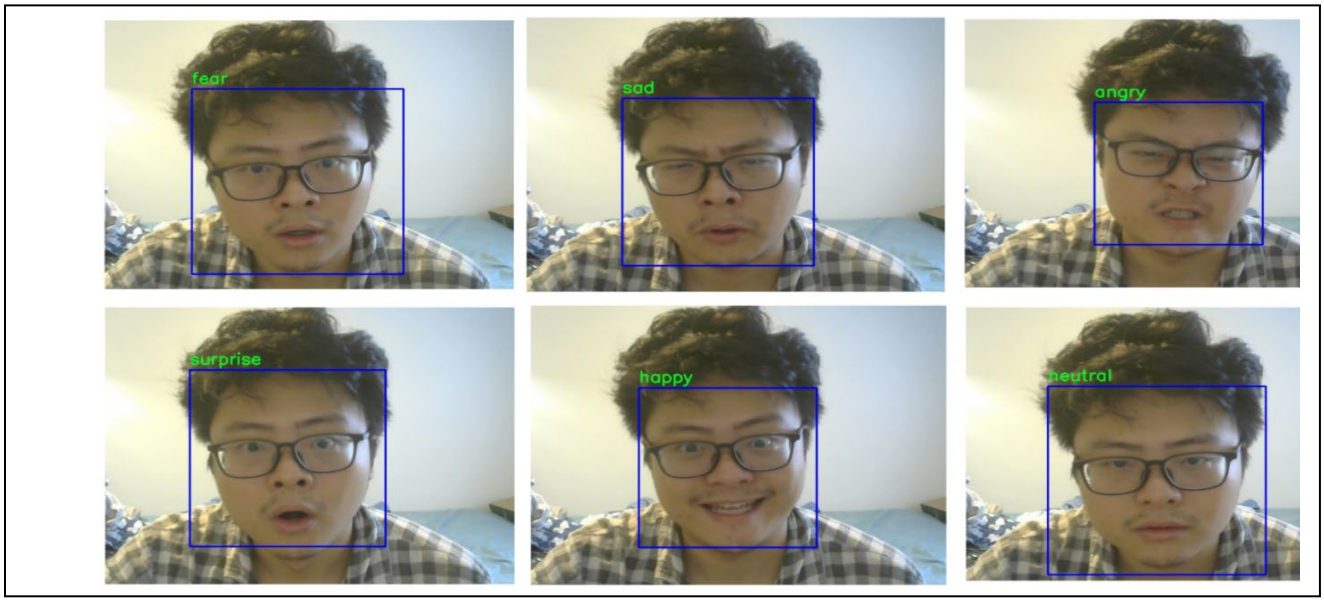

Fig. 5: Real-Time Facial Expression Detection Snapshots.

We captured the real-time performance through a series of snapshots that displayed the models' analytical interpretation of six discernible facial expressions: Angry, Fear, Happy, Sad, Surprise, and Neutral. Each snapshot matches a detected facial expression, apart from 'Disgust', which we notably omitted from our real-time analysis. The absence of 'Disgust' stems not from oversight but as a result of the initial training phase. The FER-2013 dataset, used for training our models, provides limited representation of the 'Disgust' expression, leading to the models' ineffectiveness in recognising and classifying 'Disgust' confidently in real-world situations. This limitation highlights a critical challenge in machine learning and deep learning projects: the performance of models closely links to the quality and breadth of the training data. The training set's insufficient instances of the 'Disgust' expression have created a proficiency gap, evident during live webcam testing.

Our real-time analysis not only verified the models' ability to operate interactively but also observed that the models showed variable accuracy across different expressions, particularly excelling in detecting 'Happy' and 'Surprise', which

benefited from robust datasets during training. The volume of expression data in the training set and the models' real-time detection accuracy showed a clear correlation, reinforcing the maxim that 'data is king' in machine learning. Despite the generally impressive performance under real-time conditions, the models' adaptability to less represented expressions, such as 'Disgust', was limited, indicating a need for dataset enrichment and possibly model retraining. The analysis of real-time performance offers essential insights into the models' readiness for deployment in real-world scenarios, where a variety of expressions and unpredictability are standard.

In summary, while the models demonstrated notable real-time performance, their inability to detect 'Disgust' underscores an important area for future research and development: enhancing the dataset to encompass underrepresented expressions and refining the models to improve their detection capabilities across a broader expression spectrum.

5. Conclusion

The conclusion brings together the findings of the research, emphasising the subtle interaction between model architecture complexity and performance in tasks of facial expression recognition. It underlines the effectiveness of ResNet18 in achieving a balance between computational efficiency and predictive accuracy, challenging the assumption that greater complexity leads to better results. The study further uncovers essential insights into the models' abilities and limitations in real-world settings, especially through testing with a real-time webcam. These outcomes support a strategic approach to choosing models, highlighting the significance of comprehensive datasets for training to improve model generalisation across various expressions. This research contributes to the ongoing discussion within machine learning communities on optimising model performance within practical limits, setting the stage for future developments in real-time facial expression recognition technologies.

References

[1] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 5325–5334. doi: 10.1109/CVPR.2015.7299170.

[2] Z. Ming, J. Chazalon, M. Muzzamil Luqman, M. Visani, and J.-C. Burie, "FaceLiveNet: End-to-End Networks Combining Face Verification with Interactive Facial Expression-Based Liveness Detection," in *2018 24th International Conference on Pattern Recognition (ICPR)*, Aug. 2018, pp. 3507–3512. doi: 10.1109/ICPR.2018.8545274.

[3] Y. K. Bhatti, A. Jamil, N. Nida, M. H. Yousaf, S. Viriri, and S. A. Velastin, "Facial Expression Recognition of Instructor Using Deep Features and Extreme Learning Machine," *Computational Intelligence and Neuroscience*, vol. 2021, p. e5570870, May 2021, doi: 10.1155/2021/5570870.

[4] Y. Gan, "Facial Expression Recognition Using Convolutional Neural Network," in *Proceedings of the 2nd International Conference on Vision, Image and Signal Processing*, in ICVISP 2018. New York, NY, USA: Association for Computing Machinery, Aug. 2018, pp. 1–5. doi: 10.1145/3271553.3271584.

[5] B. K. Durga and V. Rajesh, "A ResNet deep learning based facial recognition design for future multimedia applications," *Computers and Electrical Engineering*, vol. 104, p. 108384, Dec. 2022, doi: 10.1016/j.compeleceng.2022.108384.

[6] B. Li and D. Lima, "Facial expression recognition via ResNet-50," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 57–64, Jun. 2021, doi: 10.1016/j.ijcce.2021.02.002.

[7] N. Hajarolasvadi and H. Demirel, "Deep facial emotion recognition in video using eigenframes," *IET Image Processing*, vol. 14, no. 14, pp. 3536–3546, 2020, doi: 10.1049/iet-ipr.2019.1566.

[8] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Trans. Affective Comput.*, vol. 13, no. 3, pp. 1195–1215, Jul. 2022, doi: 10.1109/TAFFC.2020.2981446.

[9] H. Ali, M. Hariharan, S. Yaacob, and A. H. Adom, "Facial Emotion Recognition Based on Higher-Order Spectra Using Support Vector Machines," *Journal of Medical Imaging and Health Informatics*, vol. 5, no. 6, pp. 1272–1277, Nov. 2015, doi: 10.1166/jmihi.2015.1527.

[10] S. Sawardekar and S. R. Naik, “Facial Expression Recognition using Efficient LBP and CNN,” vol. 05, no. 06, 2018.
[11] Y. Lv, Z. Feng, and C. Xu, “Facial expression recognition via deep learning,” in *2014 International Conference on Smart Computing*, Nov. 2014, pp. 303–308. doi: 10.1109/SMARTCOMP.2014.7043872.
[12] T. Debnath, Md. M. Reza, A. Rahman, A. Beheshti, S. S. Band, and H. Alinejad-Rokny, “Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity,” *Sci Rep*, vol. 12, p. 6991, Apr. 2022, doi: 10.1038/s41598-022-11173-0.
[13] G. Zeng, J. Zhou, X. Jia, W. Xie, and L. Shen, “Hand-Crafted Feature Guided Deep Learning for Facial Expression Recognition,” in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an, China: IEEE Press, May 2018, pp. 423–430. doi: 10.1109/FG.2018.00068.
[14] L. Graumann, M. Duesenberg, S. Metz, L. Schulze, O. T. Wolf, S. Roepke, C. Otte, and K. Wingenfeld, “Facial emotion recognition in borderline patients is unaffected by acute psychosocial stress,” Journal of Psychiatric Research, vol. 132, pp. 131–135, Jan. 2021, doi: 10.1016/j.jpsychires.2020.10.007.
[15] M. Sambare. (2020) “FER-2013Learn facial expressions from an image” [Online]. Available: https://www.kaggle.com/datasets/msambare/fer2013/data