Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A High-Fidelity Facial Expression Representation Framework Using Curated Multi- Dataset Learning

Himanshu Verma
Manipal University Jaipur
Nimish Kumar
B K Birla Institute of Engineering & Technology
Pankaj Vyas
pankaj.vyas@jaipur.manipal.edu

Manipal University Jaipur

Research Article

Keywords:

Posted Date: May 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9295088/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Current facial expression recognition (FER) models have a tendency to be unstable when trained on heterogeneous public data because of annotation noise, visual artifacts, and bias depending on the data set. This paper proposes a representation-based FER model of psychological stress measurement, where automated data cleaning and feature learning that is agnostic of architecture is preferred, as opposed to architectural complexity. Various publicly available FER datasets are combined to enhance the level of expression diversity and are filtered by an automated quality pipeline to eliminate blurred, low-resolution, and ambiguous images to produce a curated corpus with 16,828 images of six emotion categories. Frozen convolutional backbones, such as ResNet18 and MobileNetV2, are used to extract deep facial embeddings that can be fairly compared with each other and avoiding effects of fine-tuning. Such embeddings are tested with various classical machine-learning classifiers with similar experimental conditions. Findings indicate that ResNet18 embeddings are more discriminative than lightweight counterparts and that the best overall results are along with Random Forests classifiers which reach a peak accuracy of 70.34%. A further comprehensive analysis of optimizer shows that AdamW has the most consistent convergence and validation consistency across folds. Results validate that the dataset integrity and fidelity of representation are a dominating factor in FER performance outperforming the effects of end-to-end architectural complexity. The proposed framework will provide a reliable and reproducible stressful FER system, which will function under limited and noisy data environments.

1. Introduction

Facial expressions provide observable signals of human affective and behavioural states. Automatic FER can therefore be regarded as one of the central tasks of affective computing and behavioural analysis [1], [2]. In practice, FER face multiple issues such as huge variations of facial appearance, variable illumination, variable poses, as well as inconsistent image quality [3]. Deep learning and specifically CNNs have significantly enhanced FER performance compared to hand-crafted feature-based methods [4], [5], [17] is due to the innate capacity of CNNs to learn hierarchical representations of faces that to some extent are independent of changes in appearance. Nevertheless, FER models that have been trained on data that are publicly accessible have a tendency to generalize poorly and give unstable results on different datasets [6]. One of the contributing factors is dataset quality. FER datasets vary in terms of acquisition conditions, resolution, annotation protocols and classes distributions [22]. Controlled datasets like CK+ is a high quality dataset but exaggerated [7], while datasets like FER2013 reflect realistic conditions but suffer from low resolution, label noise, and ambiguous expressions [8]. Training on such data propagates noise into learned representations. Prior studies indicate that label noise and low-quality samples have disastrous effects on accuracy of classification and feature discriminability [9], [10].Multiple datasets can be combined and enhance variety and decrease data-related bias. Nonetheless, there are inconsistencies in the annotation of naive dataset aggregation, as well as visual artefacts [11]. This indicates that dataset curation and automated quality control are required in FER systems. In addition to the accuracy of classification, there is also the quality of

representation. Facial embeddings of high-fidelity facilitate the stability of downstream tasks, as well as enhance resistance between classes. Deep hierarchical representations obtained by residual network architectures are known to be effective at capturing small differences in faces [12]. Hybrid systems, which take the approach of using end-to-end training, are additionally prone to perform worse than deep feature extractors, typically in settings where the number of datasets is moderate or even partially noisy [13], [14].

This work is motivated by the limitations of current FER systems and aims at the strong learning of facial expression representation through the use of multi-datasets which are carefully curated. Multiple datasets (FER2013+ Humans Face Emotions+ CK+) are merged to increase variety and is screened using an automated pipeline to filter out incorrect or mislabelled data. Deep embeddings obtained with the help of typical CNN backbones are evaluated systematically with the help of a series of machine-learning classifiers. The paper shows that dataset curation combined with representation-oriented learning is used to amplify consistent and discriminative facial features that are robust to dataset change and annotation noise and represents data quality rather than architecture complexity.

2. Related Work

Early FER models were based on geometric features and texture descriptors as Local Binary Patterns and Gabor filters. These methods achieved moderate performance when used under controlled circumstances but remarkably deteriorated when pose changed, illumination varied and when there was occlusion [11], [12]. Their performance was highly sensitive to face alignment and to manually tuned parameters. The introduction of deep CNNs marked an important turn in FER research. CNNs directly learn hierarchical representations from pixel data, hence decreasing the reliance on manual-designed features [4], [13]. VGG style networks extracted effective features through stacked convolutions and achieved strong performance on FER2013 [5], [14], [24]. Residual networks further increased representational depth by solving vanishing-gradient problems with identity mappings [12].

The features of the dataset are big determinants of FER performance. The CK+ dataset is highly resolved and labelled expressions with posed and exaggerated emotions [7]. Models that have been trained only with CK+ are also prone to fail when generalized to unconstrained settings. FER2013 presented in-the-wild facial images in large scale, but its resolution is very low and the annotations are noisy, limiting achievable accuracy, and the human performance on it is reported as 65-68% [8]. On FER2013, single-network CNN models usually reach a plateau of between 70% to 73 % without extraneous information [14], [19]. According to a number of studies, the noise in the dataset has a direct impact on the learned decision boundaries. Noise in the labels decreases the accuracy of the classification by over 5-10 percent without being mitigated [9]. Analysis of FER-specific data proves that mislabelled and ambiguous samples violate the geometry of class separability and embedding [10]. Noise-aware learning and sample filtering approaches improve robustness without full manual relabelling [9], [21], [15].

Multi dataset learning has already been investigated in an attempt to enhance the diversity of expressions and decrease datasets bias. However, naive dataset merging introduces annotation inconsistencies and duplication of samples [11]. Cross-dataset tests show over 15 % decreases in the accuracy in cases when the models are tested outside of their training sample [18]. These results highlight that dataset harmonization and quality control should be followed by representation learning. Recent advancements in FER methods, benchmarks, and datasets further emphasize the importance of standardized evaluation protocols [20]. Using automated data curation methods is a relatively new development. Laplacian variance-based blur detection is highly effective to eliminate low-quality facial images and enhance the training stability of CNN [32]. The classifier-confidence-based filtering uses the concept of detecting the repeatedly low-confidence samples across the models [21]. Such strategies have been shown to improve downstream accuracy by 2–5% in noisy datasets [9], [10]. The choice of model architecture has also an impact on the fidelity of representation. ResNet18 balance depth and computational efficiency without affecting fine grained features of the face [12]. MobileNetV2 reduce the model size and inference cost but can lose discriminative power especially to subtle expressions [37]. According to comparative studies, both ResNet18 and MobileNetV2 models show 2 to 4% accuracy differences in the same training conditions [28].

Hybrid FER pipelines combining deep embeddings with classical machine-learning classifiers have demonstrated competitive performance. Gradient-boosting classifiers,Random Forests, and Support Vector Machines usually perform better in comparison with softmax classifiers when trained on fixed deep features [13], [23]. Ensemble approaches, specifically, Random Forests, are resistant to residual noise and nonlinear decision boundaries in moderate-sized data [18], [33]. Reported gains of 1–3% over end-to-end CNN classifiers are common in hybrid settings [23], [34]. Optimization strategy further influences representation quality and convergence behavior. The RMSProp and Adam adaptive optimizers are faster but can have unstable generalization [26]. Weight-decay techniques that are decoupled like AdamW are better at regularizing decoupled models and also have more predictable performance during validation [6], [27]. According to the reports of FER studies, the use of AdamW increases the accuracy by an average of 1-2% compared to the use of standard Adam [6]. Recent efforts have extended FER research to be able to deal with not only static but also dynamic emotion representations. This is a systematic review of categorizing methods by image-based and sequence-based and identifying the difficulties related to the ambiguity of expressions, intensity variation, cross-dataset inconsistency, and uncertainty with time [25]. In order to solve the noise in annotation, ReSup concurrently learns clean and noisy labels with agreement between two networks, which demonstrate quantifiable accuracy gains on FERPlus [26].

FER has also been pointed out to be strong when there is representation diversity. Global and local feature cues learned with diversified feature learning models lead to high accuracy on the RAF-DB, FER+, and AffectNet and prove the relevance of deep embeddings in unconstrained conditions [27]. There is also Ensemble learning that enhances strength. Multi-architectural ensemble models using distillation are competitive with a lower cost to compute [28]. Construction of datasets is still a critical issue with FER evaluation. RAF-DB is a set of large-scale annotations of both basic and compound emotions in

varied demographic situations [29]. Refinement work has been demonstrated to improve the performance on benchmark tasks such as relabelling and category expansion of FER datasets [30]. There have also been architectural improvements done based on attention. The use of convolutional block attention modules enhances discriminative learning of features, and accuracy is reported to be stronger on common benchmarks [31]. Comparative studies on real-time FER highlight trade-offs between accuracy and computational efficiency, favouring lightweight models for low-latency deployment [32].

AffectNet-HQ, a large-scale dataset, make it possible to do additional benchmarking, as deep CNN models have reached over 85 % recognition accuracy [33]. The recognition of expression dynamics is enhanced with the help of temporary modeling methods that are based on integrating CNNs with RNN-LSTM networks, which emphasizes the importance of a temporal context [34]. Multi-loss and ensemble strategies are also used to make FER more robust. The feature fusion models and voting-based ensembles outperform single-model baselines on both RAF-DB and unbalanced versions of AffectNet [35]. The label distribution learning minimizes annotation ambivalency and enhances the performance in multiple datasets [36]. The self-supervised learning systems separate out identity and expression properties and achieve the same accuracy gains on in-the-wild datasets [37]. Graph attention networks encode both spatial-temporal connections and generate competition among FER datasets [38].

The lightweight attention architectures deal with the computational limitations, but are robust to pose variation [39]. The simplified cross-attention networks are state of the art and less complex to implement [40]. Hybrid attention processes also enhance their feature discriminability through a combination of both local and global attention cues[41], [42]. Comparative evaluations on benchmark datasets remain essential. CNN-based FER studies confirm strong performance on FER2013 and CK+ under real-world variability, providing practical baselines for research and deployment [43], [44].

3. Experiment

A. Dataset and Preprocessing

Experiments are conducted on a filtered multi-dataset corpus of facial expressions, created by combining CK+, FER2013, and other publicly available datasets containing labelled pictures of human face emotions. A total of 28,272 images are filtered to 16,828 images following an automated quality filtering process that eliminates blurred, low-resolution, and ambiguous images. Six emotion categories are used in this study: anger, disgust, fear, happiness, sadness, and surprise. Neutral expressions are not included. All images are resized to 224×224 in RGB format and are normalized using ImageNet statistics. During training, horizontal flipping, ±15° rotation, and brightness variation are applied as data augmentation techniques. The dataset is divided into 80% training data and 20% test data. To evaluate stability in performance, four-fold cross-validation (k = 4) is used with a constant random seed.

B. Feature Extraction and Classification

The frozen feature extractors, i.e., ResNet18 and MobileNetV2, will be utilized in the present paper with pre-trained ImageNet weights, and all the layers are maintained at the frozen level to avoid any bias that can be brought by fine-tuning. This guarantees a healthy comparison of intrinsic representational abilities of the two architectures in the same conditions of the experiment. As a result, the deep features obtained are based on the learned general-purpose visual representations and not adaptation to the task. The architectures of the two backbones are depicted in Figure 1 and Figure 2 to gain a better insight into the structural differences between them. In Figure 1, the architecture ResNet18 is shown and is distinguished by residual connections that make information bypass the intermediate layers. These skip connections are useful in gradient flow preservation during training as well as allowing the network to learn more discriminative and deeper feature representations. Consequently, ResNet18 has been specifically useful in the acquisition of minute differences in facial expressions, which is essential in the differentiation of the near related emotional states.

Figure 2, in turn, shows the structure of MobileNetV2 architecture, which is created with the emphasis on computational efficiency. It applies depthwise separable convolution and inverted residual blocks to greatly cut the quantity of parameters and computation expenses. Although such a design allows MobileNetV2 to be applicable to real-time and resource-constrained scenarios, it can limit its capability to represent fine-grained facial features in comparison to more profound residual networks. This comparison of these two architectures brings out the trade of between representational power and computational efficiency which is also seen in the experimental results.

Feature embeddings are extracted from the global pooling layer of each network. Classification is performed using Support Vector Machine, Random Forest, XGBoost, K-Nearest Neighbors, Gaussian Naive Bayes, and Decision Tree models trained on the extracted features. Among all classifiers, Random Forest achieves the best performance, with the highest test accuracy of 70.34% using ResNet18 embeddings, and shows consistent outperformance compared to MobileNetV2 across classifiers.

C. Optimization and Training Setup

Optimizers are evaluated using SGD, SGD with momentum, Nesterov, Adagrad, RMSProp, Adadelta, Adam, Nadam, AMSGrad, and AdamW under identical conditions. Training is conducted using weighted cross-entropy loss with ReduceLROnPlateau scheduling. Based on validation performance and convergence behavior, AdamW is used for final reporting. Although CNN backbones are frozen, optimizers are evaluated for training the classifier head and loss convergence stability under identical conditions. All experiments are conducted in TensorFlow under a CPU-only, deterministic execution setup.

Results

Table 1 shows the benchmark comparison of various models which were considered in this work. As one may notice, the offered hybrid scheme with the ResNet18 embeddings and the use of a Random Forest as a classifier have the highest accuracy of 70.34% and, therefore, outperform both the classical and

deep learning baselines. Although some of the deep CNN based models like VGG, ResNet50 and Inception V3 are also able to perform competitively, their performance is a bit lower when operated under the same circumstances. This underscores the efficiency of representation based learning in combination with sound classical classifiers. Fig. 3 shows that the trend in accuracy is different in various models. It can be noted that there is a definite performance disparity between lightweight architectures and more profound residual networks. Specifically, ResNet18 is always capable of generating more discriminative embeddings than MobileNetV2, and this results in a better classification score among all the considered classifiers. Fig. 5 presents both backbone models confusion matrices. As it can be noted, ResNet18 is more successful in class separation with less misclassifications between the visually similar emotions like fear and surprise.

Table 1. Model benchmark

Model Name	Accuracy (%)
ARM (ResNet)	69.5
VGG + SMI	69.1
Attentional ConvNet	68.8
CNN + SVM	68.2
MobileNetV2 + RF	67.49
VGG16	67.05
GoogLeNet	66.84
ResNet50	66.37
Inception V3	66.1
Bag of Words	63.5
Custom CNN	55.44
ResNet18 + RF	**70.34**

MobileNetV2, on the contrary, has a comparatively greater level of confusion between these classes, which suggests a weakness in the recognition of minor facial changes. The consistency of the suggested solution is also supported by cross-validation test.

The performance of all the folds has the same variance as illustrated in Fig. 4 which confirms the strength of the learned representations with varying data splits. This shows that the model is also generalized and is not highly reliant on a particular training subset.

Table 2 shows the validity acquired through various optimization strategies. Of all optimizers considered, AdamW has been found to be the most accurate (70.94) and the most stable. Although adaptive

optimizers like Adam and RMSProp have a lower convergence rate, they have a little more variance in validation performance.

Table 2. Validation accuracy of optimizers

Optimizer	Best Validation Accuracy (%)
SGD	68.86
SGD + Momentum	69.18
NAG	68.62
Adagrad	68.32
RMSprop	69.96
Adadelta	69.18
Adam	69.87
AdamW	**70.94**
Nadam	70.61
AMSGrad	70.73

Conversely, the convergence speed and total accuracy of SGD-based methods are slower. All in all, the findings prove that the quality of feature representation and dataset curation are also more important in FER performance compared to architectural complexity itself. Curation multi-dataset learning with learned high-quality embeddings and classical classifiers offers a dependable and stable system of facial expression recognition in noisy and heterogeneous settings.

Optimizer Comparison

Optimizer performance is evaluated under identical training conditions using weighted cross-entropy loss and ReduceLROnPlateau scheduling. Among the tested optimizers (SGD, SGD with momentum, Nesterov, Adagrad, RMSProp, Adadelta, Adam, Nadam, AMSGrad, and AdamW), AdamW consistently achieves the highest test accuracy and the most stable convergence behavior across folds. SGD-based optimizers show slower convergence and higher variance. Adam and RMSProp converge faster than SGD but exhibit less stable validation performance compared to AdamW. Nadam and AMSGrad provide no consistent improvement over Adam. Based on accuracy stability and convergence consistency, AdamW is selected for all final experiments.

Backbone Comparison

Backbone evaluation is performed using frozen feature extractors under identical data splits and classifiers. ResNet18 consistently outperforms MobileNetV2 across all evaluated classifiers and cross-

validation folds. The highest overall accuracy of 70.34% is obtained using ResNet18 embeddings with a Random Forest classifier, while MobileNetV2 achieves lower peak accuracy under the same settings. ResNet18 embeddings exhibit improved class separability and reduced confusion between closely related emotion classes. Results indicate that deeper residual representations provide higher discriminative capacity than lightweight architectures for multi-dataset FER under frozen-feature constraints.

Conclusion

This work presents a high-fidelity facial expression representation framework evaluated under controlled multi-dataset conditions. Automated data curation reduces noise and improves representation reliability. Frozen CNN feature extraction enables fair backbone comparison without fine-tuning bias. ResNet18 produces more discriminative embeddings than MobileNetV2 across classifiers. Random Forest gives the best classification results of all considered models. AdamW has the most consistent and stable behaviour of optimisation. The findings indicate that the quality of representation and integrity of data are the major factors to drive FER performance in limited training environments.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

The author confirms that the image presented in Figure 2 is of the author themself and provides consent for its publication.

Availability of data and materials

The datasets analyzed in this study are publicly available facial expression recognition datasets including CK+ and FER2013. These datasets are accessible from their respective official repositories subject to their licensing terms. The curated dataset generated through automated quality filtering is derived from these publicly available datasets.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors received no specific funding for this work.

Authors' contributions

H.V. conceptualized the study, designed the methodology, conducted the experiments, analyzed the results, and wrote the main manuscript text.

P.V. supervised the research work and contributed to conceptual guidance and manuscript review.

N.K. co-supervised the research and contributed to methodological refinement and technical validation.

Acknowledgements

The authors acknowledge the use of publicly available facial expression datasets used in this research.

References

1. Bartlett MS, Littlewort G, Fasel I, Movellan JR. Real-Time Face Detection and Facial Expression Recognition: Development and Applications to Human–Computer Interaction, *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 53–60, 2003.
2. Abdat F, Maaoui C, Pruski A. Human–computer interaction using emotion recognition from facial expression, *Proc. UKSim 5th European Modelling Symposium on Computer Modelling and Simulation*, pp. 196–201, 2011.
3. Fasel B, Luettin J. Automatic facial expression analysis: A survey. Pattern Recogn. 2003;36(1):259–75.
4. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90.
5. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition, *Proc. International Conference on Learning Representations (ICLR)*, 2015.
6. Li S, Deng W. Deep facial expression recognition: A survey. IEEE Trans Affect Comput. 2020;11(1):3–18.
7. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression, *Proc. IEEE CVPR Workshops*, pp. 94–101, 2010.
8. Goodfellow IJ, et al. Challenges in representation learning: A report on three machine learning contests. Neural Netw. 2015;64:59–63.
9. Frenay F, Verleysen M. Classification in the presence of label noise: A survey. Pattern Recogn. 2014;45(9):1–28.
10. Patrini G, Rozza A, Menon AK, Nock R, Qu L. Making deep neural networks robust to label noise: A loss correction approach, *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1944–1952, 2017.

11. Cohn JF, De la Torre F. Automated face analysis for affective computing. in The Oxford Handbook of Affective Computing. Oxford University Press; 2014. pp. 131–50.
12. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition, *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
13. Choi M, Lee J, Kim S. Facial expression recognition using deep features and SVM classifiers. Sensors, 21, 9, Article 3179, 2021.
14. Hasani B, Mahoor MH. Facial expression recognition using enhanced deep 3D convolutional neural networks, *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30–40, 2017.
15. Wang K, Peng X, Yang J, Lu S, Qiao Y. Suppressing uncertainties for large-scale facial expression recognition, *arXiv preprint arXiv*:2007.*03149*, 2020.
16. Oguine OC, Oguine KJ, Bisallah HI, Ofuani D. Hybrid facial expression recognition (FER2013) model for real-time emotion classification, *arXiv preprint arXiv:2203.08901*, 2022.
17. Sajjad M, et al. A comprehensive survey on deep facial expression recognition: Challenges, applications and future guidelines. IEEE Access. 2023;11:11245–70.
18. Li S, Deng W, Du J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. Neural Netw. 2016;73:70–81.
19. Reghunathan RK. Facial expression recognition using pre-trained ensemble models for FER2013. Sensors, 24, 3, Article 812, 2024.
20. Kopalidis T. Advances in facial expression recognition: Methods, benchmarks, models, and datasets, *Information*, vol. 15, no. 2, Article 98, 2024.
21. Liang C, Dong J. A survey of deep learning-based facial expression recognition research. Front Comput Intell Syst. 2025;4(1):1–20.
22. Agung ES et al. Image-based facial emotion recognition using deep learning with extended emotion categories. Sci Rep, 14, Article 6123, 2024.
23. Ullah S. Facial expression recognition survey: Edge computing deep learning approaches, *PeerJ Computer Science*, vol. 10, e1913, 2024.
24. Hasani B, Mahoor MH. Facial expression recognition using enhanced deep 3D convolutional neural networks. IEEE Trans Affect Comput. 2020;11(4):653–66.
25. Zhang Z, Zhao X, Li Y. A survey on facial expression recognition of static and dynamic emotions. IEEE Access. 2023;11:32415–38.
26. Wang K, Peng X, Yang J, Lu S, Qiao Y. ReSup: Reliable label noise suppression for facial expression recognition, *arXiv preprint arXiv:2305.17895*, 2023.
27. Li J, Zhang Y, Liu Z. Learning diversified feature representations for facial expression recognition in the wild, *arXiv preprint arXiv:2210.09381*, 2022.
28. Kollias D, Zafeiriou S, Tefas A. Recognizing facial expressions in the wild using multi-architectural ensembles with distillation, *arXiv preprint arXiv:2106.16126*, 2021.

29. Li S, Deng W, Du J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. Neural Netw. 2016;73:70–81.
30. Ben A, Ammar, et al. Introducing a novel dataset for facial emotion recognition. Heliyon. 2024;10(2):e21345.
31. Khan MS, Hussain J, Aboalsamh H, Bebis G. Attention-based facial expression recognition: A survey. IEEE Access. 2021;9:165002–28.
32. Chen Y, et al. Real-time facial expression recognition: Advances and comparative analysis. Int J Comput Vision. 2023;131(4):987–1006.
33. Li H et al. Deep learning-based facial emotion recognition on AffectNet-HQ. Int J Pattern recognit Artif Intell, 38, 5, 2024.
34. Sharma R, Singh P. Integrating CNN and RNN-LSTM for facial expression recognition, *Proc. International Conference on Computer Vision Theory and Applications*, pp. 214–221, 2024.
35. Zhou G, Xie Y, Fu Y, Wang Z. Multi-loss based feature fusion and top-two voting ensemble decision strategy for facial expression recognition in the wild, *arXiv preprint arXiv:2311.03478*, 2023.
36. Liu S, Xu Y, Wan T, Kui X. Ada-DF: An adaptive label distribution fusion network for facial expression recognition. Inf Sci. 2024;677:1–14.
37. He R, Xing Z, Tan W, Yan B. A generative framework for self-supervised facial representation learning, *arXiv preprint arXiv:2309.08273*, 2023.
38. Li Y, Zhao Y, Xia X, Jiang D. ARPGNet: Appearance- and relation-aware parallel graph attention fusion network for facial expression recognition, *arXiv preprint arXiv:2511.22188*, 2025.
39. Ezati A, Dezyani M, Rana R, Rajabi R, Ayatollahi A. A lightweight attention-based deep network via multi-scale feature fusion for multi-view facial expression recognition, *arXiv preprint arXiv:2403.14318*, 2024.
40. Mao J, Xu R, Yin X, Chang Y, Nie B, Huang A. POSTER++: A simpler and stronger facial expression recognition network, *arXiv preprint arXiv:2301.12149*, 2023.
41. Khan MS, Hussain J, Aboalsamh H, Bebis G. Attention-based facial expression recognition: A survey. IEEE Access. 2021;9:165002–28.
42. Song D, Liu C. A facial expression recognition network using hybrid feature extraction. PLoS ONE, 19, 1, e0280213, 2024.
43. Gailan MJ. Performance evaluation of facial emotion recognition through convolutional neural network model. Int J Eng Res Technol (IJERT). 2025;14(3):1–6.
44. Gan Y. Facial expression recognition using convolutional neural networks, *Proc. 2nd International Conference on Intelligent Computing and Applications*, pp. 1–6, 2018.

Figures

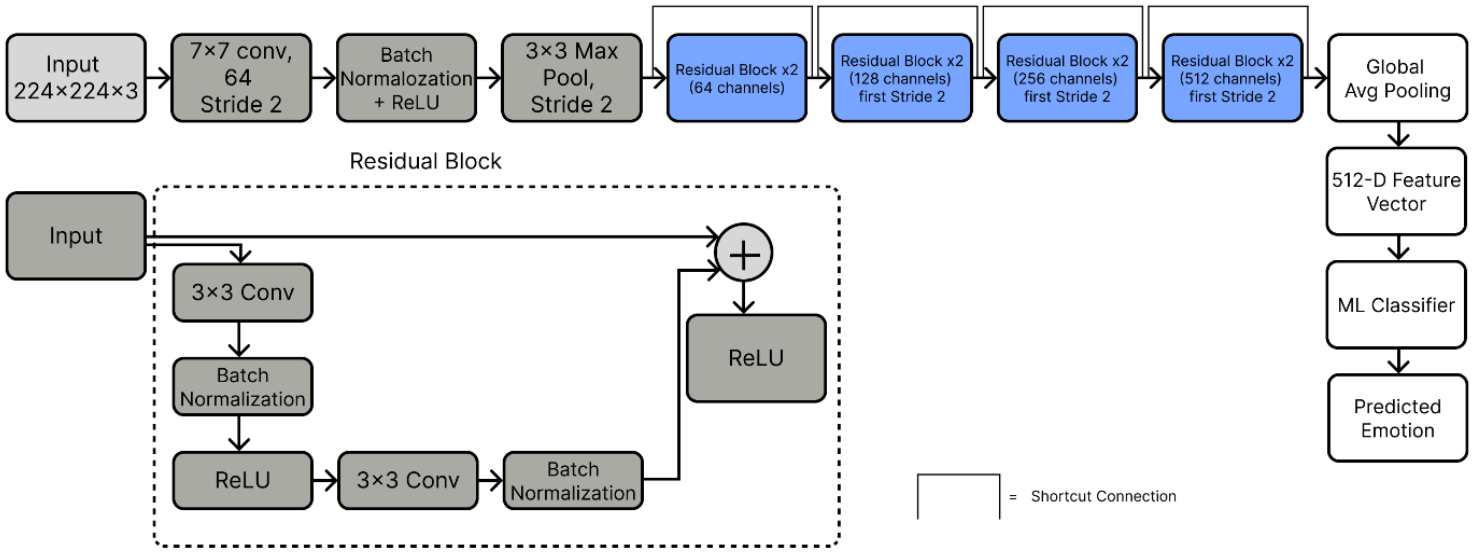


Figure 1

Architecture of the ResNet-based convolutional backbone used for feature extraction, illustrating the residual block structure and hierarchical feature learning.

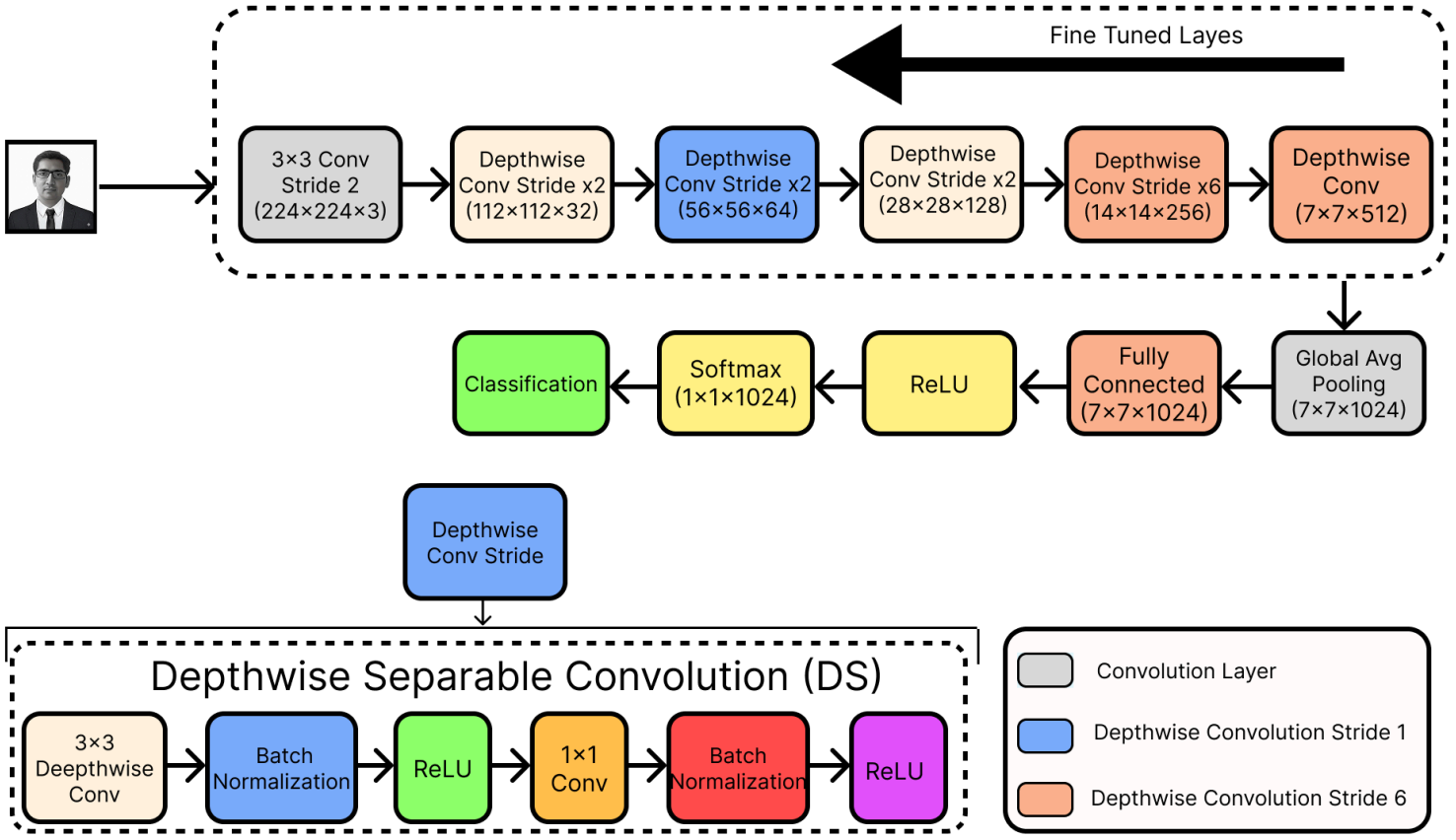


Figure 2

Architecture of the MobileNet-based convolutional backbone employing depthwise separable convolutions for lightweight feature extraction.

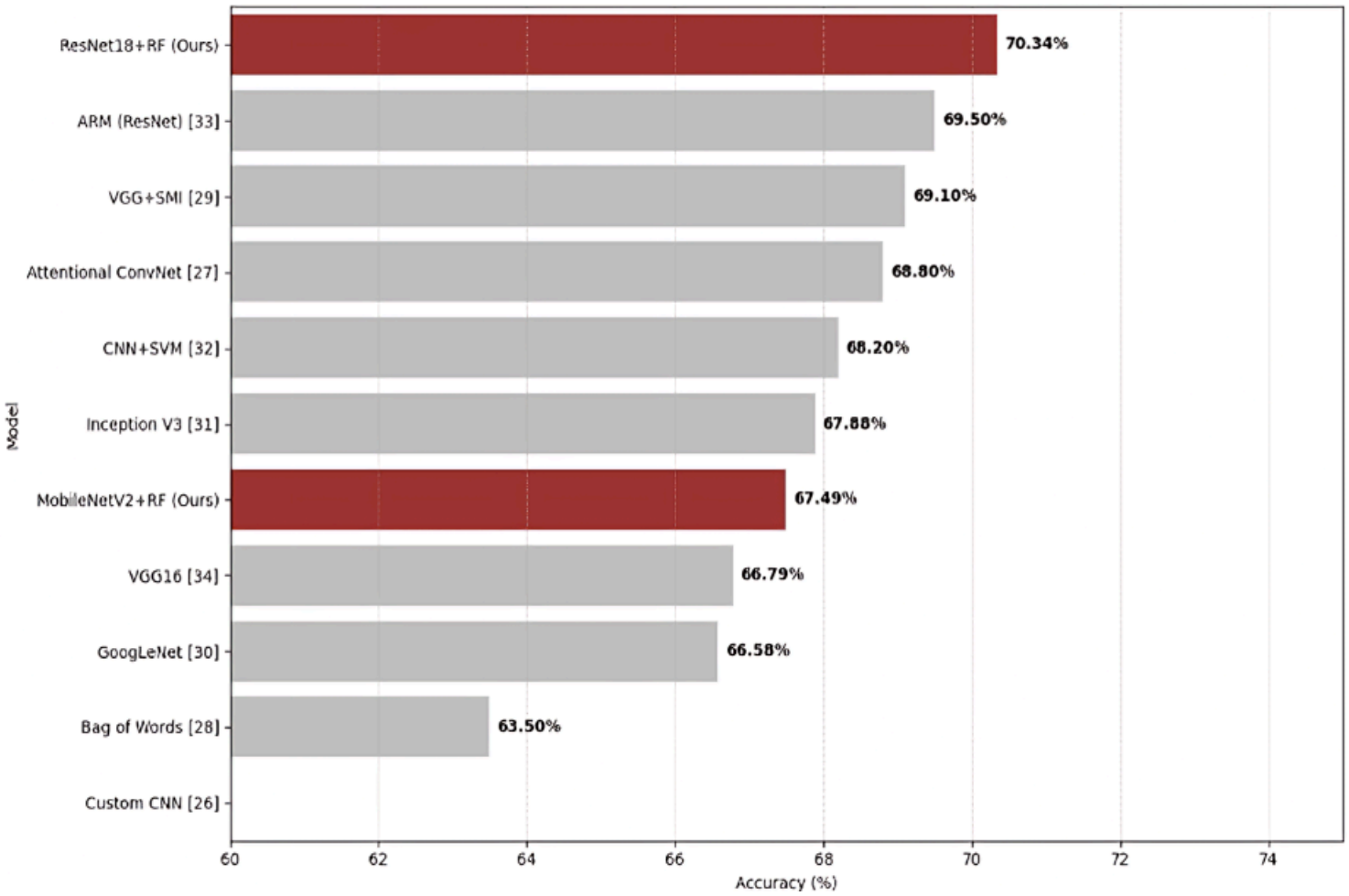


Figure 3

Accuracy comparison of models

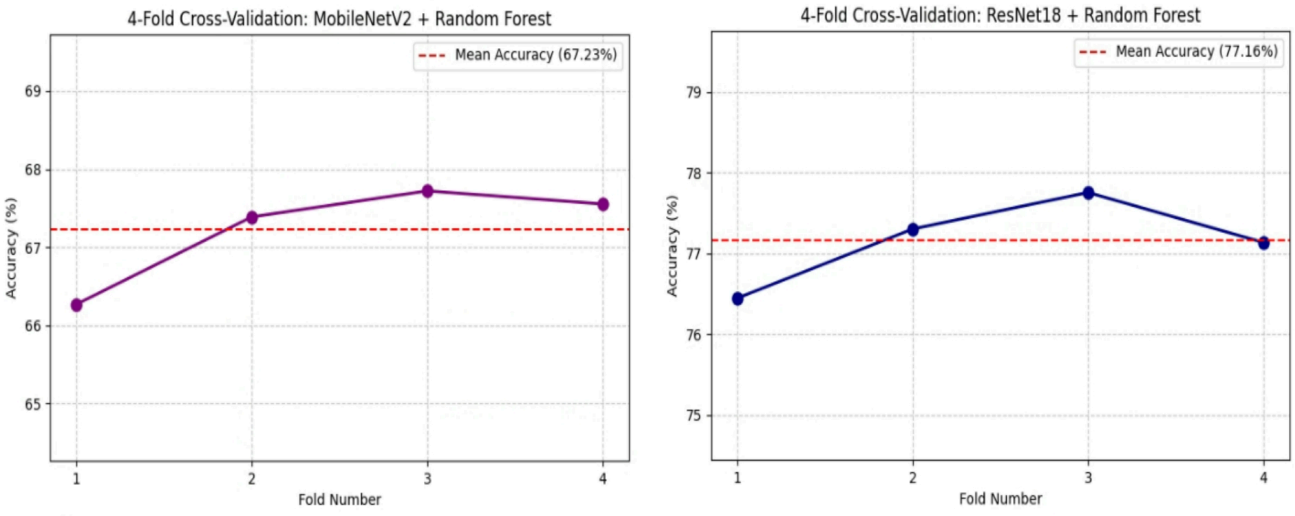


Figure 4

Cross Validation (k = 4) of MobileNetV2+RF and ResNet18+RF

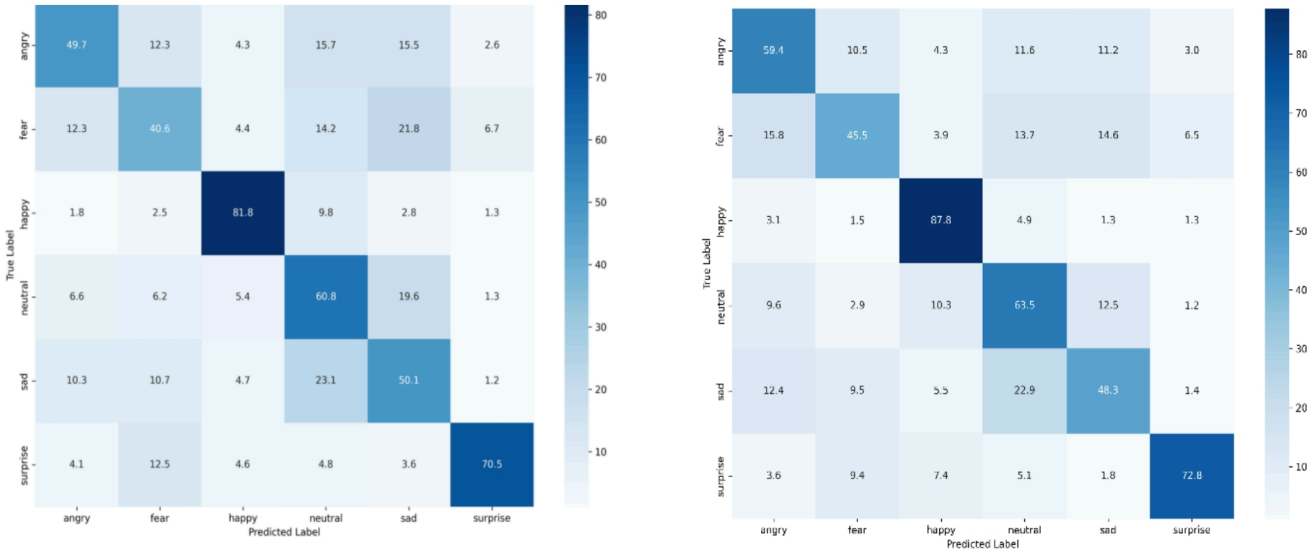


Figure 5

Confusion matrix of mobile MobileNetV2 (Left) and ResNet18 (right)