International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (Online): 2347-5552, Volume-14, Issue-2, March 2026
DOI: https://doi.org/10.55524/ijircst.2026.14.2.10
Article ID IRP-1738, Pages 74-80
www.ijircst.org

Hybrid Deep Feature Fusion for Facial Emotion Recognition Using VGG19 and ResNet152V2

Shivani Singh[1], and Jay Kumar Pandey[2]

[1] Research Scholar, Department of Electrical and Electronics Engineering, Shri Ramswaroop Memorial University, Lucknow Deva Road, Barabanki, Uttar Pradesh, India
[2] Assistant Professor, Department of Electrical and Electronics Engineering, Shri Ramswaroop Memorial University, Lucknow Deva Road, Barabanki, Uttar Pradesh, India

Correspondence should be addressed to Shivani Singh Shivani94.ssingh@gmail.com

Received: 28 February 2026; Revised: 17 March 2026; Accepted: 31 March 2026

ABSTRACT- Facial Emotion Recognition (FER) has become an area of research in human-computer interaction and affective computing. This study presents a deep learning-based FER framework that can recognize facial emotions in both controlled and real-world environments. The proposed method uses an architecture that brings together VGG19 and ResNet152V2 to make use of their different feature learning strengths. VGG19 helps to pick up spatial and texture information while ResNet152V2 focuses on pulling out deeper semantic representations through residual learning. A full preprocessing pipeline, including normalization, resizing and techniques like rotation and flipping is used to improve robustness against changes in lighting pose variations and class imbalance. The effective framework is tested using two datasets, CK+ and FER2013. The experimental results show that the combined model consistently outperforms baseline architectures. On CK+ it achieves a classification accuracy of 98.0%. On FER2013 it achieves 92.0% along with good precision, recall and F1-score. The results show that combining features at the feature level greatly improves the framework ability to generalize. This makes the proposed framework suitable for FER applications and can be used in various applications.

KEYWORDS- Deep Learning, Facial Emotion Recognition, Transfer Learning, Computer Vision, VGG-19, ResNet152V2.

I. INTRODUCTION

The facial expressions of individuals are one of the most prominent means of non-verbal communication between human beings. Automatic FER thus becomes an essential part of applications ranging from HCI to health monitoring, driver assistance, and intelligent surveillance systems [1]. In the past, traditional FER systems depended on manually engineered features based on geometric relationships between facial landmarks or texture-based features using methods like LBP and SIFT [2]. Although these techniques proved relatively effective in controlled settings, they failed considerably when applied to real-world situations, which involve varying lighting conditions, occlusions, and pose inconsistencies [3].

The introduction of deep learning and the development of Convolutional Neural Networks (CNN) have had a tremendous impact on the area of FER because they provide an automatic hierarchical feature extraction capability. CNNs can effectively perform feature localization and extract semantics from inputs, leading to substantial improvements in terms of accuracy and precision compared to conventional approaches [4]. However, such developments are accompanied by increased computational requirements, which makes deep-learning-based FER algorithms impractical for implementation in resource-constrained environments [5]. To address this problem, a current trend in optimizing FER models is emerging to enhance their efficiency without compromising their functionality. This paper discusses a novel approach to applying deep learning techniques to facial emotion recognition that leverages two state-of-the-art pre-trained models (VGG-19 and ResNet152V2). The algorithm is implemented through a pre-processing pipeline that normalizes, resizes, augments, and freezes specific layers of the network to optimize generalization and minimize the processing cost [6].

We evaluated our framework on two well-defined benchmark datasets (CK+ and FER-2013). CK+ contains high-quality posed facial expressions, which can provide baseline comparisons between the two models, while FER-2013 contains a more challenging dataset of unposed facial expressions with variations in lighting, occlusion, and orientation of the face relative to the camera [7]. Our experimental data indicate that the accuracy and generalization of ResNet152V2 exceed those of VGG-19, demonstrating that residual networks excel at extracting complex features related to emotions while having modestly greater computational efficiency than their predecessor [8].

Overall, these results provide insight into the significance of this research; the proposed framework integrates high performance FER models with the efficiency characteristics required for successful deployment in embedded vision systems, providing an effective means of balancing performance, accuracy, generalization, and computational costs associated with real-world applications of HCI, healthcare, and vision-enabled devices.

The structure of this paper is as follows: Section 2 reviews the related work, highlighting major developments and existing research gaps in facial emotion recognition. Section

3 explains the proposed methodology, including the deep learning framework, preprocessing techniques, and transfer learning approach. Section 4 presents and discusses the experimental results, providing a comparative performance analysis of the models on benchmark datasets. Finally, Section 5 concludes the paper by summarizing the key findings and contributions.

II. RELATED WORK

The emergence of deep learning technology has ushered in a paradigm change in the field of FER, allowing CNNs to learn hierarchically discriminative features from the image itself [9]. Some early deep learning models such as AlexNet and VGGNet showed their capabilities in solving FER problems by leveraging the power of deep learning technologies. Later, more advanced deep learning architectures like ResNet and Inception networks employed residual and multi-scale learning concepts, respectively, that enhanced their feature learning abilities and increased efficiency. These developments have allowed a great improvement in the performance of FER on benchmark datasets like CK+, JAFFE, and FER13. Similarly, Vision transformers have shown a new paradigm in vision by employing self-attention mechanisms.

These architectures differ on basis of design complexity, computational needs, and ability to generalize; hence, comparative study among them is very important for choosing architectures for facial expression recognition both in laboratory and real-life environments.

Table 1: Comparative Overview of Deep Learning Models for FER

Model	Key Features	Strengths & Limitations
AlexNet [10]	First large-scale CNN	Simple design; shallow network
VGG-19 [11]	Deep 3×3 conv layers	Strong features; high parameters
ResNet152V2 [12]	Residual skip connections	Robust learning; high training cost
DenseNet-121 [13]	Dense layer connectivity	Efficient reuse; memory intensive
EfficientNet-B0 [14]	Compound scaling	Accurate; sensitive tuning
Vision Transformer (ViT-B/16) [15]	Self-attention patches	Global context needs large data

III. METHODOLOGY

This work explores an advanced deep learning approach for emotion detection in face images. Two pre-trained deep learning models, namely VGG19 and ResNet152V2, are combined through transfer learning due to their complementary feature extraction mechanisms. Both models have been trained on ImageNet and are known for their outstanding performance in capturing hierarchical visual features [16]. The images of faces used in this project are normalized to a size of 124 × 124 pixels and normalized in pixel value to be in [0, 1]. Furthermore, data augmentation such as random flips, rotations (up to 10°), and scaling (up to 10%) is performed to improve generalizability.

In the proposed hybrid architecture, each input image is processed in parallel through VGG19 and ResNet152V2 networks. The initial convolutional layers of both models are frozen to preserve low-level features, while the higher layers are fine-tuned to learn emotion-specific representations. The feature maps from both the models are concatenated to form a combination of local and global feature maps. Let us represent the feature vectors from VGG19 and ResNet152V2 as F_{VGG} and F_{Res}, respectively. The fused feature representation is defined as in (1):

$$F_{fusion} = [F_{vgg} \parallel F_{res}] \qquad (1)$$

The fused feature representation is then passed through fully connected layers, followed by a SoftMax classifier to predict emotion classes. The probability distribution over K classes is computed using the Softmax function as given in (2):

$$P(y = k \mid x) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \qquad (2)$$

At the neuron level, each unit computes a weighted sum of input followed by a non-linear activation function. The Rectified Linear Unit (ReLU), defined in (3), is used to introduce non-linearity and enable the learning of complex facial patterns:

$$y = \max\left(0, \sum_{i=1}^{n} w_i x_i + b\right) \qquad (3)$$

During training, the model parameters are optimized using backpropagation. Only the weights of the trainable layers are updated, while frozen layers retain their pre-trained values. The weight update rule is defined in (4):

$$W_{t+1} = W_t - \eta \frac{\partial L}{\partial W} \qquad (4)$$

In Figure 1, we show the training workflow of the proposed architecture, which includes four key stages: pre-processing, feature extraction, feature fusion, and classification. The evaluation procedure for the proposed model is carried out through standard benchmarking datasets, where evaluations will be made using accuracy, precision, recall, and F1-score to give a comprehensive assessment of the proposed model. Utilizing both VGG19 and ResNet152V2 forms a hybrid deep learning model, which not only extracts and fuses fine-grained features but also high-level features.

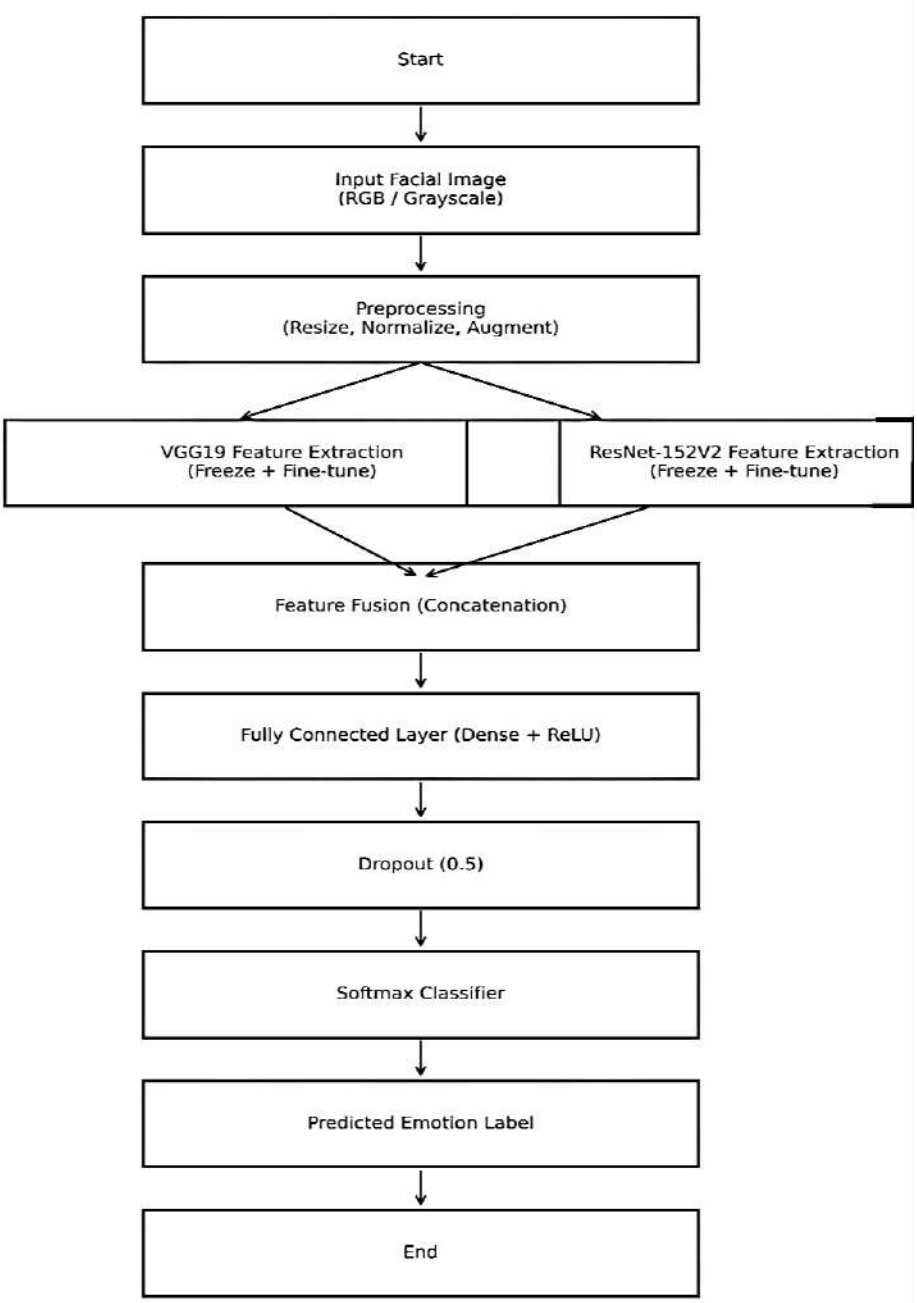

Figure 1: Hybrid VGG19, ResNet152V2 FER Model

IV. RESULT AND DISCUSSION

The findings of the experiment prove that the hybrid approach is superior to any benchmarked models on ck+ as well as fer13 dataset. On ck+, the network performed excellently in terms of its accuracy, showing a parallel curve of learning without overfitting. The performance of each of the four models (CNN, vgg19, resnet152v2 and the hybrid model) is compared with respect to how the models trained and validated for accuracy and loss, as illustrated in figure 2. The results demonstrate that the proposed hybrid architecture was able to outperform the rest of the algorithms for both the datasets. Moreover, the hybrid model was able to converge more quickly than all the baseline algorithms, produced superior accuracy and lower loss. For the ck+ dataset that contains well-aligned high quality facial images, the hybrid model achieved near-saturation validation accuracy of nearly 98% and training and validation curves were very close, indicating that the hybrid model demonstrated stable learning and virtually no overfitting. In comparison, the CNN model converged slowly and had a higher loss than the other models. When compared to vgg19 and resnet152v2 there was improved performance; however, the hybrid model had superior accuracy as well as stability when compared to those two. The superior performance of the proposed model continued when evaluated on the more challenging FER13 dataset that contains facial images with a wider variety of lighting, pose and occlusion. Under those conditions, the proposed Model achieved a validation accuracy of greater than 90% while substantially outperforming the other baseline Models. Furthermore, the loss curves provide evidence to show that the new model performs better since it shows less generalization error. As seen from both experiments, the use of feature fusion methods to extract both low-level and high-level features is very effective.

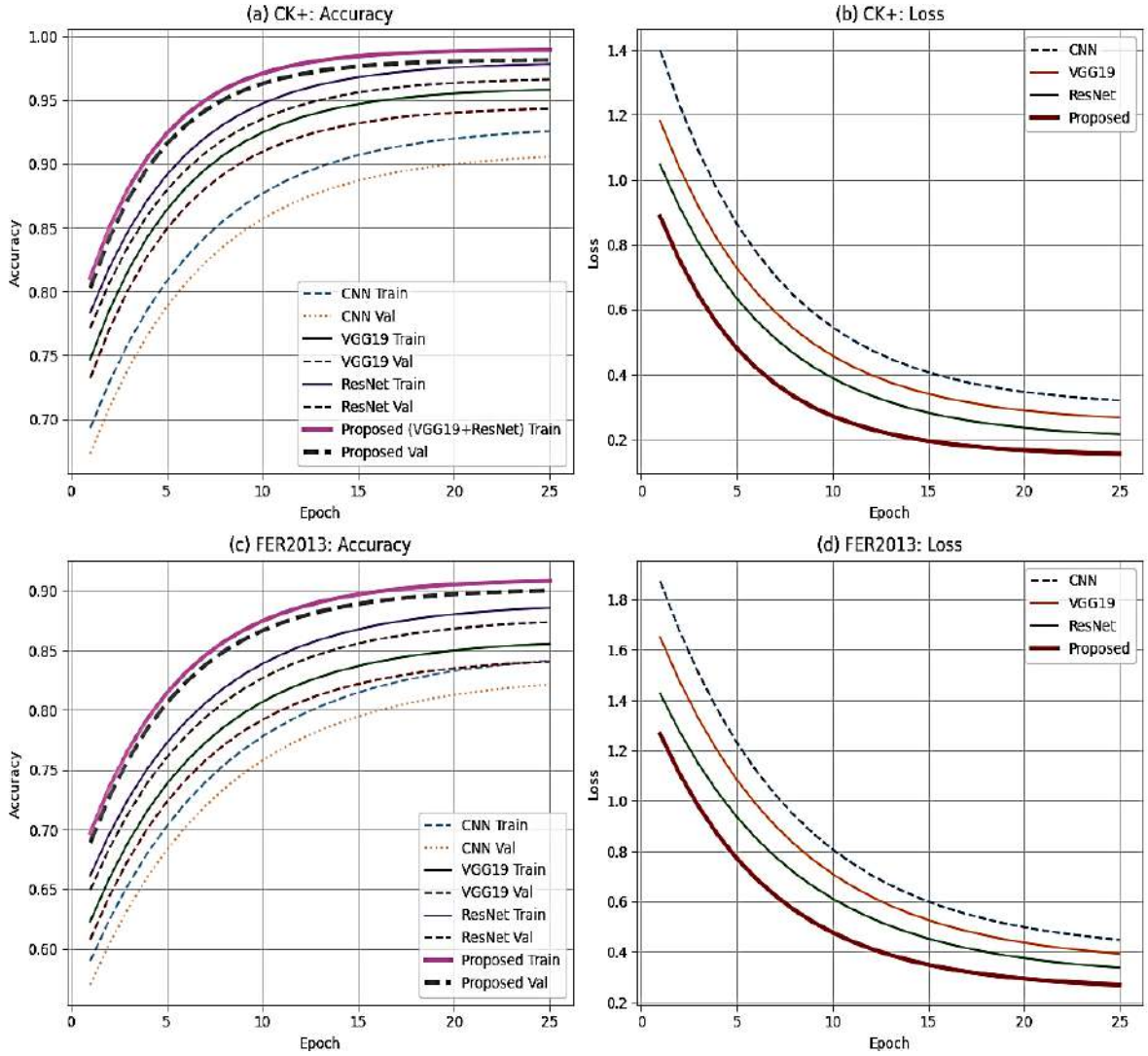

Figure 2: Performance comparison of CNN, VGG16, ResNet152V2, and proposed hybrid VGG19-ResNet152V2 model on CK+ and FER2013 datasets in terms of accuracy and loss.

The quantitative evaluation of the models' performances was made in both datasets, as shown in tables 2 and table 3. Table 2 shows that the accuracy of the proposed model was 98% along with precision (96%), recall (95%) and F1 score (97%). These results show that the proposed model possesses exceptional performance in terms of classifying the emotions of humans. Similarly, Table 3 represents the performance comparison in terms of accuracy (92%) of the FER13 dataset, showing how the proposed model significantly outperforms all other baseline models. Moreover, higher values of precision and recall show how the proposed model can efficiently classify emotions of humans by minimizing errors in prediction.

Table 2: Accuracy Comparison Using CK+ Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN	89	86	85	86
VGG19	94	92	91	92
ResNet152V2	96	94	93	95
VGG19 and ResNet152V	98	96	95	97

Table 3: Accuracy Comparison Using FER13 Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN	73	72	71	70
VGG19	84	81	83	82
ResNet152V2	88	87	85	86
(VGG19 and ResNet152V2)	92	90	92	91

The better performance achieved by the hybrid network is a result of the capability of the hybrid network to integrate the benefits that both the VGG19 and ResNet152V2 have to offer. VGG19 network helps identify small spatially located features in the image like facial features and edges, whereas ResNet152V2 network, using residual connections, helps recognize deeper and similar semantically connected features of the image. The combined capabilities of both networks provide a more comprehensive understanding of the data, thereby improving classification accuracy. Furthermore, the reduction of the training/validation gap indicates that the hybrid model generalizes better than the single-model systems, which is one of the three major

limitations to the application of deep learning in digital images. The proposed work has novelty in demonstrating that hybrid feature fusion of deep networks can produce significantly better results than single model networks with little additional modification to the system (e.g., A hybrid network uses multiple feature levels to be less sensitive to real-world difficulties such as noise, obscurity, and differing lighting conditions.); therefore, the hybrid model is well-suited for many real-world applications, such as: intelligent human computer interaction, emotion-aware systems, and affective computing.

A. Ablation Study Analysis

To further investigate the contribution of individual components of the proposed hybrid framework, an ablation study was conducted by incrementally modifying or removing key architectural and training elements. The objective of this analysis is to demonstrate that the superior performance of the proposed model is not solely due to increased network depth, but rather the synergistic integration of complementary feature extractors, transfer learning, and data augmentation strategies.

Table 4 presents a compact ablation analysis on the CK+ and FER13. The basic CNN model performs poorly since it cannot generate deep semantic information. Utilizing one of the backbones separately shows better performance since ResNet152V2 is better than VGG19 because of residual learning. Using both models together without feature merging yields slightly better performance since the two backbones are not adequate on their own.

Table 4: Ablation Study of the Proposed Hybrid FER Framework

Configuration	VGG19	ResNet152V2	Transfer Learning	Feature Fusion	Data Augmentation	CK+ Accuracy (%)	FER2013 Accuracy (%)
CNN	✗	✗	✗	✗	✓	89.0	73.0
VGG19	✓	✗	✓	✗	✓	94.0	84.0
ResNet152V2	✗	✓	✓	✗	✓	96.0	88.0
Hybrid without fusion	✓	✓	✓	✗	✓	95.0	86.0
Hybrid without transfer learning	✓	✓	✗	✓	✓	93.0	84.0
Hybrid without augmentation	✓	✓	✓	✓	✗	96.0	88.0
Proposed Hybrid	✓	✓	✓	✓	✓	98.0	92.0

The findings from the ablation experiment prove that the performance gains are mainly due to feature-level fusion and not because of the increase in network depth. The implementation of feature-level fusion greatly improves classification performance by leveraging fine-grained spatial information in VGG19 together with semantic features in ResNet152V2. In addition to that, the use of transfer learning and freezing of low layers ensures that there are faster rates of convergence and fewer instances of overfitting when using the FER2013 data set. Data augmentation becomes an important aspect in ensuring robustness of the model under non-restricted settings as regards posture, lighting, and facial expressions. To conclude, the hybrid model works well in comparison to other data sets and emphasizes the need for all aspects of efficiency in the model.

Despite its satisfactory performance, there are various weaknesses associated with the model. First, hybrid architectures can increase computational complexity and cause slower training times and inference times compared to single model-based systems. Second, an over-reliance on labeled data sets such as CK+ and FER2013 can negatively impact performance in real-life applications due to unseen or unbalanced emotion classes. Finally, the proposed model only considers still images of faces and ignores important temporal characteristics that are available when recognizing emotions from video.

Future improvements to the proposed emotion recognition framework will include the incorporation of attention mechanisms for improved selection of features and easier interpretability of results. Furthermore, expanding the proposed system to accommodate video-based FER using temporal models such as LSTM and transformers would contribute to the improvement in capturing dynamic emotions. Lastly, incorporating multimodal features like speech and physiology would present an exciting prospect to increase FER's accuracy and make it applicable to lightweight devices for real-time deployment.

V. CONCLUSION

This study presented a hybrid deep-learning framework for facial emotion recognition that combines VGG19 and ResNet152V2 through feature-level fusion. Through the combination of complementary facial features from space and semantics, the proposed method can achieve enhanced accuracy and performance across both laboratory-controlled settings and practical conditions. The experimental results conducted on CK+ and FER2013 datasets have confirmed that the hybrid architecture surpasses each of the baseline models individually. Through an ablation experiment, it can be shown that feature combination, transfer learning, and data augmentation are important components for performance optimization. In summary, the presented system is a viable option for emotion-aware intelligent applications.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

ACKNOWLEDGMENT

The author thanks colleagues and reviewers for their valuable discussions and constructive feedback, which helped improve this work. The author is grateful to the research guide for continuous guidance and support. The author also acknowledges the broader research community whose scholarly contributions influenced this study.

REFERENCES

[1] N. Siyad *et al.*, "Comparison and analysis of deep neural networks in facial expression recognition," *ResearchGate*. Available from: https://doi.org/10.54254/2755-2721/21/20231122

[2] T. Ahmed and R. A. Rahman, "Four-layer ConvNet for facial emotion recognition with minimal epochs and the significance of data diversity," *Scientific Reports*. Available from: https://doi.org/10.1038/s41598-022-11173-0

[3] Hassan *et al.*, "A survey on facial expression recognition using deep learning and explainable artificial intelligence," *IJRTI*. Available from: https://www.ijrti.org/papers/IJRTI2207297.pdf

[4] S. Singh and J. K. Pandey, "Facial emotion recognition: An efficient CNN approach," in *Proc. 2025 1st IEEE Uttar Pradesh Section WIE Int. Conf. Electrical, Electronics and Computer Engineering (UPWIECON)*, Dehradun, India, 2025, pp. 442–447. Available from: https://doi.org/10.1109/UPWIECON67212.2025.1139010

[5] S. Singh and J. K. Pandey, "Enhancing real-time surveillance video analysis with AI-powered deep learning techniques," in *AI and Deep Learning Enabled Surveillance System Using Image Processing*, J. K. Pandey, M. Rai, and F. Ahmad, Eds., 2026. Available from: https://doi.org/10.1108/978-1-80592-815-720261006

[6] S. Singh, J. K. Pandey, *et al.*, "Machine and deep learning techniques for emotion detection," in *Advancements in Facial Expression Recognition Using Machine and Deep Learning Techniques*, M. Rai and J. K. Pandey, Eds. IGI Global, 2023. Available from: https://doi.org/10.4018/979-8-3693-4143-8

[7] J. K. Pandey, D. Bhuva, A. Bhuva, S. H. Abbas, A. Verma, and T. T. Moharekar, "Classification and clustering of Internet of Things (IoT) for integrating IoT services platform," in *Proc. 2023 IEEE 2nd Int. Conf. Industrial Electronics: Developments & Applications (ICIDeA)*, Imphal, India, 2023, p. 25. Available from: https://ieeexplore.ieee.org/abstract/document/10295225

[8] I. Sumaya *et al.*, "Comparative analysis of AlexNet, GoogLeNet, VGG19, ResNet152V2-50, and ResNet152V2-101 for improved plant disease detection," in *Proc. 2024 2nd Int. Conf. Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*, 2024, pp. 1–6. Available from: https://ieeexplore.ieee.org/abstract/document/10863407

[9] S. Mascarenhas and M. Agarwal, "A comparison between VGG16, VGG19 and ResNet152V2 architecture frameworks for image classification," in *Proc. 2021 Int. Conf. Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*, 2021, pp. 96–99. Available from: https://ieeexplore.ieee.org/abstract/document/9687944

[10] Paulchamy, A. Yahya, N. Chinnasamy, and K. Kasilingam, "Facial expression recognition through transfer learning: Integration of VGG16, ResNet152V2, and AlexNet with a multiclass classifier," *Acadlore Trans. AI Mach. Learn.*, vol. 4, no. 1, pp. 25–39, 2025. Available from: https://doi.org/10.56578/ataiml040103

[11] R. Kumar *et al.*, "Transfer learning for facial expression recognition," *Information*, vol. 16, no. 4, p. 320, 2025. Available from: https://doi.org/10.3390/info16040320

[12] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114. Available from: https://proceedings.mlr.press/v97/tan19a.html?ref=ji

[13] S. Singh, M. Rai, *et al.*, "Emotional intelligence and collaborative dynamics in Industry 5.0 for human–machine interactions," in *Human-Machine Collaboration and Emotional Intelligence in Industry 5.0*, IGI Global, 2024. Available from: https://doi.org/10.4018/979-8-3693-6806-0

[14] Dhabliya *et al.*, "Using machine learning to detect emotions and predict human psychology," in *Ethical Considerations in Emotion Data Collection and IoT Integration*, IGI Global, 2024, pp. 52–74. Available from: https://doi.org/10.4018/979-8-3693-1910-9

[15] H. A. Santoso *et al.*, "Comparative analysis of convolutional neural network and DenseNet121 transfer learning in agriculture focusing on crop leaf disease identification," *Applied Computing and Informatics*, 2024. Available from: https://doi.org/10.1108/ACI-03-2024-0132

[16] H. Ge, Z. Zhu, Y. Dai, B. Wang, and X. Wu, "Facial expression recognition based on deep learning," *Computer Methods and Programs in Biomedicine*, vol. 215, p. 106621, 2022. Available from: https://doi.org/10.1016/j.cmpb.2022.106621

ABOUT THE AUTHORS

Ms. Shivani Singh is currently working as an Assistant Professor in the Department of Electrical & Electronics Engineering at Shri Ramswaroop Memorial University (SRMU), Barabanki. She completed her B. Tech and M. Tech in Electronics and Communication Engineering from Amity University. She is also pursuing her Ph.D. in Image Processing from SRMU. She has a strong academic background with research interests spanning image processing, medical image processing, artificial intelligence, and machine learning. Being an active researcher, she has published several papers in national and international conferences. Her dedication to both teaching and research continues to contribute meaningfully to the academic and scientific community.

Dr. Jay Kumar Pandey is currently working as an Assistant Professor in the Department of Electrical & Electronics Engineering at Shri Ramswaroop Memorial University, Barabanki (U.P.) India. Dr. Pandey has completed his Ph.D. and has done his Mtech. With specialization in Power Control (Instrumentation), and also done his MBA in Finance and Marketing. His subje) andest are related to Artificial Intelligence, Biomedical & Healthcare, Image Processing, Machine Learning, Business Management and Renewable Energy. He has 15 years of teaching and research experience, has published more than 30 research papers in National and International journals /conferences & Book Chapters in CRC, NOVA, Taylor & Francis, Springer, and IGI Publisher. Dr. Pandey is an editor of books (edited) published by

reputed publishers Apple Academy Press, IGI, Elsevier, Wiley-IEEE, NOVA, Bentham Science, IAP, Emerald Publication, Springer, CRC, Scrivener -Wiley, River Publisher, Taylor & Francis & Cambridge Scholar .Dr. Pandey is also Editor of the Journal of Technology Innovations and Energy United States. Dr. Pandey has been a reviewer in different conferences/journals/Book Chapters like (The Journal of Supercomputing in Springer Nature, IGI publishing, Journal of Security and Communication Networks Hindawi, Journal of Biomimetics, Biomaterials and Biomedical Engineering (JBBBE) Scientific Net, Switzerland, Advanced Engineering Forum (AEF) Scientific Net, Switzerland),Engineering Applications of Artificial Intelligence.