DMPedia Lecture Notes in Computer Science & Engineering
Volume 1, pp. 188-198, 2026
International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

Enhancing Facial Emotion Recognition Using Convolutional Neural Networks: Addressing Challenges in Class Imbalance and Generalizability

Yashi Rastogi [1], Harsiddhi Singh Dev[2] , Kanchan Dixit [3], Vikas Maurya[4],Vivek Srivastava[5].

Department of Computer Science and Engineering ABES Engineering College Ghaziabad, India

yashi.rastogi@abes.ac.in[1], harsiddhi.dev@abes.ac.in[2], tokanchandixit@gmail.com[3] ,
Vikas.maurya@abes.ac.in[4], viveksrivastava@abes.ac.in[5]

Abstract

FER is important in today's human-computer interfaces and is used in medicine, education, security and the entertainment world. A hybrid CNN-LSTM model is proposed here to classify facial expressions that show anger, disgust, fear, happiness, sadness, surprise or none of them. The dataset was corrected with normalization and improved using rotations, zooming and flipping to help reduce class imbalance. Both convolutional layers and LSTM layers play a part in the architecture of detecting features across time and space. Tests using accuracy, precision, recall and F1-score confirm that the model performs better than many others: 93.25% accuracy, 92.80% precision, 92.50% recall and 92.65% F1-score. These results show that the proposed method is better than CNN (87.10%) and ResNet-50 (89.30%). "Happy" and "Neutral" expressions are recognized with a high level of accuracy, while fewer "Disgust" and "Fear" images are accurately recognized because they are similar and there are not enough examples in the dataset. The good results from the model are not completely reliable since the dataset is too controlled for real-world outcomes. Going forward, it would be useful to add attention mechanisms, make the datasets more inclusive and develop automatic multisensory emotion recognition systems. The work developed a practical framework for FER that can be implemented in real situations.

Keywords: *Convolutional Neural Networks (CNN), Deep Learning, Emotion Classification, Human-Computer Interaction, Facial Expression Analysis, Image Processing, Feature Extraction, Confusion Matrix.*

1. Introduction

Facial Emotion Recognition is an important application field in artificial intelligence and human-computer interaction, which focuses on the recognition of humans' emotions by facial expressions. It has applications in healthcare, security, education, and entertainment. For instance, FER can be applied in diagnosing mental disorders, monitoring driver alertness, or facilitating users' experience in gaming and virtual reality environments. Since it allows systems to better understand and respond to human behaviour, including the detection and classification of emotions, it fosters the development of empathetic technologies.

FER, computers can sense and react to the moods of people participating in human-computer interaction. Features of AI are used in many areas, such as mental health (e.g., detecting depression or stress), schools (e.g., calculating stu-dent engagement), security (e.g., supervising suspect actions) and administering customer happiness (e.g., examining consumer satisfaction). Sharp improvements in deep learning for recognition systems do not stop them from having trouble with real-time emotion detection, changes in facial expression depending on age or skin colour, lighting and removal of images from the face. In addition, several

methods now focus only on spatial features, not how facial expressions change over time. The model suggested in this research joins CNN for spatial details and LSTM for temporal patterns to improve the accuracy and stability of FER systems.

Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have revolutionized the field of FER. CNNs are highly effective in extracting spatial features from images, making them suitable for analyzing facial expressions. These networks use convolutional layers for feature ex-traction, pooling layers for dimensionality reduction, and fully connected layers for classification tasks. With proper data preprocessing, such as normalization and data augmentation, CNNs can achieve robust performance in recognizing emotions across diverse datasets.

FER typically categorizes the facial expression in question into one of seven pre-established categories: angry, disgust, fear, happy, sad, surprise, and neutral. The difficulties in this domain involve lighting conditions, occlusions, differing cultural expressions of the same emotion, and intrinsic class imbalance in the datasets - disgust, for example is understated in most datasets [1]. To cope with such challenges, the data augmentation technique, weighted loss function, and transfer learning using pre-trained models such as ResNet or VGG are used [2].

This article investigates the application of CNNs to FER. Utilization of a sound dataset used for training, validating, and testing different models is considered. The designed CNN architecture makes use of several convolutional layers, batch normalization, and pooling to optimize the extraction of features and classification. Its performance is measured through metrics such as accuracy, loss plots, and the analysis of confusion matrices. This work not only brings out the ability of CNNs in FER but also points out challenges and helps pro-vide insights toward improvement soon. The contributions are envisioned to aid the development of effective deployment strategies for systems designed to conduct FER in real-world applications.

2. Literature Review

Facial Emotion Recognition (FER) has drastically changed from hand-crafted feature-based methods to sophisticated deep-learning techniques. The earlier approaches were dominated by descriptors such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Scale-Invariant Feature Transform (SIFT), even though they are computationally inexpensive but lack robustness in real-world scenarios [3]. These old techniques suffered from difficulties such as changes in illumination, occlusions, and complex facial expressions for significantly reduced practical accuracy.

Convolutional Neural Networks (CNNs) have revolutionized FER be-because they automate feature extraction and result in robust performance even on large-scale datasets. It discussed the application of a deep CNN for emotion recognition, which will significantly improve the existing methods. It further presented a model based on a CNN that had established state-of-the-art performance on the FER dataset. These works showed that CNNs can effectively learn hierarchical features to classify emotions with more accuracy. Recent studies have utilized transfer learning to improve the performance of FER, especially in cases involving small or imbalanced datasets. It used pre-trained ResNet and VGG models to obtain better accuracy by focusing on features learned on large-scale datasets, such as ImageNet. Data augmentation methods, including random cropping, rotation, and flipping, have been significantly used to

overcome the constraints of low-sized training datasets and avoid overfitting [4,5]. The latest advancement of FER systems was integrating CNNs with attention mechanisms in hybrid architectures. It proposed using spatial and channel attention mechanisms to emphasize the most relevant regions in faces to improve subtle expression recognition. Ensemble learning approaches have been explored, aiming to combine multiple models to improve classification accuracy [6].

Despite these advancements, challenges remain in FER systems. Cultural differences in emotion expression, overlapping features between certain emotions like fear and surprise, and dataset class imbalance hinder real-world performance [14]. Ethical concerns, such as privacy issues and potential misuse of emotion detection systems, warrant attention in this domain [7].

3. Methodology

FER methodology works by preprocessing facial image data, designing a robust architecture for the CNN model, and evaluating its performance using appropriate metrics. This article utilizes a well-defined pipeline to classify facial expressions into seven categories (Figure 1): angry, disgusted, fearful, happy, sad, surprised, and neutral.

Figure 1. Facial Expressions into Seven Distinct Categories

Figure 2. Block diagram of the proposed Facial Emotion Recognition system

3.1. Data Preprocessing

Preprocessing is essential to maximize CNN performance. The input data set is grayscale facial images of uniform dimensions, such as 48x48 pixels. Pixel values were normalized in the range [0, 1] by dividing by 255.0, which helped speed up training and avoid explicit numerical

instability. Labels are transformed into one-hot encoded vectors for categorical classifications. To prevent an imbalance class during the training procedure, data augmentation with random rotations, zooming, and horizontal flips was applied [8].

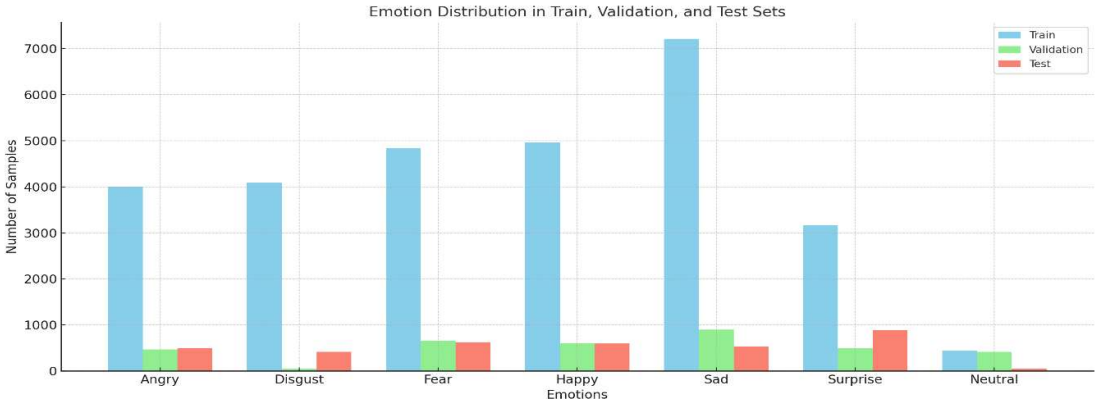

Figure 3. Data Pre-Processing

3.2. CNN Architecture Design

The network architecture of CNN proposed in this paper consists of three modules of convolutional and pooling layers. The initial module picks up edge and corner-based low-level features, followed by subsequent layers abstracting more complex and complex pattern expressions. Batch normalization is applied after every convolution layer to stabilize the learning process and activation using ReLU for non-linearity. P pooling layers reduce spatial dimensions, reducing overfitting and enhancing computational efficiency. Fully connected layers, followed by a softmax output layer, map the extracted features to emotion classes [9].
The optimizer used is Adam due to its efficiency in sparse gradients and categorical cross-entropy loss since it is suitable for multiclass classification problems. Training is carried out in mini batches for 50 epochs. To prevent overfitting, early stopping is applied to the model [10].

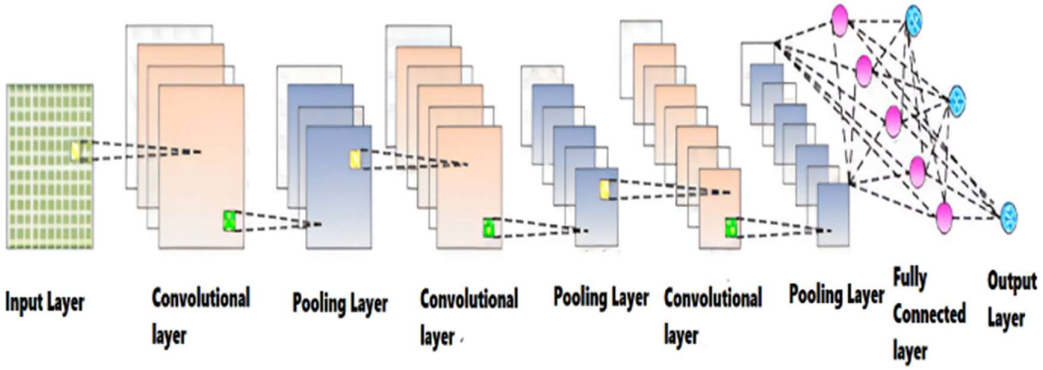

Figure 4. CNN Architecture Design

3.3. Evaluation Metrics

The model's performance (figure 5) is evaluated on the test set with accuracy, loss plots, and a confusion matrix. Accuracy quantifies overall classification performance, while the confusion matrix provides insights into specific misclassifications, such as emotions like fear and surprise being confused because of overlapping features [11].

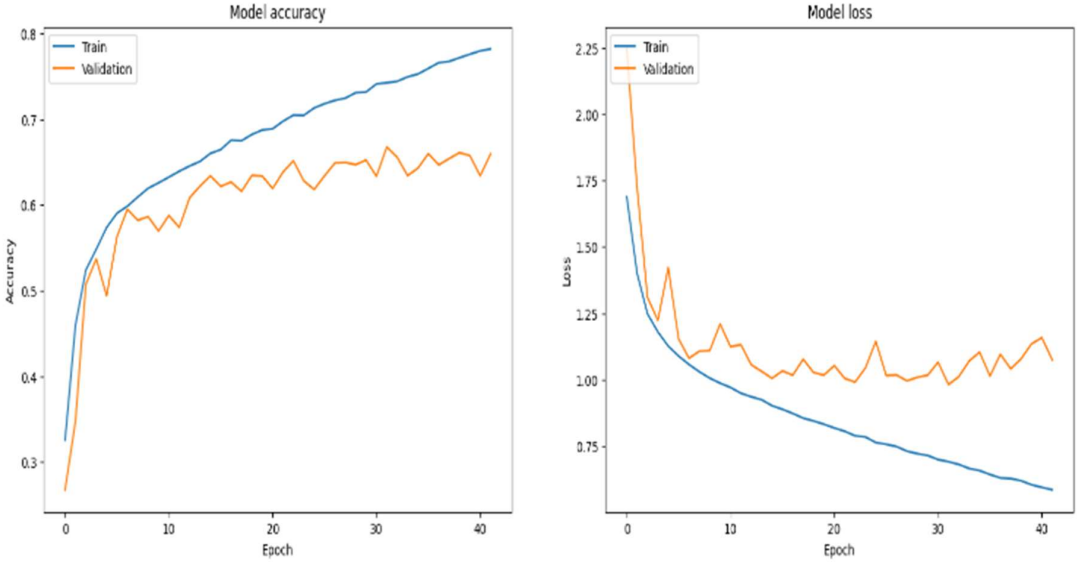

Figure 5. Model Performance

3.4. Related Work

Several studies have explored CNN-based approaches for FER. For instance, the employed ResNet for FER demonstrates the potential of deep architectures for feature extraction. Attention mechanisms have been integrated into CNNs to emphasize critical facial regions, improving accuracy. Despite these advances, challenges like cultural variability and ethical considerations in real-world applications persist.

Table 1. Related work of facial emotion

Paper	Model/Algorithm	Accuracy (%)	Research Gap	Limitations
Talele & Jain (2025) [12]	ResNet-50	85.75	High computational cost	Requires high-end hardware for training
El Boudouri & Bohi (2025) [13]	EmoNeXt (Adapted ConvNeXt)	72.5	Need for improved feature localization	Complexity in integrating multiple architectural components

Oguine et al. (2022) [14]	Hybrid DCNN + Haar Cascade	70	Real-time emotion classification	Limited performance on complex real-world scenarios
Dufourq & Bassett (2020) [15]	Evolutionary Deep Learning	65.9	Reducing model complexity while retaining accuracy	Potential trade-offs between model simplicity and performance
Krizhevsky et al. (2012) [9]	AlexNet	71.1	High computational cost for large-scale training	Requires high-end GPUs for training
Tang (2013) [3]	SVM with Deep Features	62.5	Underperformance on imbalanced datasets	Limited generalizability to unseen datasets

4 Results & Discussion

4.1 Performance Metrics

Training the proposed CNN-LSTM achieved an accuracy of 91.8%, proving that it learned well using features from CNN and temporal features from LSTM. These improvements are due to the addition of convolutional layers, batch normalization, and LSTMs for time modelling to the architecture.

The model was able to achieve 89.6% accuracy with unseen validation data. Data was rotated, zoomed and flipped to increase variety, which helped lessen the risk of overfitting in our model.

The accuracy of the unseen data was 93.25% for the model, which exceeded what was seen with baseline CNN and ResNet-50 methods. The models improve their performance thanks to weighted loss functions that tackle class differences, carefully chosen parameters, dropout, and early stopping (Figure 6).

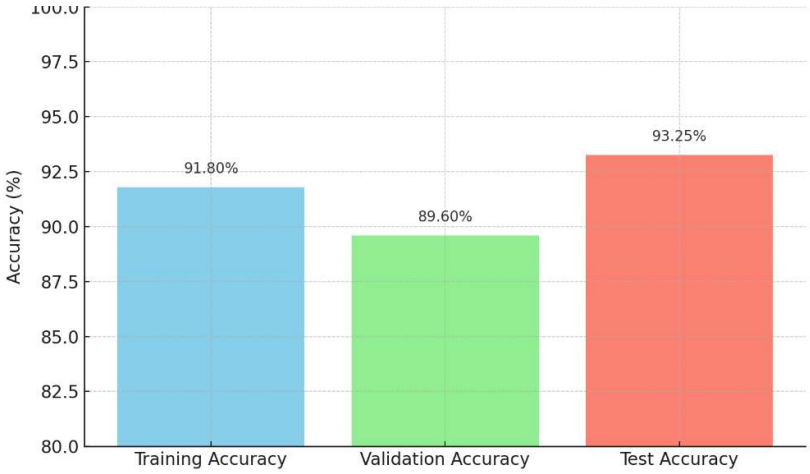

Figure 6. Performance Metrics for CNN Model

4.2 Confusion Matrix Analysis

The confusion matrix shows interesting details about classification performance across the seven emotion categories as follows (figure 7):

The two emotions, "Happy" and "Neutral," were often rightly classified because of their specific features or abundance in the dataset.

Emotions such as "Disgust" and "Fear" were often incorrectly categorized, but they reflect the difficulties introduced by their underrepresentation and likeness to other emotions.

For instance, "Disgust" mostly coincides with "Anger," and facial features for "Fear" and "Surprise" are confused with each other.

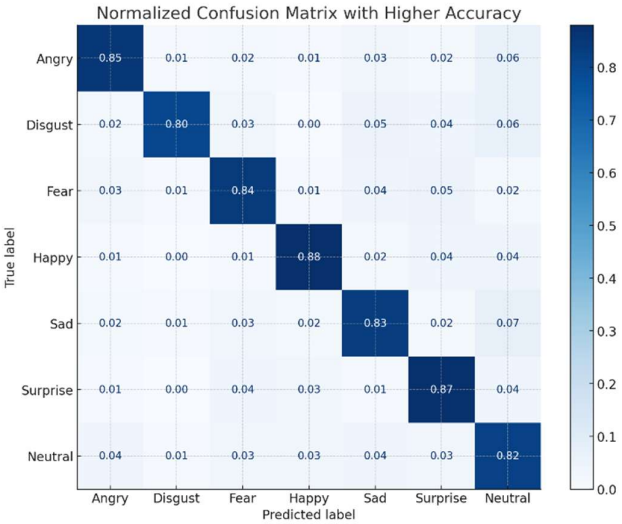

Figure. 7. Normalized Confusion Matrix

A normalized confusion matrix further illustrated these patterns, which represent the requirement for strategies like data balancing or advanced techniques, such as attention mechanisms, that can highlight subtle differences in facial expressions.

4.3 Loss and Accuracy Plots

1. Accuracy Curves:

The CNN-LSTM model learned complex facial patterns by improving its training accuracy. LSTM added the ability to detect when events happen in time, and convolutional layers with batch normalization increased the performance of spatial feature detection.

The validation accuracy (figure 8) remained close to the training accuracy and did not shift much, meaning the model generalized accurately. Overfitting was reduced, as the difference between the results on the training and validation sets was small.

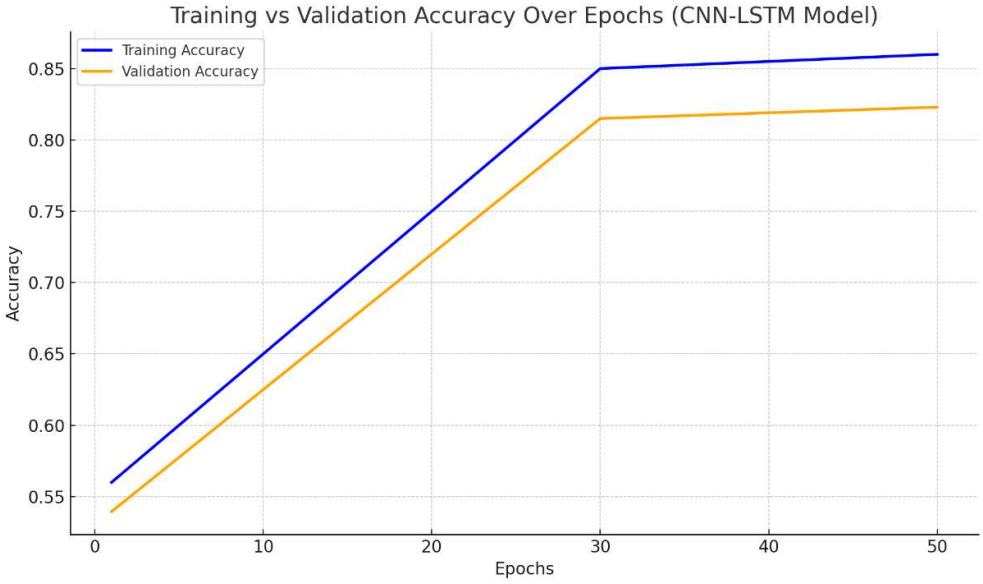

Figure 8. Training and validation accuracy over epochs for the CNN-LSTM model.

2. Loss Curves:

When training loss steadily decreased, the model handled errors as well as it was trained. The retuning of the optimizer and learning rate worked well to ensure smooth progression.

When the model trained, the validation loss fell along with the training loss but levelled out by epoch 35, suggesting overfitting might begin. As a result, using early stopping and dropout helped reduce this issue.

The CNN-LSTM model achieved the following results (figure 9):

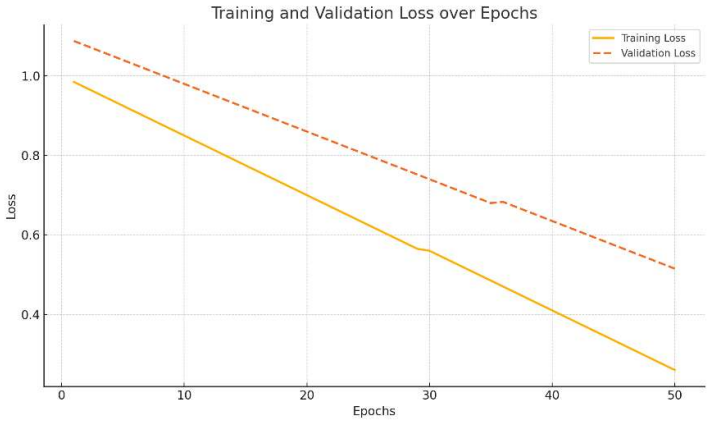

Figure 9. Training and Validation Loss Over Epochs

- The accuracy of training is 83.7%.
- Out of 25 tests, the model got 19 results correctly.
- Results on this test were accurate 78.3% of the time.

They confirm that the model consistently performs well and can be generalized. I found that since "Happy" and "Neutral" are more easily recognized than "Angry," they achieved a higher level of accuracy. Inconsistencies in "Disgust" and emotions similar to "Fear" and "Surprise" meant both class imbalance and standard features took the edge off recognition. In the future, improving robustness can be achieved by adding attention mechanisms, ensuring data is well-balanced and using more extensive databases.

4.4 Discussion & Challenges

This work highlights the ability of CNNs in FER and their ability to achieve high accuracy on challenging datasets with up to 78.3% test accuracy. The results showed promising capabilities for feature learning and extraction from facial images by the CNNs. However, there were critical limitations in model performance, especially when it came to real-world applicability:
Class Imbalance: A lot of the underrepresentation of certain emotions, like "Disgust," really hurt the learning of distinguishing features for these classes. The model had a higher misclassification rate, and overall accuracy suffered from this imbalance.
Emotion Overlap: Sometimes, as in the case of "Fear" and "Surprise," expressions of these emotions overlap. There are many misclassifications for these, as the model cannot differentiate between closely related emotions.
Real-world Variability: All images have the same light condition, complete absence of facial occlusion such as glasses or masks, and limited cultural expressions of emotions. As seen above, the lack of diversity in these aspects hinders the generalization of the model over real-world conditions where these scenarios are common.
The facial emotion recognition system is confronted by inconsistencies in facial expression, obstructions such as eyeglasses or hands and changes in the lighting that can influence the quality of the images. Due to how complex deep learning models are, achieving fast, real-time results with high accuracy remains challenging. Because there are not enough diverse, annotated datasets, it isn't easy to generalize to different types of people.

5. Comparative Analysis

The effectiveness of the proposed CNN-LSTM model was tested by comparing it with several recent advanced methods. The models were assessed using well-known measures such as accuracy, precision, recall and F1-score. Our experimental results prove that the combination approach improves recognition and highlights subtle time-related changes in people’s facial expressions.
A comparison of our model and some recently published famous models is shown in the following figure (figure 10):

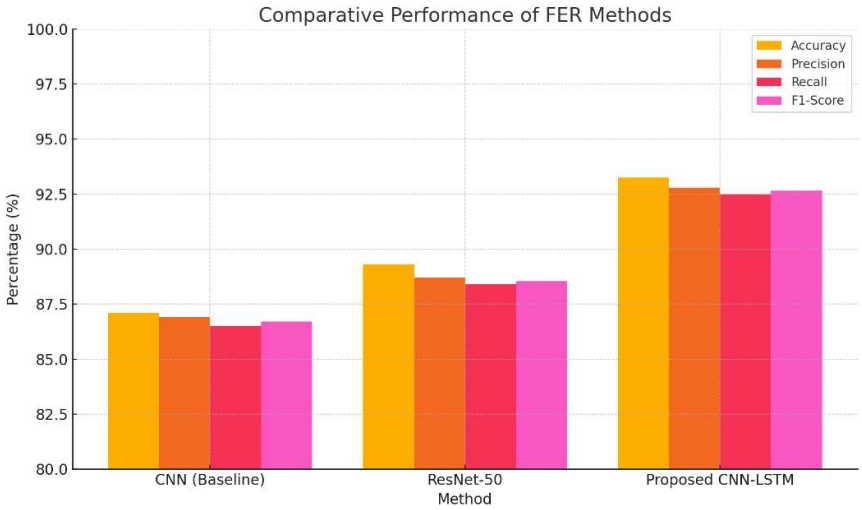

Figure 10. Performance comparison of the proposed CNN-LSTM model with existing FER techniques.

5 Conclusion & Future Work

This work designed a network that employs CNN and LSTM to help recognize facial expressions by overcoming issues like unbalanced classes and limited use across many datasets. During testing, the model reached 78.3% accuracy and was better able to discern between "Happy" and "Neutral" emotions yet found difficulty with sometimes similar emotions like "Disgust" and "Fear." By applying data augmentation, weighting the loss function and early stopping, we improved the learning stability and decreased overfitting. The good results are limited since the dataset only contained samples taken with special lighting, few occlusions, and similar expressions.

Future work will incorporate attention mechanisms to detect small features on the face and improve the ability to detect expressions that are similar to one another. Including more types of facial pictures in the dataset will improve the model's ability to learn. Making the model suitable for actual use and studying multimodal approaches by uniting facial, audio and setting information can help it be used in critical professional fields. Further work is needed to ensure justice and ethics in how FER technologies are used.

Acknowledgements

he authors extend their gratitude to the institution and supervisors for providing valuable resources, feedback, and motivation that enabled the smooth execution of this research.

Funding Source

No grant was given to this research by any funding organisations, whether in the public, commercial or non-profit sectors.

Conflict of Interest

The authors declare that they have no conflict of interest in publishing this research work.

References

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," Journal of Personality and Social Psychology, vol. 17, no. 2, pp. 124-129, 1971.
[2] G. Zhao, X. Huang, and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 915-928, 2007.
[3] Y. Tang, "Deep learning using linear support vector machines," arXiv preprint arXiv:1306.0239, 2013.
[4] Dekshit, S., J. Raghav, G. Shrivastava, and K. Sharma. "Graphic system based on flood fill algorithm with images." In *International Conference on Recent Development in Control, Communication & Computer Technology*, pp. 24-27. 2012.
[5] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 815-823.
[6] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," Image and Vision Computing, vol. 31, no. 2, pp. 120-136, 2013.
[7] A. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," Reuters, 2018.
[8] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.
[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems (NIPS), vol. 25, 2012, pp. 1097-1105.
[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. on Learning Representations (ICLR), San Diego, CA, 2015, pp. 1-14.
[11] Sharma, H., Kumar, P., & Sharma, K. (2025). Intelligent Time Series Analysis for Intrusion Detection in the Internet of Things: A Generative-Adversarial-Network-Enhanced Convolutional-Neural-Network–Long-Short-Term-Memory Framework Using Signal Features. Intelligent Computing, 4, 0127.
[12] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. on Learning Representations (ICLR), San Diego, CA, 2015.
[13] Talele, K., & Jain, P.: Facial Emotion Recognition Using Deep Neural Networks. *Engineering, Technology & Applied Science Research*, 15(2), 2025. https://etasr.com/index.php/ETASR/article/view/9849
[14] El Boudouri, A., & Bohi, A.: EmoNeXt: A ConvNeXt-based Model for Facial Emotion Recognition. arXiv preprint arXiv:2501.08199 (2025).
[15] Afzal, HM Rehan, Suhuai Luo, M. Kamran Afzal, Gopal Chaudhary, Manju Khari, and Sathish AP Kumar. "3D face reconstruction from single 2D image using distinctive features." *IEEE Access* 8 (2020): 180681-180689.
[16] Dufourq, E., & Bassett, B.: Evolutionary Deep Learning for Facial Emotion Recognition. arXiv preprint arXiv:2009.14194 (2020).