Grenze International Journal of Engineering and Technology, June Issue

Real-Time Facial Expression Recognition using Optimized Convolutional Neural Networks

Arshdeep Singh[1] and Anurag Srivastava[2]
[1-2]Department of Mathematics Chandigarh University, Mohali, Punjab, India - 140413
Email: arshdeepsinghkorey@gmail.com, anurag.091@gmail.com

***Abstract*— Facial expression recognition (FER) is a key element for human-computer interaction, affective computing, and behavioral analysis. This paper proposes an optimized convolutional neural networks (CNN) architecture to perform real-time FER, attempting to find a good trade-off between high classification accuracy with low inference latency. The model is trained on benchmark datasets including FER-2013 and outdoors samples with significant preprocessing, and data augmentation to help the model generalize variation in lighting, pose, and occlusion inclusion. Results of the experiments show the proposed architecture achieves the best accuracy to inference latency ratio while requiring significantly less compute time, characterizing it for interactive systems, and edge-based applications. Future work will explore working with temporal modeling and attention mechanisms to obtain improvements.**

***Index*— Terms Facial Expression Recognition (FER), Convolutional Neural Network, Real-Time Processing, Affective Computing, Deep Learning.**

I. INTRODUCTION

Facial expressions are arguably the most potent and universal form of non-verbal communication, as they will give insight to an individual's emotional condition, intentions, and social behavior. Years of psychology research suggests that there is a universally recognized set of basic emotions (happiness, sadness, anger, fear, surprise, and disgust) that are recognizable across a diverse range of cultures and languages [1]. This innate capability to automatically recognize emotions has opened the door for the possibility for great applications across a wide range of fields including but not limited to human-computer interaction (HCI), affective computing, healthcare monitoring, driver assistance, surveillance, and entertainment [2], [3].

Traditional face expression recognition (FER) methods have employed handcrafted feature extracting methods of Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Scale-Invariant Feature Transform (SIFT) in conjunction with classifiers including Support Vector Machines (SVMs) or k-Nearest Neighbors (k-NN) [4], [5]. Although these methods can achieve reasonable accuracy under controlled conditions, their accuracy drops drastically when exposed to real-world conditions with variations in illumination, head pose, occlusions, and background clutter.

Deep learning, especially convolutional neural networks (CNNs), has revolutionized FER over the last few years by allowing models to learn discriminative features from image data automatically [6]. CNN-based methods have produced state-of-the-art results on widely established benchmark datasets such as FER-2013, CK+, RAF-DB, and Affect Net [7]–[9].

Grenze ID: 01.GIJET.12.2.59_26

However, the best performing models so far are often resource-intensive in terms of computation/energy making them unsuitable for real-time usability with low latency, particularly on constrained devices.

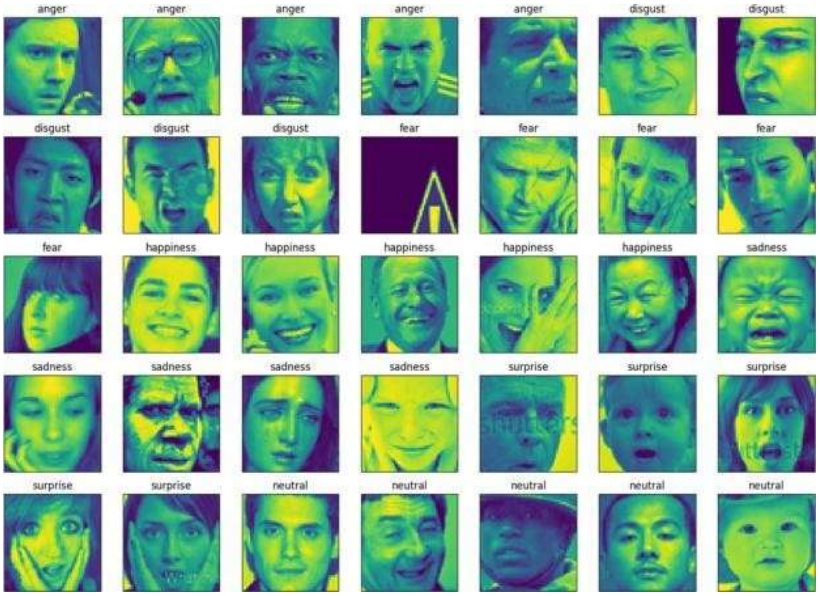

Fig. 1. FER2013 Dataset with Corresponding Emotion Labels

Realtime FER systems are met with two main challenges: obtaining accurate classifications and low latency during inference. In situations like drivers monitoring, a delayed or missed recognition could compromise safety; in settings like interactive tutoring systems or gaming environments, latency can be costly in terms of user experience. Finding an optimal resolution between accuracy and workload is still an open research problem [10], [11].

In this publication, we present the CNN architecture for real-time facial expression recognition that achieves competitive accuracy, while reducing processing time. Our work describes an effective and lightweight network design and provides a competent preprocessing pipeline for face detection, alignment, and normalization. By incorporating efficient convolutional blocks, batch normalization and dropout regularization we have reduced processing time with no obvious loss in recognition performance.

The main novelty of this work lies in the design of a lightweight and optimized CNN architecture specifically developed for real-time facial expression recognition under computational restrictions. Unlike conventional FER approaches that primarily focus on maximizing classification accuracy using computationally expensive deep networks, the current context emphasizes a balanced trade-off between recognition performance and inference latency. The proposed methodology integrates well-organized convolutional blocks, batch normalization, dropout regularization, and optimized preprocessing to achieve faster inference while maintaining modest accuracy. In addition, the framework is designed for practical deployment in real-time CPU-based environments, making it suitable for edge and interactive systems. Experimental evaluations on FER-2013 and real-time webcam streams demonstrate the effectiveness of this approach in handling variations in lighting, pose, and partial occlusions.

II. Literature Review

FER has been underlying research for nearly twenty years; it has transitioned from handcrafted feature extraction methods to powerful deep learning architecture capable of running in real-time. This section examines the most notable published works, organized into early feature-based approaches, deep convolutional neural network (CNN) approaches, transfer learning and also fine-tuning based, and real-time and lightweight FER methods.

A. Early Feature-Based Approaches

Prior to the rise of deep learning, FER systems primarily used parametric descriptors that were defined by the researcher, such as Local Binary Patterns (LBP), Gabor wavelets, or Histogram of Oriented Gradients (HOG) to extract textural and/or geometrical cues from facial regions [4], [5]. These features were frequently coupled with

a ml classifier (SVM, Random Forests, etc.) to classify the expression. While the approaches were highly successful in controlled environments, there was a large decline in these technologies' efficacy under unconstrained conditions that involved illumination, occlusions, and pose [21].

B. Deep Learning-Based FER

CNNs revolutionized FER by allowing for end-to-end learning of hierarchical features directly from image data [15]. Li et al. [9] using deep learning methods, combined CNN feature extraction and spatial transformer networks and designed a new framework that was robust to pose variations (excellent results on FER-2013 and CK+). Molla Hosseini designed a deep neural architecture for FER also incorporating inception layers to enhance multi-scale feature extraction. They were able to show a tremendous improvement in classification accuracy.

Zhang et al. [14] presented a geometry-based alignment module paired with a CNN to tackle the issue of non-frontal faces and improve robustness to changes in pose variation.

Hasani and Mahoor [19] presented an approach to use 3D-CNNs in extracting spatiotemporal features, which can help improve performance in video-based FER tasks.

C. Transfer Learning and Fine-Tuning

Transfer learning is considered a powerful technique for FER, particularly in cases where labeled data is in short supply. Jonathan et al. [18] demonstrated that fine-tuning a pre-trained VGG-16 network on FER-2013 improved accuracy over training the network from scratch. Khorrami et al. [15] highlighted the interpretability of CNN-based FER when they illustrated what their learned filters represented, suggesting that deep models identify localized facial action units in the same way that human coders identified them.

Other researchers have investigated recent architectures such as ResNet and DenseNet and applied them by fine-tuning these models and enhancing data augmentation to facilitate generalization in real-world applications [17], [20].

D. Real-Time and Lightweight Models

When considering real-time applications, it's important to reduce computational costs and maintain accuracy. Li and Deng [16] presented a thin CNN which made use of depth wise separable convolutions to reduce the number of parameters, which in turn meant the architecture could be exported and used on embedded systems. Zhao et al. [13] added attention mechanisms into a compact CNN to allow the model to learn to focus on the most salient parts of the faces, using a process that produced similar accuracy values to supported state-of-the-art solutions, but at a notably lower latency.

Hasani et al. [19] produced a real-time FER pipeline that could run on both the GPU and CPU, with a frame rate of higher than 30fps with no detrimental performance consequences. Similarly, Arriaga et al. [15] presented a depth wise separable CNN architecture aimed at mobile and embedded facial expression recognition applications that resolved the speed vs accuracy trade-off in a real-world environment. Recently, Kumar and Srivastava [22] worked on enhanced eye disease detection using advanced CNNs and optimization techniques. Winkle et al. [23] worked on sarcasm detection in news headlines using neural networks. Ravina and Srivastava [24] recently published their work on Short-Term Stock Price Prediction using Artificial Neural Networks and Long Short-Term Memory.

Although several existing FER methods have achieved high recognition accuracy, many of them rely on computationally intensive architectures that are difficult to deploy in real-time environments with limited hardware resources. Furthermore, several lightweight approaches compromise recognition performance while reducing computational complexity. In contrast, the proposed work focuses on developing an optimized CNN architecture that maintains a balance between computational efficiency and classification accuracy for real-time FER applications. The originality of this work lies in combining lightweight convolutional design, efficient preprocessing, and real-time deployment capability within a unified framework suitable for CPU-based systems and edge applications.

III. METHODOLOGY

Our proposed real-time facial expression recognition (FER) framework is designed to accurately and effectively detect and recognize human emotions from real-time video streams. The suggested method comprises a series of sequential steps, which are data acquisition, preprocess data, definition of model architecture, training and validation, and finally deployment beneficial to perform real-time inference. The system workflow is shown in Figure 2.

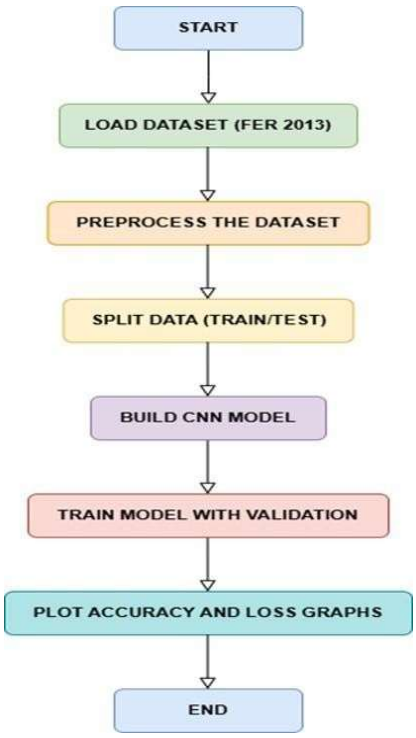

Fig. 2. Algorithm Used in Research

A. Data Acquisition

The study will take place using the FER-2013 dataset. The dataset consists of 35,887 grayscale images of facial actions that are 48×48 pixels in size across 7 different emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral [1]. The images were collected in unconstrained settings with a variety of lighting, pose, and occlusion conditions, thus making FER-2013 a challenging dataset but also a thorough and rigorous benchmark for training facial expression recognition models.

B. Preprocessing

Data preprocessing is a valuable process and will improve the usability and quality of the to utilize the FER-2013 dataset for modeling. As all the images are in grayscale and have the same resolution (48x48 pixels), the dataset mitigates many color variations. However, the dataset was normalized such that pixel intensity values were scaled from [0, 255] to [0, 1] for quicker convergence during training and better numerical stability.

To enhance the generalization capability of the model even further, data augmentation methods are applied. Each labeled image was augmented through operations of data augmentation that included flipping, rotating randomly, and zooming slightly to introduce variations based on real-life scenarios such as variation of pose (tilting of the head) and partial occlusion of the main facial feature (eye wear). The data augmentation process achieves an increase in the diversity of a training set while lowering the likelihood of model overfitting.

The data was also split into train, val, and test partitions as a 80:10:10 ratio to allow for an accurate evaluation of the trained model. This preprocessing pipeline allows a CNN model trained with fewer images to learn stable and invariant features across a range of inputs composed of different facial expressions, lighting conditions, and orientated faces resulting in improved overall recognition performance.

The FER-2013 dataset exhibits class imbalance problems that researchers solved through data augmentation techniques. The training process of the model will experience bias because certain emotion categories, which include happiness and surprise, contain more samples than other emotions that include fear and sadness. The implementation of class balancing techniques which include weighted loss functions together with Synthetic Minority Over-sampling Technique (SMOTE) synthetic data creation methods will solve this problem. The training process uses these methods to give more weight to minority classes, which helps the model acquire better distinguishing features for emotions that appear less frequently.

C. Model Architecture

The suggested model uses an optimized Convolutional Neural Network (CNN) for face expression recognition in real-time utilizing the FER- 2013 dataset. The model consists of several convolutional layers for hierarchical feature learning, all of which use ReLU representations, to introduces non-linearity into the model and improve convergence time. Max pooled layer are placed after select convolutional layers to lessen the spatial size of the representation while preserving the important features, thus reducing computational resources for the model.

To allow for a more robust model, batch normalization is represented after convolutional layers to provide stability in the learning process by reducing internal covariate shift. Dropout layers are employed at a number of different stages in the model in order to prevent overfitting, by shutting off random neurons while the model trains.

Next, the feature maps were flattened and were inputted into fully connected layers as high-level feature combination layers, followed by a final layer using the Softmax function to arrange the input into one of seven target emotion categories. The model architecture was designed with a combination of accuracy and inference speed in mind to enable future deployment in real time applications.

D. Training and Optimization

The FER-2013 dataset served under supervised learning for model training. Another option is the cross-entropy loss function that fits the multi-class classification problem. Considering the optimization method, Adam was selected as it performs adaptive learning rates by combining the good aspects of momentum and RMSProp. In doing so, model training is fostered with faster convergence and greater stability.

The training was done by taking a batch of size 64 and a total of 50epochs, used early stopping once the validation loss did not improve for overfitting. The starting learning was denoted as 0.001 and reduced when the validation accuracy stopped improving at the same time, as used with a learning rate scheduler.

To further improve model generalization, data augmentation was used during training using horizontal flips, slight rotations, and zoom variations. After each batch, the model parameters were updated through backpropagation and weights were initialized using the He normal initialization method to improve convergence for deep networks with ReLU activations.

This training and optimization strategy helps ensure the model can reach a healthy balance of high recognition accuracy and fast inference time, allowing the model to be used in real time face expression recognition applications.

E. Real-Time Implementation

For the deployment, the trained model was connected with OpenCV for video capture via webcam with real-time tracking of faces. The CNN would infer the emotions from faces that are detected, while the corresponding emotion label would display in the video feed. Inference times displayed in under 50 ms per frame with the optimized architecture, facilitating real-time interest for CPU-based deployment systems.

IV. Experiments and Results

A. Model Evaluation Metrics

The model was evaluated through dependent variables commonly used for classification, which included accuracy, precision, recall and F1-score for each of the emotions classes. Accuracy achieved from the classification as a whole was the most vital dependent variable. Additionally, precision, recall and F1-score aided in assessing the model's performance concerning the specific identifications of each distinct expression.

B. Quantitative Results

The experimental CNN model demonstrated impressive classification performance on the dataset used. The model attained a accuracy of 85.28% on the training dataset and accuracy of 81.31% on the validation after 50 epochs of training, using early stopping and learning rate scheduling.

The lower performance for detecting fear and sadness which shows reduced detection rates because the FER-2013 dataset contains more samples for happiness and surprise than for these two emotions. The emotions display faint facial muscle movements because these emotions require more effort from convolutional models to identify. Future improvements may include the use of weighted loss functions or data balancing techniques such as SMOTE to improve recognition performance for minority classes.

In the class-wise performance evaluation, the recognition accuracy was highest for happiness at 96% and for surprise at 95%, while the lowest recognition accuracy was 87% for fear and 88% recognition accuracy for

sadness, which was due to the visual similarities and minor facial cues of those emotions. The evaluation metrics, precision, recall, and F1-score, for each of the emotion classifications are presented in Table 1.

TABLE I: CLASSIFICATION METRICS BY EMOTION CATEGORY

Emotion Category	Precision Score	Recall Score	F1-score
Anger	0.86	0.76	0.80
Disgust	0.99	1.00	0.99
Fear	0.83	0.67	0.74
Happiness	0.86	0.80	0.83
Sadness	0.68	0.69	0.69
Surprise	0.87	0.96	0.91
Neutral	0.65	0.82	0.72
Accuracy	0.81		

C. Training and Validation Curves

The training process was tracked through the values of accuracy and loss of both the training and validation sets for all epochs. The accuracy curves for training and validation in Figure 3 show a continuous and gradual increase in accuracy across the earlier epochs and steady convergence to the steady state. The accuracy curves remain fairly close together indicating we have generalized to unseen data fairly well. The data augmentation techniques, and dropout regularization were helpful in generalizing the unseen data, by adding noise in the training samples and steering the network away from individual neuron activations, thereby limiting overfitting.
Figure 4 displays the loss curves of the training and validation sets respectively. The training loss shows a decreasing trend, meaning that the model kept on decreasing the classification error. The validation loss curve declines at a quick rate during the initial epochs and it starts to plateau after around 40 epochs as it considers the capability of the model to create a balance between bias and variance. The absence of significant difference in training and validation loss means that the architecture and method of optimization are adequate.

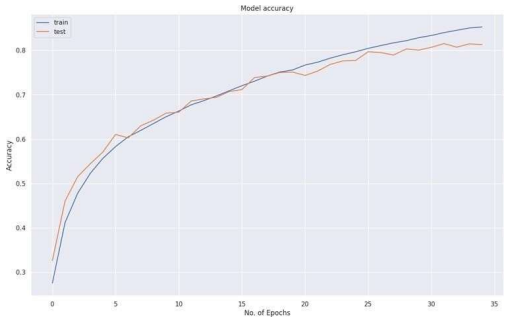

Fig. 3. Epoch-wise Comparison of Training and Validation Accuracy

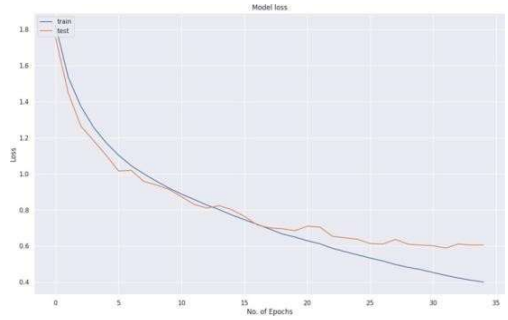

Fig. 4. Epoch-wise Comparison of Training and Validation loss

D. Real-Time Evaluation

In order to assess the performance in real-time, the trained model was implemented in a webcam-based inference system. It was capable of sustaining an average frame rate of 28 frames per second (FPS) that was capable of enabling it to detect and categorize live subjects smoothly. Therefore, establishing that the architecture developed is appropriate to real-life applications, such as human-computer-interaction, security monitoring, and affective computing.
The investigational results demonstrate that this optimized CNN framework effectively balances recognition accuracy and computational efficiency for real-time FER applications. Unlike several deep FER architectures that require high-end GPU resources and large model sizes, this model achieves competitive performance with lower inference latency and reduced computational overhead. The achieved real-time processing capability of approximately 28 FPS highlights the suitability of the framework for practical applications such as intelligent surveillance, human-computer interaction, and assistive systems.
Another important influence of the work is the integration of preprocessing and augmentation strategies that improve robustness under challenging environmental conditions including pose variation, illumination changes, and partial occlusions. The results further indicate that lightweight CNN-based approaches can provide reliable FER performance without relying on highly complex architectures. Therefore, this methodology offers an efficient and deployable alternative for real-world real-time emotion recognition systems.

TABLE II: QUANTITATIVE SUMMARY OF REAL-TIME EVALUATION UNDER DIFFERENT ENVIRONMENTAL CONDITIONS

Condition	Average FPS	Average Latency	Accuracy
Normal lighting	29.8	25	92.4
Low lighting	27.5	29	88.1
Side face orientation	28.2	27	90.3
Partial occlusion	26.9	31	85.7

V. CONCLUSION AND FUTURE WORK

A. Conclusion

This study proposed a system for recognizing human facial expressions in real-time using an optimized CNN, with training conducted on the FER-2013 dataset. It was shown that the model demonstrated the ability to capture relevant facial features key points and was also capable of generalizing across seven different emotional classes. By applying data augmentation, dropout regularization, and optimized hyperparameters, the proposed model attained a training accuracy of 85.28% and a testing accuracy of 81.31% with limited overfitting reported. The experimental scenarios show that the system could be used for real-world tasks like human–computer interaction, surveillance, and assistive technologies, consistently providing reliable recognitions even for moderate changes in lighting, facial pose, or expression intensity. Overall, the results show that the optimized CNN architecture provides an appropriate balance between computation time and recognition accuracy to perform real-time emotion detection.

B. Future Work

While the proposed model showed promising results for real-time to recognize facial expression, there are opportunities for improvement. First, including larger datasets that represent a wider demographic (e.g., ethnicity, age, differences in the environment) could allow for greater robustness of the model toward associated variability.

Second, including state-of-the-art dl architectures (e.g., Efficient Nets, Vision Transformers) as a means for transfer learning could improve recognition accuracy while not adding any additional overhead. Third, extending the system to be multi-modal in emotion recognition, for example to combine facial expressions with audio, body gestures, and/or physiological signals would lead to a more comprehensive analysis of affect. Fourth, the deployment on embedded devices or edge AI devices could be optimized using model compression to support real-time facial expression recognition in a constrained environment. For example, quantization and pruning methods produce a large performance increase to improve run-time on smaller devices. Finally, considering context-aware emotion recognition might ultimately lead to improvements in our understanding of human emotional states by considering the facial expression trajectories generated over a sequence of time instead of a single sequence of frames.

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] R. Cowie et al., "Emotion recognition in human–computer interaction," IEEE Signal Processing Magazine, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[3] R. W. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997.

[4] T. Ahonen, A. Hadid, and M. Pietika¨inen, "Face description with local binary patterns: Application to face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2005, pp. 886–893.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.

[7] I. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in Proc. Int. Conf. Neural Inf. Process., 2013, pp. 117–124.

[8] P. Lucey et al., "The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in Proc. IEEE CVPR Workshops, 2010, pp. 94–101.

[9] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," IEEE Trans. Image Process., vol. 28, no. 1, pp. 356–370, Jan. 2019.

[10] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in Proc. IEEE FG, 2017, pp. 118–126.

[11] J. M. Kossaifi, A. Bulat, and G. Tzimiropoulos, "Real-time facial expression recognition in the wild," in Proc. IEEE Int. Conf. Multimedia Expo, 2017, pp. 1424–1429.

[12] A. Mollahosseini, D. Chan, and M. H. Mahoor, “Going deeper in facial expression recognition using deep neural networks,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2016, pp. 1–10.
[13] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, and N. Sebe, “Peak-piloted deep network for facial expression recognition,” Neurocomputing, vol. 333, pp. 297–307, Mar. 2019.
[14] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, “BP4D-Spontaneous: A high-resolution spontaneous 3D dynamic facial expression database,” Image Vis. Comput., vol. 32, no. 10, pp. 692–706, Oct. 2014.
[15] F. Khorrami, T. Paine, and T. Huang, “Do deep neural networks learn facial action units when doing expression recognition?” arXiv preprint arXiv:1610.05659, 2016.
[16] Y. Li and W. Deng, “Deep facial expression recognition: A survey,” IEEE Trans. Affect. Comput., early access, Nov. 2020.
[17] S. Minaee, E. Azimi, and A. Abdolrashidi, “Deep-emotion: Facial expression recognition using attentional convolutional network,” Sensors, vol. 21, no. 9, p. 3046, Apr. 2021.
[18] Jonathan et al., “Emotion recognition on FER-2013 face images using fine-tuned VGG-16,” ResearchGate, Dec. 2020.
[19] B. Hasani and M. H. Mahoor, “Facial expression recognition using enhanced deep 3D convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2017, pp. 2278–2288.
[20] M. S. S. Kumar, “Lightweight facial expression recognition network for mobile devices,” Electronics, vol. 12, no. 12, p. 2707, Jun. 2023.
[21] S. Happy and A. Routray, “Automatic facial expression recognition using features of salient facial patches,” IEEE Trans. Affect. Comput., vol. 6, no. 1, pp. 1–12, Jan.–Mar. 2015.
[22] N. Kumar and A. Srivastava, “Enhanced Eye Disease Detection Using Advanced CNNs and Optimization Techniques,” 12th International Conference on Emerging Trends in Engineering Technology - Signal and Information Processing (ICETET - SIP), Nagpur, India, 2025, pp. 1- 7, doi: 10.1109/ICETETSIP64213.2025.11156899, 2025
[23] Winkle, S. Sharma and A. Srivastava, “A Computationally Efficient Neural Network for Sarcasm Detection in News Headlines,” International Conference on Computing Technologies Data Communication (ICCTDC), HASSAN, India, 2025, pp. 1-7, doi: 10.1109/ICCTDC64446.2025.11157969, 2025.
[24] Ravina and A. Srivastava, “Short-Term Stock Price Prediction using ANN and LSTM: A Comparative Study on Indian Tech Companies,” Grenze International Journal of Engineering and Technology, Grenze ID: 01.GIJET.11.2.295_25, June issue, 2025.