International Journal of AI, BigData, Computational and Management Studies
Noble Scholar Research Group | Volume 6, Issue 2, PP 01-10, 2025
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V6I2P101

Original Article

Edge-Optimized Facial Emotion Recognition: A High-Performance Hybrid Mobilenetv2-Vit Model

Susmith Barigidad
Lam Research USA.

Received On: 15/02/2025 ***Revised On: 08/03/2025*** ***Accepted On: 21/03/2025*** ***Published On: 03/04/2025***

Abstract - *Computer vision applications span various fields including healthcare, security, autonomous vehicles, and augmented reality, enabling machines to interpret and analyze visual data. Facial Emotion Recognition (FER) is a subclass of healthcare applications that leverages computer vision to analyze and interpret human emotions from facial expressions. Facial emotion recognition also plays a vital role in human-computer interaction, with applications in security, and affective computing. This study suggests a deep learning (DL) based hybrid model integrating MobileNetV2 for efficient feature extraction and a Vision Transformer (ViT) for capturing global facial dependencies. The dataset obtained from Kaggle is used for training, which is then preprocessed and augmented. The trained model is deployed on a smartphone as an edge device, enabling real-time emotion recognition with improved privacy, low latency, and minimal computational overhead. During testing, facial images captured by the smartphone are preprocessed using the Haar Cascade algorithm before being fed into the model for classification. Performance evaluation using accuracy, recall, precision and F1-score demonstrates a high classification accuracy of 98.51%, confirming the model's effectiveness. The proposed approach enhances on-device FER capabilities, making it a promising solution for emotion-aware applications in mobile healthcare and intelligent human-computer interactions.*

Keywords: *Computer vision application, MobileNetV2, Facial Emotion Recognition, Vision Transformer, Haar Cascade algorithm*

1. Introduction

Computer vision is revolutionizing various industries by enabling machines to interpret, analyze, and process visual data. In autonomous driving, computer vision plays a crucial role in object detection, lane tracking, and pedestrian recognition, enabling self-driving vehicles to navigate safely in dynamic environments. Additionally, in retail and marketing, computer vision is utilized for customer behavior analysis, automated checkout systems, and inventory management, improving operational efficiency and customer experience. In security and surveillance, it is widely used for facial recognition, anomaly detection, and real-time threat monitoring, enhancing public safety and automated access control systems [1]. With the increasing adoption of edge computing, deploying computer vision models on smartphones, IoT devices, and embedded systems ensures faster processing, reduced dependence on cloud infrastructure, and improved data privacy, making it a key technology for real-time applications in various domains.

FER is crucial for enhancing human-computer interaction, enabling applications such as virtual assistants, sentiment analysis, and adaptive user interfaces [2]. In healthcare, FER aids in mental health assessment, early diagnosis of psychological disorders, and improving patient care through emotion-aware monitoring systems. Additionally, FER plays a significant role in security and surveillance, enabling intelligent authentication systems and real-time threat detection based on emotional cues. The rapid advancement of DL has revolutionized computer vision applications, enabling intelligent systems to analyze and interpret visual data with high accuracy. However, deploying DL models in real-world scenarios, particularly on edge devices, presents challenges due to computational constraints, latency requirements, and energy efficiency. Traditional DL architectures rely on cloud-based processing, which introduce communication delays and privacy concerns [3]. To address these issues, edge-based computer vision solutions have gained traction, allowing models to process data locally on resource-constrained devices while maintaining efficiency and security.

Conventional FER models typically employ convolutional neural networks (CNNs) to extract spatial and temporal features. However, these models struggle with generalization in diverse environments and require extensive computational resources. Recent advances in transformer-based architectures, such as Vision Transformers (ViTs), have demonstrated superior feature extraction capabilities.

This study proposes a hybrid DL model that integrates MobileNet V2 and ViT for edge-based FER. The contributions of this work are as follows:

- Development of a hybrid DL model for robust facial emotion recognition.

- Optimization for edge-based deployment, ensuring low latency and efficient resource utilization and evaluate the performance of the model.

2. Related Works

Zhang *et al.* [4] proposed a FER method using CNN combined with image edge detection to eliminate the need for explicit feature extraction. The FER-2013 dataset, mixed with LFW data, was used in the work. The method normalized facial images and extracted edge information during convolution, preserving texture details. Max pooling was used for dimensionality reduction, and classification was performed using a Softmax classifier. The proposed method achieved an 88.56% recognition rate; however, the network complexity was the major limitation. Yang *et al.* [5] presented a lightweight, edge computing-based facial action unit (AU) detection system for real-time emotion recognition.

The method utilized optimized algorithms on Raspberry Pi to process raw image data locally, reducing network overhead and improving efficiency. A distributed system was designed, separating front-end and back-end processing to minimize delay. The AU-based recognition algorithm was deployed in Docker containers for efficient execution. Results showed that response time was one-third of ordinary computers, with accuracy up to 30% higher. The approach significantly reduced network bandwidth consumption and costs compared to cloud-based methods.

Chen *et al.* [6] proposed a deep CNN-based method using edge computing to address class imbalance and overfitting in expression databases. The dataset consisted of facial expression images enhanced using GAN-based augmentation framework. Experimental results demonstrated that the method achieved a higher recognition rate and improved classification accuracy. However, constraints in data augmentation using neutral expressions as the source domain limited the performance of the model. Hossain *et al.*

[7] proposed a privacy-preserved automatic FER system, where IoT devices captured facial images and speech signals. A multi-secret sharing scheme was used to distribute the signals across edge clouds for preprocessing before transmission to the core cloud. A pre-trained CNN extracted deep-learned features, which were fused using deep sparse autoencoders (DSAE) to introduce non-linearity, followed by an SVM classifier. The system was evaluated on the eNTERFACE'05 and RML databases, achieving recognition accuracies of 87.6% and 82.3%, respectively.

Wu *et al.* [8] proposed an Edge-AI-driven framework for efficient and accurate FER on resource-constrained edge devices. To address challenges such as pose variations, occlusion, illumination, and scale changes, two attention mechanisms were introduced namely Scalable Frequency Pooling (SFP) and Arbitrary-oriented Spatial Pooling (ASP). The ASP module captured spatial information in multiple directions, while the SFP module operated in the wavelet frequency domain to enhance feature representation. The framework was evaluated on FER2013, RaFD, and SFEW datasets, demonstrating improved accuracy and computational efficiency.

Partial cloud offloading was implemented to balance inference speed and accuracy. However, the system's performance relied on cloud connectivity, but lack of optimization for real-time applications. Pascual *et al.* [9] proposed Light-FER system designed for resource-constrained edge devices. The model was based on compression techniques, including pruning to remove less important connections and quantization to half-precision for reduced memory consumption. The system was evaluated on the FER2013 dataset. DL compilers, such as TensorRT, were utilized to enhance inference speed. However, the face detector struggled with steep-angle poses.

Makhmudkhujaev *et al.* [10] proposed the Local Prominent Directional Pattern (LPDP), an edge-based descriptor for FER to address issues like noise and positional variations. LPDP analyzed the statistical information of a pixel's neighborhood to extract significant edges, ensuring robustness against distortions while avoiding noisy edges. Extensive experiments were conducted on well-known facial expression datasets. However, the study did not explore its performance across different lighting conditions or real-time applications.

Ajay *et al.* [11] proposed a real-time facial emotion recognition system for passenger safety using Local Binary Pattern (LBP) and a Binary Neural Network (BNN) feature, implemented on an FPGA. The system utilized the Viola-Jones algorithm for facial detection, followed by LBP feature extraction and classification using a quantized BNN. The model was trained on the FER-2013 dataset and deployed on FPGA, significantly improving inference speed compared to software-based implementations. The method classified emotions into six states and demonstrated superior performance over CNN/BNN models without preprocessing.

Xu *et al.* [12] proposed a lightweight CNN-based neural network for FER on edge devices by enhancing the Visual Geometry Group 19 (VGG-19) model with residual learning. Each block's input was added to its output to improve feature propagation. Model compression techniques, including pruning and post-training quantization, were applied to reduce size while maintaining accuracy. The model achieved better accuracy compared to mainstream lightweight models. However, the study did not fully address inner-class classification challenges or further parameter reduction while preserving accuracy.

Pathak *et al.* [13] proposed a low-cost IoT edge computing system combined with a multi-headed 1D-CNN model for real-time monitoring and classification of baby facial expressions. A dataset of 600 images (200 per category) was used. The method involved face detection, cropping, feature extraction using a deep neural network, and training a 1D-CNN model on 128-dimensional embeddings. The optimized model was deployed on an edge device and operated as a REST API web service. However, memory and computational constraints of the edge device limited model size, and prolonged offline storage required periodic cloud synchronization.

Wang *et al.* [14] proposed a DL model for multimodal emotion recognition by fusing EEG signals and facial expressions. DEAP and MAHNOB-HCI datasets were used by the pre-trained CNN for facial feature extraction, with an attention mechanism enhancing crucial expression frames. CNNs were also applied to EEG signals using local and global convolution kernels to capture spatial features from different brain regions. The fused features were classified to predict valence and arousal labels. The model achieved 96.63% and 97.15% accuracy for valence and arousal on DEAP, and 96.69% and 96.26% on MAHNOB-HCI. Chaudhari *et al.* [15] investigated FER using a fine-tuned ViT and compared its performance with ResNet-18. The researchers merged three datasets after addressing class imbalances. The dataset was split into training, validation, and testing sets. The study highlighted the potential of ViT models for FER but noted limitation, including high computational costs.

Umer *et al.* [16] proposed a FER system using a CNN and data augmentation techniques using CK+ and GENKI-4k database. The system consisted of four components namely face detection, DL-based feature extraction, data augmentation, and fine-tuning of the trained CNN model. Facial images were converted to grayscale to improve processing speed. Data augmentation techniques were applied to enhance learning parameters and reduce overfitting. The method attained an accuracy of 94.67 using GENKI-4k database.

2.1 Research Gap

Despite significant advancements in FER using deep learning, several challenges remain unaddressed. Existing studies often face network complexity issues, limiting their deployment on resource-constrained edge devices. While some approaches reduce bandwidth consumption and costs compared to cloud-based methods, their reliance on cloud connectivity hinders real-time performance. Additionally, constraints in data augmentation and the use of neutral expressions as the source domain affect model generalization [6]. Many models struggle with facial detection under steep-angle poses and varying lighting conditions, impacting robustness. Furthermore, inner-class classification challenges and the need for parameter reduction while maintaining accuracy remain unresolved. Memory and computational limitations of edge devices further restrict model size, necessitating periodic cloud synchronization for prolonged offline storage.

3. Materials And Methods

The proposed FER system leverages a hybrid DL model combining MobileNetV2 for lightweight feature extraction and a ViT for capturing global facial dependencies. The dataset obtained from Kaggle is used for training. Preprocessing and augmentations are performed on the dataset to enhance feature learning. After training, the model is deployed on a smartphone as an edge device, ensuring real-time emotion detection with minimal computational overhead. During testing, the smartphone captures facial images, applies Haar Cascade-based face detection, and processes them through the trained model for emotion classification. The block schematic of the proposed method is given in Figure 1.

3.1 Dataset Description

The dataset used in this study is sourced from the Kaggle repository and consists of images representing seven basic human emotions namely Fear, Disgust, Anger, Happiness, Sadness, Contempt, and Surprise [17]. Each image is preprocessed to 48x48 pixel grayscale format. The dataset serves as a benchmark for FER models, supporting DL-based classification tasks. The images capture various expressions under different lighting conditions, facial orientations, and skin tones, making the dataset suitable for training robust emotion recognition models. Sample images from the dataset are shown in Figure 2.

3.2 Data preprocessing and augmentation

Several preprocessing steps are applied to improve the quality of images for the FER model. Image normalization is performed by scaling pixel values to a range of [0,1] or [-1,1], ensuring stable model training. To improve model generalization and robustness, data augmentation methods, namely zooming, random rotations, horizontal flipping, brightness adjustments, and slight translations are applied. These augmentations help mitigate overfitting by introducing variations that mimic real-world conditions, allowing the model to learn diverse facial expressions across different perspectives and environments. Once the image is captured, preprocessing occurs on the smartphone to prepare the raw data for emotion recognition. The first step in preprocessing is face detection, where algorithms like Haar Cascades, detect the face within the captured frame. Haar Cascades is a feature-based face detection algorithm that uses Haar-like features to identify patterns such as edges and textures in an image. It employs an integral image for fast computation and an AdaBoost classifier to select the most relevant features. The detection process follows a cascade structure, where simple classifiers eliminate non-face regions early, while complex ones refine detection. The feature F is computed as per Equation 1.

$$F = \sum(Pixel\ intensity\ in\ white\ region) - \sum(pixel\ intensity\ in\ black\ region) \quad (1)$$

The integral image is used for rapid computation as per Equation 2.

$$I(x,y) = I(x-1,y) + I(x,y-1) - I(x-1,y-1) + P(x,y) \quad (2)$$

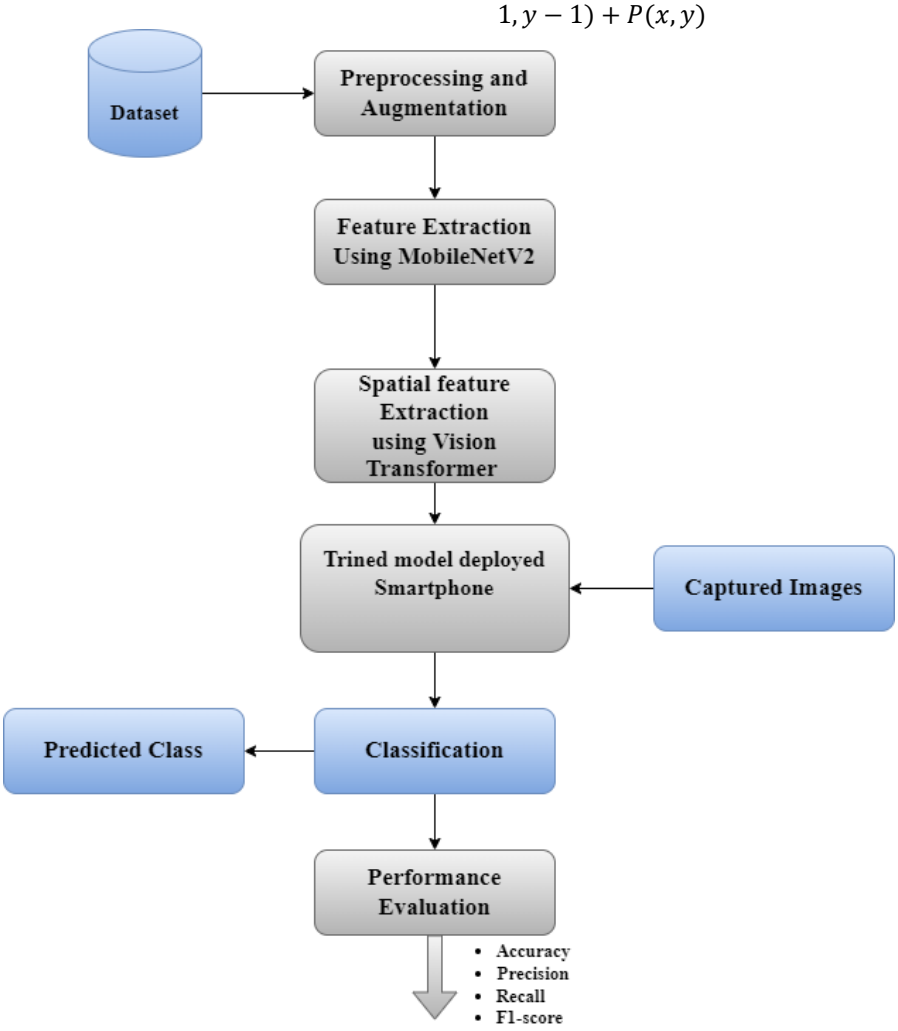

Figure 1. Block schematic of the suggested system

Figure 2. Sample images in the dataset

Where $I(x, y)$ is the integral image at point (x, y) and $P(x, y)$ is the original pixel intensity. This hierarchical approach ensures real-time detection, making it efficient for edge-based facial emotion recognition on smartphones. This ensures that only the face is processed, isolating it from the background or other objects. Following face detection, face alignment takes place, where facial landmarks such as the nose, eyes, and mouth are identified and used to align the face to a standard orientation, improving the consistency of the analysis. The face is then cropped from the image, and the resulting region is resized to a fixed dimension, typically 224x224 pixels, suitable for feeding into the emotion recognition model. Lastly, the pixel values are normalized to a standard to help the model perform more efficiently. The detected faces are given to the proposed hybrid model for facial emotion recognition.

3.3 Model Development

To achieve efficient and accurate FER on edge devices, the proposed model combines MobileNetV2 for feature extraction with a ViT for enhanced attention-based feature learning. MobileNetV2 transforms input facial images into a set of low-dimensional, high-level feature representations. These extracted features are then passed to the ViT for further processing.

3.3.1 MobileNet V2

MobileNetV2 is a lightweight CNN designed for edge devices. It employs depthwise separable convolutions to reduce computational complexity while maintaining high accuracy [18]. Figure 3 illustrates the basic architecture of MobileNetV2. The MobileNetV2 architecture consists of three main components. Depthwise separable convolutions reduce computational cost by splitting the convolution into a depthwise and a pointwise operation. Inverted residual blocks allows efficient feature learning by connecting input and output through a lightweight bottleneck structure.

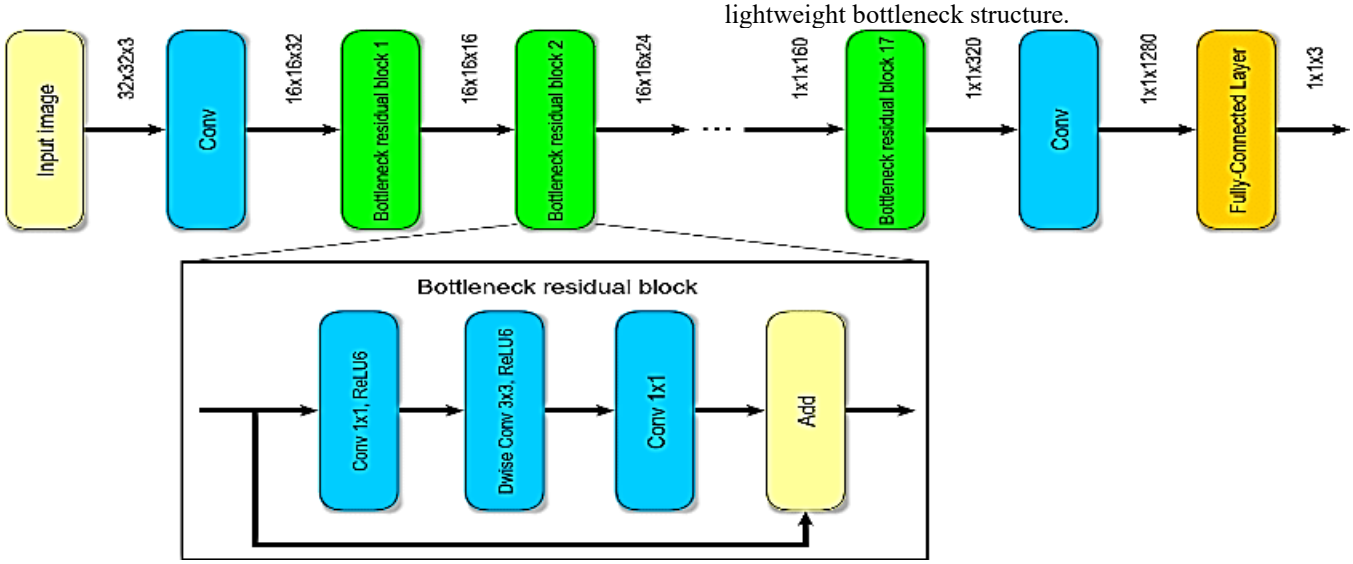

Figure 3. Basic architecture MobileNetV2

Linear bottleneck layers prevent loss of information by maintaining low-dimensional feature maps. Mathematically, the output of a depthwise separable convolution is computed as per Equation 3.

$$Y = \sigma(W_d * X) + \sigma(W_p * Y_d) \quad (3)$$

Where the depthwise convolution kernel is denoted by W_d, W_p denotes the pointwise convolution kernel, input feature map is denoted by X , Y_d is the depthwise convolved output, $\sigma(\cdot)$ represents the activation function. The inverted residual block is defined as per Equation 4 to Equation 6.

$$Z = \sigma(W_{1\times1}^{expand} \cdot X) \quad (4)$$

$$Y_d = \sigma(W_d * Z) \quad (5)$$

$$Y = W_{1\times1}^{reduce} * Y_d) \quad (6)$$

Where $W_{1\times1}^{expand}$ is the expansion layer, W_d is the depthwise convolution, $W_{1\times1}^{reduce}$ is the reduction layer.

3.3.2 Vision Transformer

The ViT employs a self-attention mechanism to capture global dependencies in images [19]. Figure 4 depicts the general architecture of the ViT, which processes images by dividing them into patches, encoding them with positional embeddings.

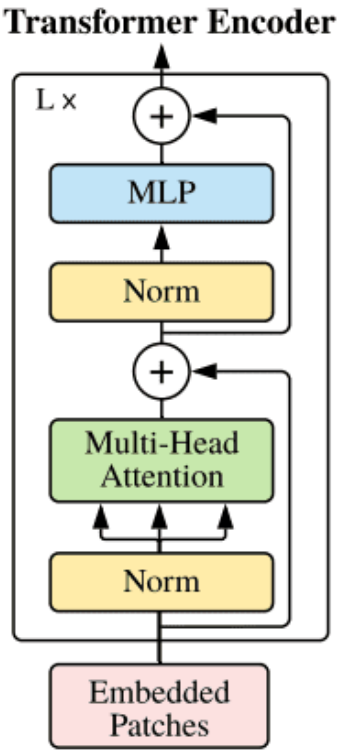

Figure 4. General architecture of Vision transformer

ViT divides an image into fixed-sized patch and then embeds them into a lower-dimensional space using a trainable linear projection as per Equation 7.

$$z_0 = [x_1E;\ x_2E; \dots; x_NE] + E_{pos} \quad (7)$$

Where x_1 represents the flattened image patches, E is the learnable embedding matrix, E_{pos} is the positional encoding that retains spatial information. Each patch embedding is given to transformer encoder. Self-attention allows the model to focus on different facial regions based on their importance for emotion recognition. The Multi-Head Self-Attention (MHSA) mechanism is defined as per Equation 8.

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (8)$$

Where $Q,V,K,$ are the query, value and key matrices obtained by projecting the input feature maps, d_k is the dimension of key vectors for scaling. MHSA enables ViT to learn relationships between different facial regions without losing contextual information. Each transformer encoder block also contains a position-wise Feed-Forward Network (FFN), as per Equation 9.

$$FFN(x) = \sigma(xW_1 + b_1)W_2 + b_2 \quad (9)$$

Where W_1 and W_2 are weight matrices, b_1 and b_2 are biases, σ is a non-linear activation function. Each encoder layer applies layer normalization (LN) to stabilize training as per Equation 10.

$$LN(x) = \frac{x-\mu}{\sigma+\epsilon} \quad (10)$$

Where σ and μ are the standard deviation and mean of the input.

3.3.3 Proposed Hybrid Model

The model incorporated MobileNetV2 and ViT by fusing their feature representations before classification. The CNN captures local spatial details, while the ViT captures model global dependencies. Mathematically, the feature fusion is defined as per Equation 11.

$$F_{Hybrid} = \alpha F_{MobileNetV2} + (1-\alpha)F_{ViT} \quad (11)$$

Where $F_{MobileNetV2}$ represents extracted features from CNN, F_{ViT} represents global transformer-based features, α is a learnable parameter controlling feature importance. The fused features are processed by fully connected layers a per Equation 12.

$$Y = softmax(W_f.F_{hybrid} + b) \quad (12)$$

Where weight matrix is denoted by W_f, b is the bias, softmax activation converts the final output into

probability scores across multiple emotion classes. Hyperparameters such as learning rate, number of epochs, and batch size are critical in training the models, as they influence convergence speed, prevent overfitting, and ensure optimal model performance in FER.

Table 1: Hyperparameters in the proposed model

Parameters	Values
Optimizer	Adam
Learning rate	0.0001
Loss	Categorical Cross-entropy
Activation function	ReLU
Epoch	50

4. Result And Discussion

To assess the effectiveness of the proposed Hybrid DL Model for Edge-Based Facial Emotion Recognition, standard evaluation metrics such as accuracy, precision, recall, and F1-score are used. Accuracy measures the overall correctness of the model by computing the ratio of correctly classified samples to the total number of samples. It is defined as per Equation 13.

$$Accuracy = \frac{T_P+T_N}{T_P+T_N+F_P+F_N} \quad (13)$$

Where, T_P indicates true positive, T_N denotes true negative, F_N signifies false negative, and F_P indicates false positive. Precision, also called Positive Predictive Value, measures the proportion of correctly predicted positive instances among all instances classified as positive. It is given as per Equation 14.

$$Precision = \frac{T_P}{T_P+F_P} \quad (14)$$

Recall measures the model's ability to correctly identify all relevant instances of a given class. It is defined as per Equation 15.

$$Recall = \frac{T_P}{T_P+F_N} \quad (15)$$

F1-score is the harmonic mean of precision and recall, balancing both metrics in cases where there is an imbalance in class distribution. It is computed as per Equation 16.

$$F1-score = 2 * \frac{Precision \times Recall}{Precision+Recall} \quad (16)$$

A high F1-score signifies that the model maintains a good trade-off between precision and recall, meaning it is both accurate and sensitive in detecting emotions. When training and evaluating DL models for edge-based facial emotion recognition, two key visualization tools are used to assess performance: the accuracy plot as shown in figure 5 and the loss plot as given in Figure 6. These plots help in monitoring the model's learning behavior and detecting potential issues like overfitting or underfitting. In the initial epoch, the model achieved a training accuracy of 85.78% and a validation accuracy of 88.11%. As training progressed, it attained a training accuracy of 95% by the 33rd epoch, while the validation accuracy remained the same. Finally, in the last epoch, the model reached a training accuracy of 96.01% with an improved validation accuracy of 99%.

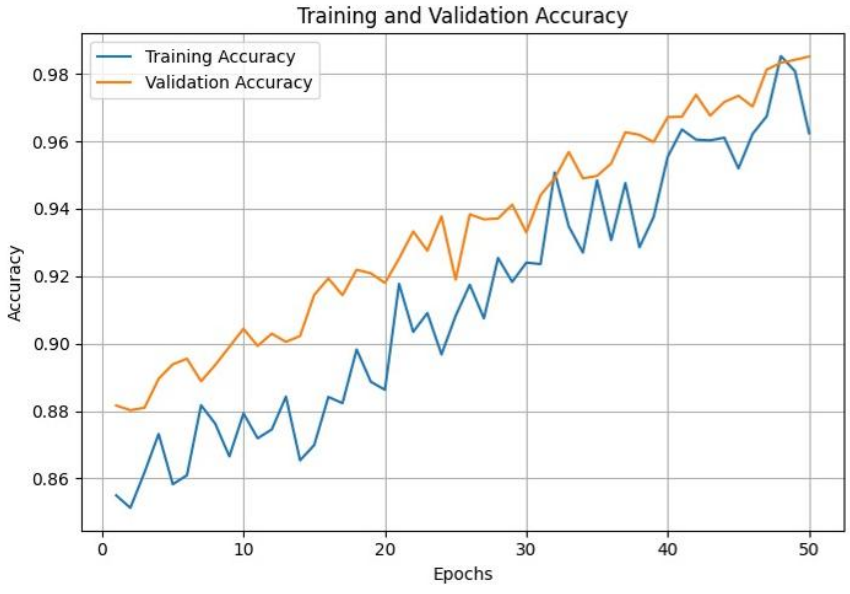

Figure 5. Accuracy plot of the suggested system

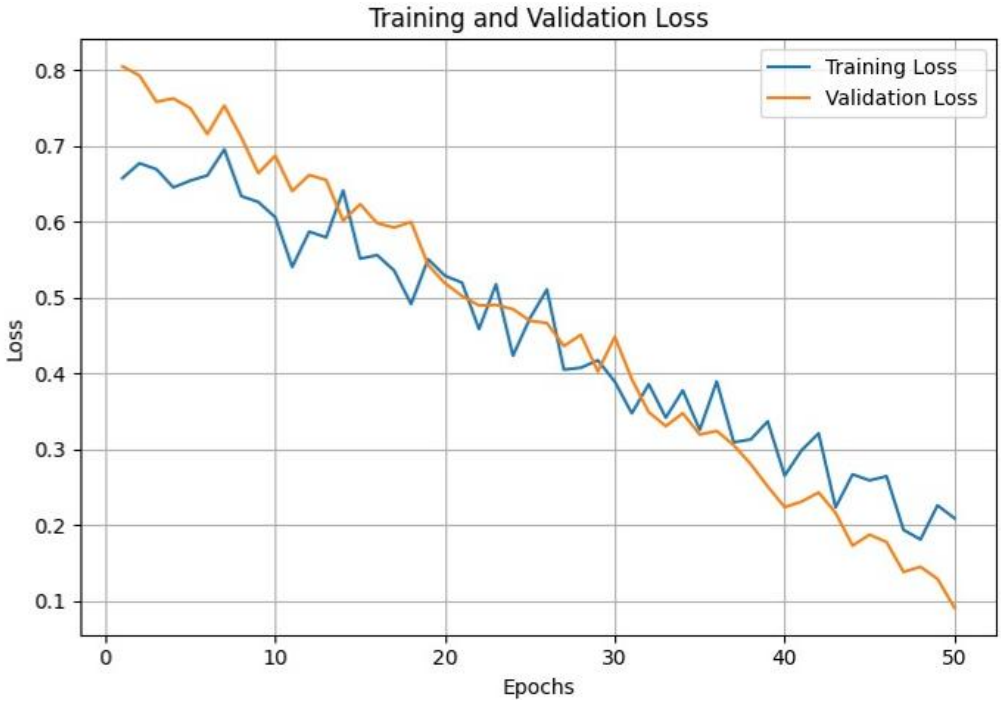

Figure 6. Loss plot of the suggested system

The model initially started with a training loss of 0.675 and a validation loss of 0.812. As training progressed, the training loss decreased to 0.392, while the validation loss showed a slight increase to 0.467. Finally, the model further reduced the training loss to 0.213, achieving a validation loss of 0.101. The confusion matrix as given in Figure 7 is a performance evaluation tool that visualizes the model's classification results. Figure 8 presents the classification report of the suggested system, showcasing key performance metrics such as accuracy, precision, recall, and F1-score, which validate the model's effectiveness in facial emotion recognition.

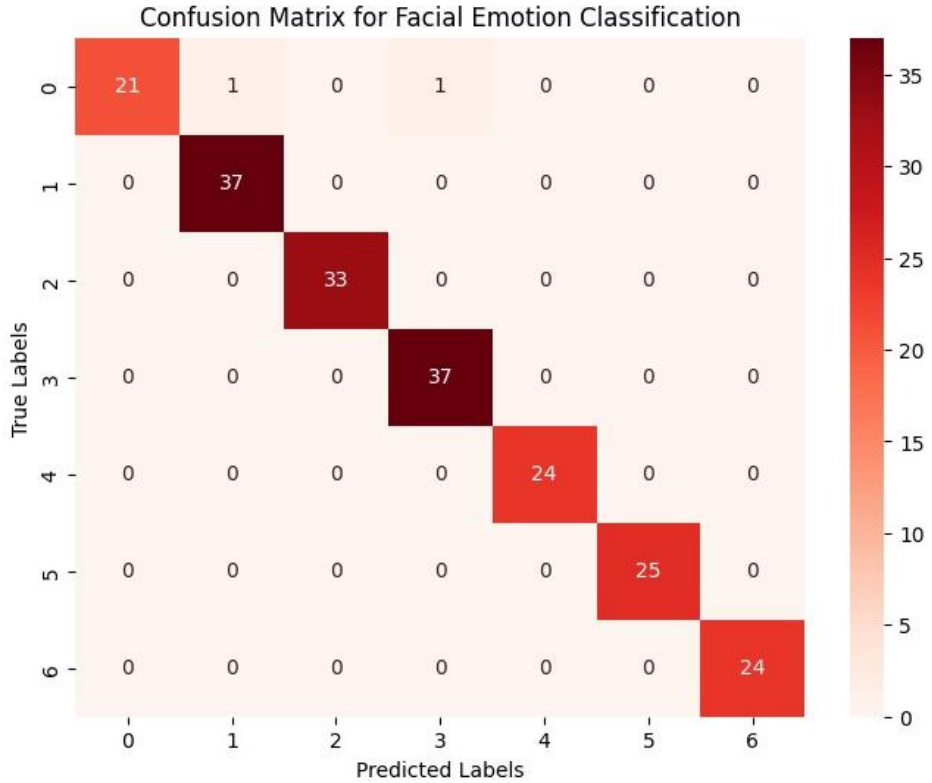

Figure 7. Confusion matrix of the suggested system

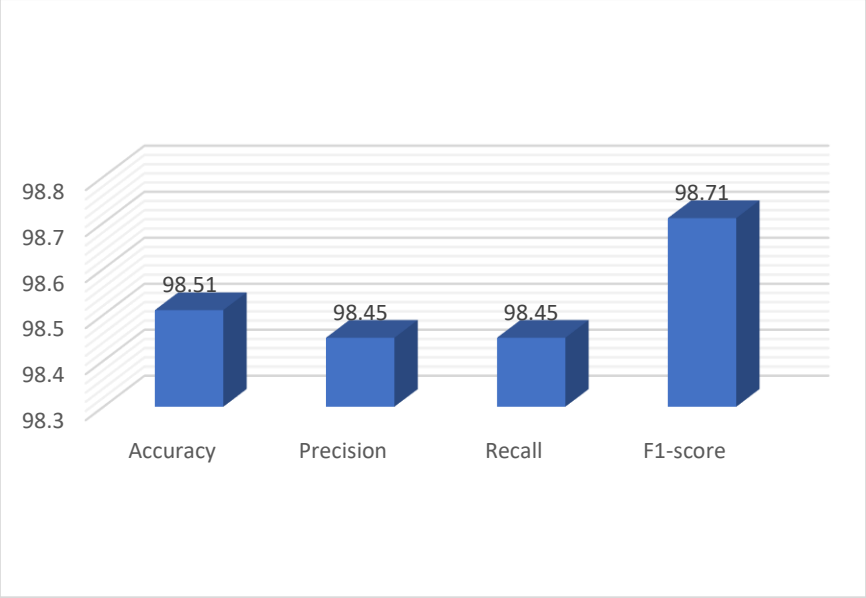

Figure 8. Classification report of the suggested system

The model achieved an accuracy of 98.51%, indicating a high overall correctness in classification, while the precision of 98.45% reflects its ability to minimize false positives. Additionally, the recall of 98.99% shows the model's effectiveness in capturing true positives, and the F1-score of 98.71% demonstrates a strong balance between precision and recall, ensuring reliable emotion recognition. Figure 9 illustrates the prediction output of the suggested system, displaying the detected facial expressions along with their corresponding emotion.

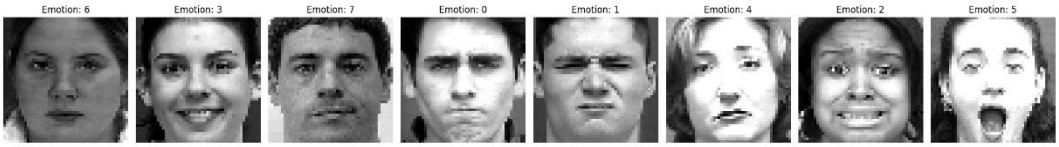

Figure 9. Prediction output of the suggested system

5. Conclusion

Computer vision applications, particularly in Facial Emotion Recognition (FER), play a crucial role in various domains such as healthcare, human-computer interaction, mental health monitoring, and security surveillance by enabling automated analysis of human emotions from facial expressions. The relevance of FER lies in its ability to enhance personalized user experiences, assist in early mental health diagnosis, and improve intelligent systems by enabling emotion-aware responses, making it a vital tool for modern AI-driven applications. This study presents a hybrid DL model integrating MobileNetV2 and a ViT for efficient and accurate Facial Emotion Recognition (FER) on edge devices. The proposed model leverages MobileNetV2 for lightweight yet powerful feature extraction, while the ViT captures global facial dependencies, enhancing the model's ability to recognize subtle emotional variations. To enable real-time FER, the trained model was deployed on a smartphone as an edge device, ensuring low latency, privacy preservation, and reduced computational overhead compared to cloud-based solutions. During testing, facial images captured by the smartphone underwent preprocessing before being classified by the model. The system was evaluated using standard performance metrics, achieving a high classification accuracy of 98.51%, along with superior precision, recall, and F1-score. These results validate the effectiveness of the proposed approach in recognizing facial emotions with minimal computational requirements, making it ideal for real-world applications.

Reference

[1] Rezaee, K., Rezakhani, S. M., Khosravi, M. R., & Moghimi, M. K. (2024). A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing, 28(1), 135-151.

[2] Zadeh, E. K., & Alaeifard, M. (2023). Adaptive Virtual Assistant Interaction through Real-Time Speech Emotion Analysis Using Hybrid Deep Learning Models and Contextual Awareness. International Journal of Advanced Human Computer Interaction, 1(1), 1-15.

[3] Wu, H., Li, X., & Deng, Y. (2020). Deep learning-driven wireless communication for edge-cloud computing: opportunities and challenges. Journal of Cloud Computing, 9(1), 21.

[4] Zhang, H., Jolfaei, A., & Alazab, M. (2019). A face emotion recognition method using convolutional

neural network and image edge computing. IEEE Access, 7, 159081-159089.
[5] Yang, J., Qian, T., Zhang, F., & Khan, S. U. (2021). Real-time facial expression recognition based on edge computing. IEEE Access, 9, 76178-76190.
[6] Chen, A., Xing, H., & Wang, F. (2020). A facial expression recognition method using deep convolutional neural networks based on edge computing. Ieee Access, 8, 49741-49751.
[7] Hossain, M. S., & Muhammad, G. (2019). Emotion recognition using secure edge and cloud computing. Information Sciences, 504, 589-601.
[8] Wu, Y., Zhang, L., Gu, Z., Lu, H., & Wan, S. (2023). Edge-AI-driven framework with efficient mobile network design for facial expression recognition. ACM Transactions on Embedded Computing Systems, 22(3), 1-17.
[9] Pascual, A. M., Valverde, E. C., Kim, J. I., Jeong, J. W., Jung, Y., Kim, S. H., & Lim, W. (2022). Light-FER: a lightweight facial emotion recognition system on edge devices. Sensors, 22(23), 9524.
[10] Makhmudkhujaev, F., Abdullah-Al-Wadud, M., Iqbal, M. T. B., Ryu, B., & Chae, O. (2019). Facial expression recognition with local prominent directional pattern. Signal Processing: Image Communication, 74, 1-12.
[11] Ajay, B. S., & Rao, M. (2021, February). Binary neural network based real time emotion detection on an edge computing device to detect passenger anomaly. In 2021 34th International Conference on VLSI Design and 2021 20th International Conference on Embedded Systems (VLSID) (pp. 175-180). IEEE.
[12] Xu, G., Yin, H., & Yang, J. (2020, December). Facial expression recognition based on convolutional neural networks and edge computing. In 2020 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS) (pp. 226-232). IEEE.
[13] Pathak, R., & Singh, Y. (2020, October). Real time baby facial expression recognition using deep learning and IoT edge computing. In 2020 5th International conference on computing, communication and security (ICCCS) (pp. 1-6). IEEE.
[14] Wang, S., Qu, J., Zhang, Y., & Zhang, Y. (2023). Multimodal emotion recognition from EEG signals and facial expressions. IEEE Access, 11, 33061-33068.
[15] Chaudhari, A., Bhatt, C., Krishna, A., & Mazzeo, P. L. (2022). ViTFER: facial emotion recognition with vision transformers. Applied System Innovation, 5(4), 80.
[16] Umer, S., Rout, R. K., Pero, C., & Nappi, M. (2022). Facial expression recognition with trade-offs between data augmentation and deep learning features. Journal of Ambient Intelligence and Humanized Computing, 1-15.
[17] https://www.kaggle.com/datasets/shareef0612/ckdataset
[18] Dong, K., Zhou, C., Ruan, Y., & Li, Y. (2020, December). MobileNetV2 model for image classification. In 2020 2nd International Conference on Information Technology and Computer Application (ITCA) (pp. 476-480). IEEE.
[19] Li, J., Yan, Y., Liao, S., Yang, X., & Shao, L. (2021). Local-to-global self-attention in vision transformers. arXiv preprint arXiv:2107.04735.