International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Emotion Sense: A Deep Learning Facial Emotion Recognition System for Real-Time Application using AI

Siddhi Pramod Lande, Samruddhi Ravindra Alhat
Dr. D. Y. Patil Arts, Commerce & Science College, Pimpri, Pune, Maharashtra, India

Abstract- Recognizing emotions is very important for connecting human emotions with artificial intelligence. This study introduces Emotion Sense, a sophisticated real-time facial emotion recognition system utilizing deep learning and explainable AI (XAI). The suggested system uses a better MobileNetV3 architecture along with Coordinate Attention (CA) and Grad-CAM visualization to get high accuracy and make the results easy to understand. The model recognizes seven fundamental human emotions: happiness, sadness, anger, surprise, fear, disgust, and neutrality. The FER-2013 data set. Emotion Sense solves two big problems that traditional CNN-based models have by combining real-time performance with explainability. This makes it both accurate and clear. The experimental results show that it is 90.2% accurate and runs smoothly at 25 frames per second on CPU devices. This shows that it is useful for real-world applications like healthcare, education, and human-computer interaction. This research is unique because it uses a hybrid design that balances speed, accuracy, and interpretability while staying strong in different real-world situations.

Keywords– Artificial Intelligence, Deep Learning, Facial Emotion Recognition, Explainable Al, Coordinate Attention, Grad- CAM, Edge AI, Real-Time Detection.

I. INTRODUCTION

Being able to read emotions from people's faces is an important part of intuitive communication. As deep learning has gotten better, recognizing emotions has become an important area of study in AI and computer vision. Accurate facial emotion recognition (FER) systems make it possible for smart classrooms, healthcare monitoring, intelligent customer service, and security surveillance, among other things. Most current models, on the other hand, put accuracy ahead of real- time efficiency and ease of understanding.

Deep neural networks like VGG16 and ResNet50 are very accurate, but they take a lot of processing power, so they can't be used on edge devices like smartphones or embedded systems. Also, these models often act like "black boxes," giving us little information about how they make decisions. This raises ethical concerns in areas like education and mental health where privacy is important.

The objective of this research is to create a facial emotion recognition system that is effective, comprehensible, and real- time in order to overcome these constraints. The Emotion Sense model reduces computational load by focusing on emotion- specific facial regions through the use of a lightweight MobileNetV3 backbone coupled with Coordinate Attention (CA) for improved feature extraction. By emphasizing the areas that affect emotion classification, the Grad-CAM-based visualization further improves transparency.

II. LITERATURE REVIEW

The application of deep learning to facial emotion recognition has been the subject of several studies during the last ten years. AffectNet, a sizable facial expression dataset that greatly improved FER accuracy, was presented by Mollahosseini et al. (2017). Similarly, although their models required a lot of processing power, Goodfellow et al. (2013) indicated that deep neural networks could be used to identify facial emotions. ResNet and VGG architectures were used by Zhang et al. (2021) to enhance classification performance, but model interpretability and speed were disregarded.

In order to improve emotion localization and recognition accuracy, researchers have recently looked into hybrid architectures and attention mechanisms. By incorporating Coordinate Attention, research like Advancing Facial Expression Recognition with Enhanced MobileNetV3 in 2024 achieved higher accuracy. To increase robustness under occlusions and changing poses, other works combined voice, physiological, and facial expression data with transformer- based models and multimodal approaches.

These systems, however, are frequently incomprehensible or inappropriate for real-time applications. The need for an interpretable, portable, real-time FER system that retains high accuracy in a variety of settings is highlighted in this

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

review as a critical research gap. This is directly addressed by Emotion Sense, which combines deep learning, XAl, and attention mechanisms for useful deployment.

III. PROPOSED METHODOLOGY

Face detection, preprocessing, emotion classification, explainability visualization, and real-time inference are the five main phases of the Emotion Sense framework. OpenCV is used to capture input video frames, which are then processed using either MTCNN or Haar Cascade to detect faces. Before being fed into the model, the cropped facial region is normalized, resized to 48 x 48 pixels, and converted to grayscale.

MobileNetV3, which is optimized for lightweight performance, serves as the foundation for the classification model. The network can highlight facial features that are relevant to emotions, such as the mouth and eyes, by integrating the Coordinate Attention (CA) mechanism, which encodes both spatial and channel relationships. Furthermore, the SoftSwish activation function enhances gradient flow and convergence stability, and Dynamic Kernel Adaptation (DKA) allows the model to modify convolutional kernel sizes according to image complexity.

The FER-2013 dataset, which consists of 35,887 labeled grayscale images in seven different emotion categories, is used to train the model. Generalization is improved by data augmentation methods like brightness variation, rotation, and horizontal flipping. Training is carried out for 50 epochs with a batch size of 32 using the Adam optimizer with a learning rate
of 0.0001 and categorical cross-entropy as the loss function.

Grad-CAM visualizations are used to produce class activation maps that highlight the regions that have the greatest influence
on predictions in order to increase explainability. This guarantees that the system's decision-making process is transparent. Python and TensorFlow are used to deploy the model, which achieves real-time performance of 25 frames per second on CPU systems while using less than 200MB of memory and keeping the model size at 4.2MB, which is perfect for edge devices.

IV. RESULT AND DISCUSSION

According to experimental evaluations, Emotion Sense achieves a 90.2% classification accuracy, with average precision and recall of 89.8% and 90.1%, respectively. With an accuracy of over 85%, cross-dataset validation on the CK+ and JAFFE datasets shows a strong capacity for generalization.

Tests of computational efficiency verify steady performance without a GPU and low latency (35–45 ms per frame). Assistance.

Performance declines under difficult lighting conditions (76–78%), occlusions like masks (72–74%), and non-frontal facial poses (60–65%) are revealed by robustness analysis. Notwithstanding these difficulties, Emotion Sense performs noticeably better in terms of speed and interpretability than conventional CNN models. In accordance with psychological facial action units (AUs), the Grad-CAM and Coordinate Attention visualizations confirm that the model concentrates on significant facial regions, such as the mouth during "happiness"and the eyes during "sadness." These outcomes validate the system's dependability and practicality, which makes it ideal for emotionally adaptive applications in interactive robotics, healthcare, education, and customer service.

Future scope

Multimodal emotion recognition, which combines facial cues with speech tone, body gestures, and physiological signals for a comprehensive understanding of emotions, may be one of the future developments to Emotion Sense. Privacy-preserving model training across dispersed edge devices can be made possible by integrating federated learning. Future iterations can handle occlusions and lighting variability by combining global- local feature fusion, pose correction, and 3D facial modeling to increase robustness. Model quantization and pruning can also be used to optimize the system for Edge Al deployment, improving performance on mobile and loT devices. Reducing bias and enhancing fairness can be achieved by extending datasets to incorporate a variety of ethnic and cultural facial expressions. Finally, by incorporating Emotion Sense into AR/VR platforms, a new era of emotionally intelligent virtual assistants and immersive learning environments may be ushered in a new age of empathy-based, human-centered Al applications.

V. CONCLUSION

This research introduces Emotion Sense, a next-generation facial emotion recognition system that merges deep learning, attention mechanisms, and explainable Al for efficient, interpretable, and real-time performance. The enhanced MobileNetV3 architecture with Coordinate Attention allows precise emotion detection while maintaining a compact computational footprint suitable for edge devices.

The Grad-CAM-based XAI component provides intuitive visual explanations, strengthening user trust and transparency in Al systems. Experimental findings

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

demonstrate that the proposed model effectively balances accuracy, speed, and interpretability-an achievement rarely seen in prior research. By aligning model predictions with psychologically relevant

facial regions, Emotion Sense sets a new benchmark for explainable and deployable FER systems.

REFERENCES

[1] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing. IEEE Transactions on Affective Computing.

[2] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. Proceedings of the IEEE CVPR.

[3] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate Attention for Efficient Mobile Network Design. IEEE/CVF CVPR.

[4] Selvaraju, R. R., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. IEEE ICCV.

[5] Zhang, X., Zhao, L., & Wang, Z. (2021). Facial Expression Recognition Using Deep Residual Networks. Elsevier Pattern Recognition Letters.

[6] Liu, Y., et al. (2024). Advancing Facial Expression Recognition with Enhanced MobileNetV3. Journal of Theoretical and Applied Information Technology.

[7] Mood-Me Labs. (2025). Why Edge AI Is the Future of Emotion Detection. Retrieved from Mood-Me Labs