Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 5 | May-2026
ISSN (O) 3107-6696

Facial Emotion Recognition using Ensemble Deep Learning and SVM

Saarthak Shivam[1], Dr Archana Kumar[2]

saarthakshivam04@gmail.com, profdrarchanakumar@gmail.com

[*] Scholar, Btech (AI&DS) 4th Year, [**] Department of Artificial Intelligence and Data Science, Dr. Akhilesh Das Gupta Institute of Professional Studies, New Delhi

---

Abstract - Facial Emotion Recognition (FER) is an important application of Artificial Intelligence, Computer Vision, and Deep Learning that enables machines to identify human emotions through facial expressions. Human emotions such as happiness, sadness, anger, fear, surprise, disgust, and neutrality play a major role in communication and behavioral understanding. However, accurately classifying emotions remains challenging due to variations in facial appearance, image quality, lighting conditions, and similarities between emotional expressions.

This project presents a hybrid Facial Emotion Recognition system using ensemble Deep Learning and Machine Learning techniques for accurate emotion classification. The system utilizes the FER-2013 Dataset consisting of grayscale facial images categorized into seven emotion classes. Multiple Convolutional Neural Network architectures including LeNet-5, ResNet-50, and VGG-16 are used for deep feature extraction. The extracted features are combined using ensemble learning and classified using a Support Vector Machine classifier.

The proposed CNN-SVM hybrid architecture improves feature representation, reduces overfitting, and enhances classification robustness compared to traditional single-model systems. The developed FER system has potential applications in healthcare, surveillance systems, driver monitoring, gaming, robotics, and human-computer interaction.

Index Terms: Facial Emotion Recognition, Deep Learning, Convolutional Neural Network, Ensemble Learning, Support Vector Machine, Computer Vision, FER-2013 Dataset, Image Classification, Feature Extraction, Emotion Detection

Abbreviations-

FER : Facial Emotion Recognition
CNN : Convolutional Neural Network
SVM : Support Vector Machine
AI : Artificial Intelligence
ML : Machine Learning
DL : Deep Learning
CV : Computer Vision
GPU : Graphics Processing Unit
ANN : Artificial Neural Network
ReLU : Rectified Linear Unit
CSV : Comma Separated Values
ROC : Receiver Operating Characteristic
AUC : Area Under Curve
MSE : Mean Squared Error
RMSE : Root Mean Square Error
GUI : Graphical User Interface
OpenCV : Open Source Computer Vision Library
PCA : Principal Component Analysis
VGG : Visual Geometry Group

I. Introduction

Facial expressions are one of the most important forms of non-verbal communication used by humans to express emotions such as happiness, sadness, anger, fear, surprise, disgust, and neutrality. With the rapid growth of Artificial Intelligence and Computer Vision, machines are now capable of understanding emotions through facial analysis. This field is known as Facial Emotion Recognition (FER).

Traditional emotion recognition systems often suffer from low accuracy because of image noise, lighting variations, pose differences, and similarities between emotions. Single-model approaches may fail to capture all important facial features effectively.

The proposed Facial Emotion Recognition system addresses these challenges using an ensemble-based Deep Learning architecture. The system combines LeNet-5, ResNet-50, and VGG-16 for feature extraction and uses SVM classification for final emotion prediction. The hybrid CNN-SVM approach improves feature learning, robustness, and classification performance.

1.1 Challenges

Developing an accurate Facial Emotion Recognition system involves several challenges related to image quality, feature extraction, and emotion classification.

One major challenge is the similarity between certain emotions such as fear and surprise or sad and neutral. These emotions share similar facial characteristics, making classification difficult even for Deep Learning models.

Another challenge is dataset imbalance and low-resolution facial images in the FER-2013 dataset. Variations

in lighting conditions, facial orientation, and image noise further reduce model performance.

Training multiple CNN architectures also increases computational complexity and requires GPU acceleration for efficient execution.

1.2 Need of System

Human emotions play a very important role in communication, decision-making, and behavioral understanding. In traditional human-computer interaction systems, machines are unable to naturally understand human emotions, which limits the ability of intelligent systems to interact effectively with users. This creates a strong need for automated Facial Emotion Recognition (FER) systems capable of identifying human emotions accurately through facial expressions.

With the rapid growth of Artificial Intelligence, Deep Learning, and Computer Vision technologies, emotion-aware systems are becoming increasingly important in modern applications. Existing traditional emotion recognition systems often rely on handcrafted features and single-model architectures, which usually suffer from low accuracy, poor robustness, and weak generalization capability under real-world conditions such as lighting variations, image noise, and facial pose changes.

An intelligent FER system can significantly improve human-computer interaction by allowing machines to respond according to user emotions and behavior.

The proposed ensemble-based FER system addresses these limitations by combining multiple CNN architectures with SVM classification to improve feature extraction, robustness, and overall emotion classification accuracy. The system provides a more reliable and efficient approach for recognizing human emotions from facial images

1.3 Applications

1. Healthcare Systems: FER systems can help monitor patient emotions, stress levels, and mental health conditions.
2. Driver Monitoring Systems: Emotion detection can identify driver fatigue, stress, or distraction to improve road safety.
3. Smart Surveillance: FER systems can analyze suspicious or unusual emotional behavior in public places.
4. Human-Computer Interaction: Emotion-aware systems improve communication between humans and machines.
5. Gaming and Robotics: Emotion recognition enables intelligent interaction in gaming environments and robotic systems.
6. Online Education: FER systems can analyze student engagement and emotional response during virtual learning.

II. LITERATURE REVIEW

[1] LeCun et al. (1998), in their work on LeNet-5, introduced one of the earliest Convolutional Neural Network architectures for image classification tasks. The architecture demonstrated that CNNs could automatically learn important image features such as edges, textures, and patterns directly from raw pixel data without relying on handcrafted feature extraction techniques. This work became the foundation for modern Deep Learning-based image recognition systems and inspired the use of CNNs in Facial Emotion Recognition applications.

[2] Simonyan and Zisserman (2015), in their research on VGG-16, proposed a deep convolutional network architecture capable of learning highly detailed spatial image representations. The use of multiple convolutional layers with small filter sizes significantly improved image classification performance. VGG-16 became widely used in facial recognition, object detection, and emotion classification tasks because of its strong feature extraction capability and ability to capture detailed facial structures.

[3] He et al. (2016), in their paper on ResNet-50, introduced residual learning to solve the vanishing gradient problem in deep neural networks. Residual connections enabled the successful training of deeper CNN architectures by improving gradient flow during backpropagation. ResNet models achieved remarkable performance in image classification and computer vision applications and became highly effective for extracting deep semantic facial features in FER systems.

[4] Cortes and Vapnik (1995) developed the Support Vector Machine classifier, one of the most widely used machine learning algorithms for classification problems. SVM performs effectively in high-dimensional feature spaces and provides strong decision boundaries between classes. Several recent FER research studies have combined CNN-based feature extraction with SVM classification to improve classification accuracy, reduce overfitting, and enhance generalization capability compared to standalone CNN models.

[5] The FER-2013 Dataset is one of the most commonly used benchmark datasets for emotion recognition research. The dataset contains grayscale facial images categorized into seven emotion classes: angry, disgust, fear, happy, sad, surprise, and neutral. Researchers widely use FER-2013 for training and evaluating Deep Learning models because it provides large-scale labeled facial expression data suitable for benchmarking CNN architectures and emotion classification systems.

[6] Recent research studies in Facial Emotion Recognition indicate that ensemble learning approaches provide better performance compared to single-model systems. Different CNN architectures learn different facial representations and emotional patterns from images. Combining multiple CNN models improves feature diversity, robustness, and classification accuracy. Ensemble-based FER systems have shown improved capability in handling image noise, lighting variations, facial pose changes, and similarities between emotional expressions.

III. METHODOLOGY

3.1 Data Collection and Knowledge Base

Dataset: The proposed Facial Emotion Recognition (FER) system uses the FER-2013 Dataset consisting of grayscale facial images of size 48 × 48 pixels. The dataset contains seven human emotion categories:

- Angry
- Disgust
- Fear
- Happy
- Sad
- Surprise
- Neutral

The dataset was loaded in CSV format from Google Drive into the Google Colab environment. Each image in the dataset is represented through pixel intensity values, which were converted into image matrices during preprocessing. FER-2013 is widely used as a benchmark dataset for training and evaluating Deep Learning-based emotion recognition systems because it contains a large number of labeled facial expression samples with diverse emotional patterns.

The dataset was divided into training and testing sets using an 80-20 split ratio. The training dataset was further prepared for individual CNN model training and feature extraction. The FER-2013 dataset served as the primary data source for model learning, testing, and performance evaluation.

3.2 Data Preprocessing and Pipeline

The data preprocessing pipeline is one of the most important stages of the proposed FER system because raw image data cannot be directly used for Deep Learning model training. The preprocessing stage improves model convergence, training stability, and overall classification performance.

The preprocessing process involved:

- image reshaping,
- normalization,
- label encoding,
- and train-test splitting.

During image reshaping, pixel values from the CSV dataset were converted into 48 × 48 grayscale image matrices suitable for CNN processing. Pixel intensities were normalized from the range 0–255 to 0–1 to improve gradient optimization and reduce training instability.

Emotion labels were encoded into categorical numerical format for classification. After preprocessing, the dataset was divided into training and testing sets. The processed images were then passed into the CNN architectures for feature extraction and training.

3.3 System Architecture

Architecture: The proposed FER system follows an ensemble-based Deep Learning architecture consisting of multiple stages for emotion classification.

The overall workflow of the system is organized as follows:

1. Input Facial Images
2. Data Preprocessing
3. CNN Feature Extraction
4. Feature Concatenation
5. SVM Classification
6. Emotion Prediction Output

The system utilizes multiple CNN architectures including LeNet-5, ResNet-50, and VGG-16 for feature extraction. Each CNN model learns different emotional patterns and facial representations from the input images.

LeNet-5 extracts low-level image features such as edges and textures, VGG-16 captures detailed spatial facial information, while ResNet-50 extracts deep semantic features using residual learning. The extracted feature vectors from all three models are combined using feature concatenation techniques to create a stronger emotional representation.

The concatenated feature vectors are then passed into a Support Vector Machine classifier for final emotion classification.

Technology Stack used:

1. Frontend: React 18.2 + Vite 5 + TypeScript 5.2 + Tailwind CSS 3.4 + shadcn/ui + Framer Motion 12 + React Router DOM 7.7
2. Backend: Node.js + Express.js 4.18 + Mongoose 8.17 + JWT 9.0.2 + bcryptjs 3.0.2 + Nodemailer 7.0.5
3. AI: Google Gemini Pro via @google/generative-ai 0.24.1 + SimpleVectorRAG + EmbeddingServic.
4. Travel APIs: ONDC/EMT Flight API + EMT Hotel API + eRail API + ondc-crypto-sdk-nodejs 2.1.1

5. Security: Helmet 8.1 + express-rate-limit 8.0 + xss-clean + express-mongo-sanitize + hpp
6. Infrastructure: Docker + Nginx + Vercel + MongoDB Atlas

+ NodeCache 5.1 + node-cron + Winston

Metric	Single CNN Model	Ensemble CNN + SVM
Feature Representation	Moderate	High
Classification Accuracy	Moderate	Improved
Overfitting	Higher	Reduced
Generalization Capability	Moderate	Improved

3.4 Model Training

The CNN models were independently trained using the preprocessed FER-2013 dataset. Each architecture was configured with convolutional layers, pooling layers, flattening layers, and dense layers for learning emotional patterns from facial images.

The training process involved forward propagation, backpropagation, and weight optimization using Deep Learning optimization techniques. The CNN models learned important facial features including:

- facial textures,
- emotional structures,
- and spatial image patterns.

After training, feature vectors were extracted from intermediate layers of:

- LeNet-5,
- ResNet-50,
- and VGG-16.

These extracted features were concatenated into a single combined feature vector. The final combined feature representation was then used to train the SVM classifier for emotion prediction.

The ensemble learning approach improved classification robustness, reduced overfitting, and enhanced overall system performance compared to traditional single-model FER systems.

3.5 Evaluations

The proposed FER system was evaluated using classification accuracy, prediction performance, and confusion matrix analysis. The testing phase focused on analyzing the system's capability to correctly classify unseen facial images into their corresponding emotion categories.

The ensemble CNN-SVM architecture demonstrated improved emotion recognition performance compared to traditional single-model systems. The combination of multiple CNN architectures increased feature diversity and improved generalization capability.

The following table presents the comparative performance of the proposed system:

Table 1: Performance Comparison — Single CNN vs Ensemble CNN-SVM Model

The evaluation results demonstrate that the ensemble-based hybrid architecture provides better emotion classification capability and more robust feature learning compared to traditional approaches.

:

IV. CONCLUSION

The proposed Facial Emotion Recognition (FER) system provides an effective and intelligent solution for recognizing human emotions from facial expressions using Deep Learning and Machine Learning techniques. The system combines multiple CNN architectures including LeNet-5, ResNet-50, and VGG-16 with a Support Vector Machine classifier to improve feature extraction capability, classification robustness, and overall prediction performance.

The use of the FER-2013 Dataset enabled the system to learn and classify seven different human emotions including angry, disgust, fear, happy, sad, surprise, and neutral. The ensemble learning approach improved feature diversity and reduced the limitations associated with traditional single-model FER systems.

The preprocessing pipeline including image reshaping, normalization, and label encoding improved model convergence and training stability. Feature extraction using multiple CNN architectures helped the system capture both low-level and high-level emotional representations from facial images, while the SVM classifier effectively separated emotional classes using combined feature vectors.

The experimental evaluation demonstrated that the hybrid CNN-SVM architecture achieved improved emotion

recognition performance, better generalization capability, and reduced overfitting compared to standalone CNN models. The system showed effective classification capability even under challenging conditions such as image noise and similarities between emotional expressions.

Overall, the project successfully demonstrated the practical implementation of Artificial Intelligence, Computer Vision, Deep Learning, Ensemble Learning, and Machine Learning techniques for Facial Emotion Recognition. The developed FER system has potential applications in healthcare, surveillance systems, driver monitoring, online education, gaming, robotics, and human-computer interaction systems.

V. FUTURE SCOPE

The proposed Facial Emotion Recognition (FER) system can be further improved and extended through several advanced research and development directions. Although the current system demonstrates effective emotion classification performance using ensemble Deep Learning and SVM classification, there are multiple opportunities for enhancing accuracy, scalability, and real-world usability.

First, the system can be extended to support real-time emotion recognition using webcams or CCTV video streams. Integrating live video processing would enable the FER model to analyze emotions continuously in real-world environments such as surveillance systems, classrooms, healthcare centers, and driver monitoring applications.

Second, more advanced Deep Learning architectures such as Vision Transformers (ViT), EfficientNet, and attention-based neural networks can be implemented to improve feature extraction capability and classification accuracy. These architectures may provide better handling of complex facial variations and improve robustness under different lighting and pose conditions.

Third, the current system works only on static facial images. Future systems can integrate multimodal emotion recognition using:

- facial expressions,
- speech signals,
- and text sentiment analysis

to achieve more accurate and context-aware emotion prediction.

Fourth, larger and more balanced datasets can be used to improve model generalization capability and reduce classification bias between emotion classes. Data augmentation techniques and synthetic image generation can also be explored to address dataset imbalance problems.

Fifth, the FER system can be deployed as:

- web applications,
- mobile applications,
- or cloud-based AI services

to make emotion recognition accessible in real-world environments. Mobile deployment can enable applications in smart devices, virtual assistants, and wearable systems.

Finally, future work may focus on reducing computational complexity and improving inference speed for lightweight real-time deployment. Techniques such as model compression, pruning, and quantization can help optimize the system for low-resource devices while maintaining classification performance.

VI. REFERENCES

[1] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. "Gradient-Based Learning Applied to Document Recognition." *Proceedings of the IEEE*, vol. 86, no. 11, 1998, pp. 2278–2324.

[2] Simonyan, K., and Zisserman, A. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *International Conference on Learning Representations (ICLR)*, 2015.

[3] He, K., Zhang, X., Ren, S., and Sun, J. "Deep Residual Learning for Image Recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[4] Cortes, C., and Vapnik, V. "Support-Vector Networks." *Machine Learning*, vol. 20, no. 3, 1995, pp. 273–297.

[5] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning.* MIT Press, 2016.

[6] FER-2013 Dataset Documentation. "Challenges in Representation Learning: Facial Expression Recognition Challenge." Kaggle Dataset Repository.

[7] Chollet, F. *Deep Learning with Python.* Manning Publications, 2018.

[8] Krizhevsky, A., Sutskever, I., and Hinton, G. "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

[9] Mollahosseini, A., Hasani, B., and Mahoor, M. H. "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild." *IEEE Transactions on Affective Computing*, 2017.

[10] TensorFlow Documentation. "TensorFlow Deep Learning Framework." Available: https://www.tensorflow.org/