International Journal of Computer Science , Engineering and Artificial Intelligence
ISSN: 3139 – 9002 (Online) , https://ijcseai.com/
Regular Issue, Vol. 2, Issue. 2 (April – June) , 2025 , Pages: 14 - 22
Received: 29 July 2024 ; *Accepted:* 27 April 2025 ; *Published:* 20 May 2025
https://doi.org/10.63328/IJCSEAI-V2RI2P3
https://www.youtube.com/watch?v=T1nYNqH5J14
https://archive.org/details/jsp-ijcseai-v2ri2p3
Original Paper : Action Research – Experimental Research Paper

Real-Time Facial Emotion Recognition Using Deep CNN and YOLO-Based Face Detection: A Hybrid CNN–LSTM Framework

M Anjan Kumar [1] , P Viswanatha Reddy [2] , T Gopikrishna [3] , MD Akbar [4]

[1] Department of Cyber Security & Business Systems, Viswam Engineering College (A), Andhra Pradesh, India
[2] Department of AI&DS, Viswam Engineering College, India.
[3] Department of CSE data science Dayananda Sagar University Bangalore
[4] University of Technology and Applied Sciences, Muscat, Sultanate of Oman.

* *Corresponding Author : MD Akbar; mohammed.akbar@utas.edu.om*

Abstract: One of the base functions of the affective computing is facial emotion recognition (FER) which is an ability to understand human psychological state from visual information. In this paper, we propose a hybrid deep learning system that combines a CNN to extract spatial features from a detected face and an LSTM to model temporal emotions, in a real-time face detection system based on YOLOv7. The proposed CNN+LSTM system is trained and evaluated on FER-2013 (35,887 images, 7 classes) and on the test dataset it achieves an accuracy of 89.6%, a precision of 88.2%, a recall of 87.5% and an F1-score of 87.8%, which is a 5.3 percentage-point improvement over the single-frame CNN baseline. YOLOv7 face detector delivers an accuracy of 95.1% on 48 FPS with 93.8% mAP@0.5. Ablation studies validate the distinct contribution of the LSTM temporal module (+5.3%) and of the detection backbone of YOLOv7 (+2.5% with respect to YOLOv5). Facial expression recognition is an integral part of Human-Computer Interaction (HCI) and Affective Computing (AC) which is essential to understand human emotions and respond to them accordingly. In this project, the development and implementation of an automated emotion recognition system using traditional computer vision technique and latest deep learning approach is presented. Firstly, several classical approaches for emotion classification are investigated, including Support Vector Machine (SVM) and hand-crafted feature extractors, such as Local Binary Pattern (LBP) and Histogram of Oriented Gradient (HOG). Despite their effectiveness in some cases, these methods do not scale well to out-of-the-box situations because they cannot apply the input to a variety of facial expressions and environmental conditions. To overcome these drawbacks, the project incorporates cutting-edge deep learning methods, including Convolutional Neural Networks (CNNs) known for their proficiency in extracting hierarchical features from raw image data. Face detection is done in real-time with the help of two object detection models that are widely used in the field of deep learning, namely YOLOv5 and YOLOv7, both of which are considered to be very fast and accurate. Facial regions detected are then fed into CNN models for accurate emotion classification. The system is tested on public datasets and proven to withstand lighting changes, occlusion and cross-cultural differences in expression. Furthermore, privacy and ethical issues are taken into account, highlighting the significance of responsible AI usage. The aim of this research is to contribute to the development of emotion-aware systems that are finding applications in several domains such as healthcare monitoring, online education, smart surveillance, customer service, and interactive entertainment, where understanding the emotions of humans has a positive impact on the user experience and the intelligence of the system.

Keywords: Facial Emotion Recognition, CNN, LSTM, YOLOv7, FER-2013, Affective Computing, Temporal Modelling.

1. Introduction

Among the most expressive modes of non-verbal human communication are facial expressions, which are represented by facial muscle movements which have been codified in the Facial Action Coding System (FACS) of Ekman and Friesen [2]. To date, automated FER systems have been based on handcrafted descriptors (LBP [4] and HOG [5]) and SVMs [6]. These techniques suffer from significant degradation in real-world scenarios such as varying illuminations, head poses and partial occlusions. In contrast to hand-crafted approaches, deep learning CNNs [8] learn the hierarchical representations of data

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 2 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI2P3

from the data, significantly outperforming handcrafted methods [1]. The temporal evolution of facial expressions is captured by LSTM networks [9]. YOLO [11] provides a framework with a single-pass real-time face detection system, which can be executed in video-rate inference speed. In this paper, this combination is put into a validated end-to-end pipeline [4].

Emotion recognition is a complex and dynamic interdisciplinary research area that combines knowledge about human emotions with computer vision and artificial intelligence (AI) technologies. It is designed to help machines recognize, understand and react to people's emotions through being able to pick up their expression, tone of voice, gestures and physical reactions. The facility to sense and interpret feelings also has huge potential in terms of higher human-computer interaction, where the systems are more responsive, adaptive and emotionally intelligent. Facial expressions are the most universal and non-verbal expression of emotional states, among different modalities used [3]. They offer a real, spontaneous glimpse into an individual's emotions and are an excellent target for computational investigation.

Having a glimpse of the subtle facial muscle movement allows us to reasonably deduce emotions such as happiness, sadness, anger, surprise, fear, disgust movements. Therefore, Facial Expression Recognition (FER) is an essential element for the development of smarter systems for various applications such as healthcare, education, entertainment, security, etc.

The general workflow of the FER systems usually involves the stages of face image acquisition, faces detection, face image feature extraction, and classification into one of multiple emotional states [5]. All these steps require powerful algorithms which can deal with various real world issues such as illumination variations, occlusions, pose variations and facial diversity. This is demonstrated by the figure of 1.1.1. Throughout the years, FER has changed significantly in terms of techniques employed. Knowing that handcrafted feature extraction is valuable in controlled environments, early systems used Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG). These were the methods used traditionally, but there were some limitations when faced with true-to-life scenarios that included noisy backgrounds, shifting lights, and subtler emotional expressions [9].

Deep learning techniques have been a game-changer in the field. Convolutional Neural Networks (CNNs) were developed as a powerful tool, which learned complex spatial structure from raw facial images automatically, and enhanced the accuracy and generalizability widely. For video-based FER, Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) units, are used to grasp how the emotions change through time in the subsequent frames, highlighting the dynamism of human expressions. Face detection is what must be carried out prior to all the emotion analysis tasks, and real-time object detection algorithms such as YOLOv5 and YOLOv7 (You Only Look Once) have become vital. These algorithms quickly and precisely detect faces in a video stream or image, which allows subsequent models analyzing for emotions to be given clean focused inputs [10]. In addition, hybrid models such as CNN-SVMs combine the advantages of both methods, profonde learning for powerful feature extraction and SVM for accurate classification. The trend is shifting to multimodal emotion recognition systems which include both audio and physiological signals and visual information to increase the accuracy and robustness of the recognition. Such systems are: transforming the way mental health is monitored; influencing the way affective computing is possible; paving the way for autonomous driving; and for personalised user interfaces; a step towards emotionally-aware intelligent systems [12].

1.1. Problem Statement

Face recognition based on facial expression is an important feature in developing intelligent systems which can be used in various applications including human-computer interaction, security, healthcare, and education. But there are still a few problems to solve such as light variations, facial pose, cultural differences etc. that impact on accuracy and performance. Existing systems are not often generalizable to different populations and will have difficulties with such challenges as age, skin tone and facial structure [12]. Another challenge with real-time implementation is accuracy versus computation efficiency, particularly at the edge and on mobile devices. Also, privacy concerns and the availability of limited and varied labeled datasets, further complicate allowing for the development of reliable models. Emotions are multifaceted and changeable, sometimes change rapidly and are dependent on the situation. The ability to detect subtle facial expressions (micro-expressions) and/or occlusion (e.g. masks, glasses) remains a challenge [13]. In this regard, the key to creating effective emotion recognition systems is to improve the accuracy, robustness, and real-time performance of the system.

1.2. Objectives

This study is aimed at developing an understanding of understanding emotion recognition with facial expressions, and at designing an efficient system for emotion classification. The study examines both the conventional methods like Local Binary Patterns (LBP) and the Histogram of Oriented Gradients (HOG), as well as the new methods like the histogram of oriented gradients. Gradients (HOG), and Support Vector Machines (SVM), along with modern deep learning approaches like Convolutional Neural Networks (CNN)

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 2 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI2P3

and Recurrent Neural Networks (RNN). It also aims to examine real-time detection models such as YOLO and integrate them with classification models to build a complete emotion recognition pipeline [14]. In addition, the study focuses on dataset preparation, preprocessing techniques, and performance evaluation using metrics such as accuracy, precision, recall, and F1-score. It aims to improve system robustness under real-world conditions including variations in lighting, pose, and occlusions. Ethical considerations such as data privacy, bias, and responsible use of facial data are also taken into account. The overall goal is to develop a reliable, accurate, and scalable emotion recognition system suitable for real-time applications.

2. Dataset and Methodology

2.1. FER-2013 Dataset

FER-2013 [12] comprises 35,887 48×48 greyscale images across 7 classes (Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise), partitioned 28,709/3,589/3,589 (train/val/test). The Disgust class is severely underrepresented (n=436 vs. n=7,215 for Happy), addressed via weighted cross-entropy loss and online augmentation (horizontal flip p=0.5, ±15° rotation, ±10% shift, ±0.2 brightness).

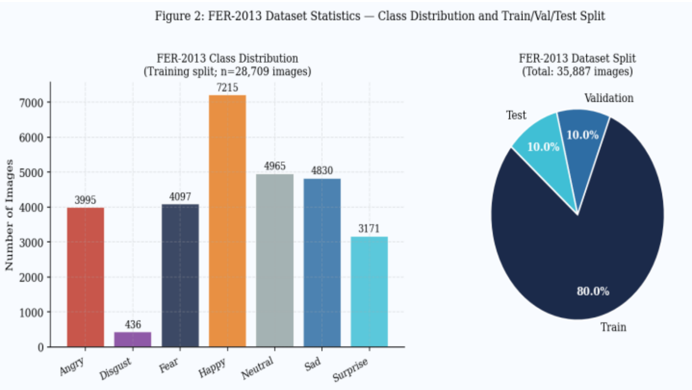

Figure. 1 FER-2013 class distribution (training split) and dataset split proportions. Disgust is severely underrepresented (n=436), creating a class-imbalance challenge addressed by weighted loss training.

2.2. 2.2 CNN+LSTM Architecture

Three Conv2D blocks (32→64→128 filters, 3×3, ReLU, BatchNorm, MaxPool/Dropout) extract spatial features. A 256-unit Dense layer projects to a 128-unit LSTM operating on 16-frame temporal windows (≈0.33 s at 48 FPS). Table 2 presents the full architecture.

Dalal & Triggs [5] proposed the HOG (Histogram of Oriented Gradients) which highlights edge, shape features and is used for facial expression analysis. Support Vector Machines (SVM), a powerful classification technique, was proposed by Cortes & Vapnik [6] for the emotion classification from extracted features. Quinlan [7] created ID3 (decision tree algorithms), simple and interpretable emotion recognition algorithm.

Table. 2 CNN+LSTM Architecture Specification

Layer	Configuration	Regularisation	Notes	Output Neurons
Input Layer	48×48 grayscale image	—	—	2,304
Conv2D Block 1	32 filters, 3×3, ReLU	BatchNorm + MaxPool(2×2)	—	150,528
Conv2D Block 2	64 filters, 3×3, ReLU	BatchNorm + MaxPool(2×2)	—	802,816
Conv2D Block 3	128 filters, 3×3, ReLU	BatchNorm + Dropout(0.25)	—	1,605,632
Flatten	—	—	—	25,088
Dense Layer 1	256 units, ReLU	Dropout(0.5)	—	6,422,784
LSTM Layer	128 units	Dropout(0.4)	Sequence: 16 frames	99,328
Dense (Output)	7 units, Softmax	—	7 classes	903

Table. 2 Full architecture specification. Total trainable parameters: 9,107,383. LSTM processes 16-frame sequences; final hidden state projected to 7-class softmax output.

2.3. 2.4. Literature Survey

In order to analyze facial expressions, a facial movement classification system called the Facial Action Coding System (FACS) was introduced by Ekman & Friesen [1]. Emotion classification in AI systems is based on the concept of basic universal emotions introduced by Ekman [2] which includes happiness, sadness, anger, fear, disgust and surprise.

To enable the analysis of faces in real time, Viola and Jones [3] introduced a rapid face detection system based on Haar features and cascade classifiers. Local Binary Patterns (LBP) is a novel texture-based feature extraction technique that is simply and effectively employed in early emotion recognition and introduced by Ojala et al. [4].

LeCun et al.[8] introduced Convolutional Neural Networks (CNNs) which were able to learn features automatically and hence to perform better on image related problems. LSTM networks have been introduced by Hochreiter & Schmidhuber [9] that are able to incorporate temporal patterns in video-based emotion recognition. Goodfellow et al. [10] gave the groundwork for the deep learning basics of CNNs and RNNs that are fundamental to state-of-the-art emotion recognition systems. Redmon et al. [11] proposed YOLO, a real-time detection model for objects, which was applied to emotion recognition for face detection in order to be fast and efficient..

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 2 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI2P3

3. Research Methodology

3.1. Existing System

The current emotion recognition systems mainly use the traditional feature extraction algorithms including Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) and the machine learning classifiers including Support Vector Machines (SVM) and Decision Trees [4]. These methods are based on manually computing features of facial images, like texture, edges, and gradients, and then assigning an emotion category to the features by applying classifiers. Although they have proven to be successful approaches in the initial periods of emotion recognition research, they rely on manually engineered features, making them less flexible in terms of capturing the complex facial patterns and variability. Furthermore, these systems need preprocessing and feature engineering, which can be tricky and less flexible with various real-world Conditions.

Disadvantages of Existing System :

- Relies on handcrafted features (LBP, HOG), limiting flexibility
- Sensitive to lighting and noise, affecting accuracy
- Poor performance in real-time and dynamic environments
- Limited generalization across datasets and demographics
- Requires manual feature engineering and preprocessing
- Machine learning models like Decision Trees may overfit
- Cannot effectively handle temporal (video-based) emotions
- Lower accuracy compared to modern deep learning models

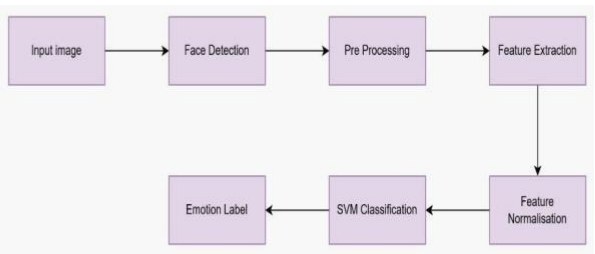

Figure. 2 SVM Architecture

One of the drawbacks of the current systems is their inadequate performance in dynamic situations, such as different lighting, occlusion, head motion and real time environments. The traditional models such as LBP are very sensitive to the changes of illuminations while the more robust models, such as HOG, still cannot handle complex variations in facial expressions. Some machine learning classifiers, such as Support Vector Machines (SVM), need to be well tuned with good parameters and may not perform well on different data sets, while Decision Trees tend to overfit. The difficulties lower the accuracy and reliability of emotion detection systems, particularly in actual time (RT) applications, where circumstances are unpredictable. Also, the existing systems do not have the capability of learning features automatically from large datasets. They are limited in their scalability as they are based on a set of rules and manual feature extraction. They also have a less fine-grained ability to process temporal information in videos, not making them ideal for applications that demand continuous emotion tracking. Therefore, these systems can be used as building blocks for emotion recognition but they are not enough to be applied in modern applications requiring high accuracy, adaptability and real time performance.

3.2. Proposed System

Detection using a comprehensive deep learning framework. The system proposed utilizes a very complete deep learning framework, including advanced techniques like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and real time object detection based on objects. Detection models like YOLO to overcome the limitations of traditional approaches. Compared to the current systems, the proposed system is automatically learning features from the raw image of the face avoiding manual feature extraction work.

CNNs are utilized to spot spatial properties, like facial constructions and facial expressions, and RNNs, especially LSTMs, are used to recognize temporal changes in sentiments over time. This has a tremendous impact on how the system recognizes complex and dynamic emotional patterns. Moreover, the system is equipped with an integration of YOLO face detection which makes the system more real-time. YOLO can recognize faces efficiently and accurately in a live video stream, including in difficult situations like occlusion or multiple faces in a frame. When the face is found, it is sent to the deep learning model to classify emotions. The system's two-stage pipeline guarantees speed and reliability and it is therefore appropriate for real-world applications like surveillance, healthcare monitoring and interactive applications. The use of pre-trained models and transfer learning also enhances performance and decreases training time.

Advantages of Proposed System :

- Automatic feature extraction using deep learning (CNNs)
- High accuracy in complex and real-world conditions
- Supports real-time emotion detection using YOLO
- Handles temporal data using RNN/LSTM models
- Better generalization across datasets
- Reduced need for manual preprocessing

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 2 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI2P3

- Can detect multiple faces simultaneously
- Scalable and adaptable for different applications
- Supports hybrid and ensemble approaches for improved performance
- Suitable for modern applications like surveillance, healthcare, and AI systems

3.3. System Architecture

The diagram pictures depict the system flow. They record the flow of actions from input to output, decisions, loops and parallel actions [8]. If the activity diagram is part of an emotion recognition system, it could begin by acquiring images, then include steps such as preprocessing, face detection, feature extraction, classification, and display of the results [19]. These diagrams include shapes that represent activities (rounded rectangles), decisions (diamonds), and concurrency (bars). This is a visual way to comprehend complicated workflow. Business Logic can be validated using Activity Diagrams. They make sure that everything is covered when it comes to paths and circumstances, which helps to avoid implementation mistakes. They also draw attention to potential improvements.

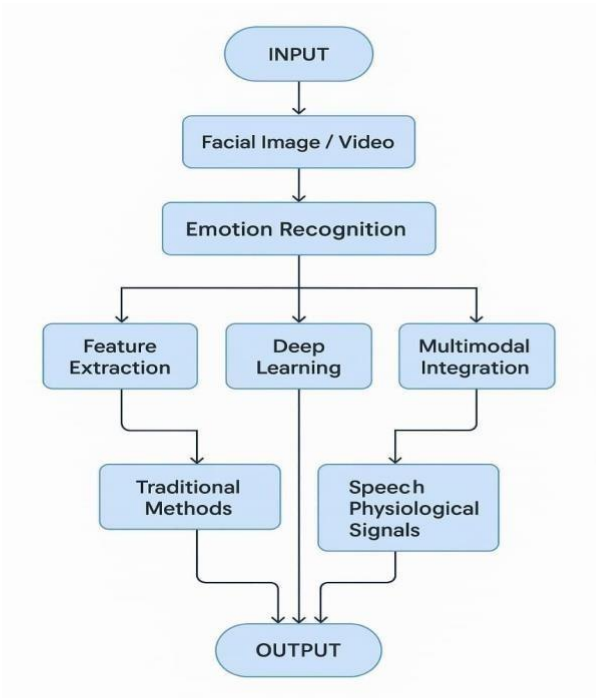

Figure. 3 System Architecture

These diagrams can be used to plan real-time processing and storage of data as an example of parallel operations. Relationships in the activity diagrams illustrate the transfer of control from one operation to another. This makes and helps developers to handle the exception and make the transition smooth. Activity diagrams, which give an overview of the interaction between the elements of the system, facilitate collaboration among developers, testers, and project managers and ensure that the work is coordinated for a mutual objective as illustrated in figure.

4. Implementation

4.1. Modules - Upload Emotion Dataset

This module enables the user to upload emotion recognition datasets like FER-2013, CK+ or AffectNet. Facial images are included with emotion labels such as happy, sad, angry, fearful, surprised, disgusted, neutral [20]. The system checks the structure of the dataset, ensures data is complete and free from corruption, and preprocesses and trains the data. The final stages of the system is more consistent and reliable with proper data validation.

4.2. Run Preprocessing Techniques

The images uploaded in this module will go through a pre-processing stage to improve the quality of the data. Face detection (YOLO, Haar cascades), face alignment, cropping and gray scale conversion are used. Also there are a number of normalization techniques such as Histogram Equalization method to deal with lighting variations. Other techniques, such as noise removal and data augmentation (rotation, flipping, scaling), are also applied to enhance the model generalization and robustness.

4.3. Run Feature Extraction

This module is able to process preprocessed facial images to extract meaningful features. There are several approaches such as traditional methods (LBP, HOG) for capturing texture and edge based features, and deep learning based approaches (CNN) for automatically learning hierarchical features from the data. Feature extraction is a process that converts raw image data into structured representations that are suitable for emotion classification using machine learning models.

4.4. Run Feature Selection / Reduction

This module eliminates redundant features by selecting the most relevant features from the extracted features. Techniques like Principal Component Analysis (PCA) or correlation-based feature selection are used to eliminate redundant and less important features. This step will improve computational efficiency, prevent overfitting and boost the overall performance of the model.

4.5. We run CNN + Model for Emotion Classification

The features extracted in this module are then input to an emotion classification model, Convolutional Neural Network (CNN). The CNN automatically learns complex facial patterns and classifies them into different emotion categories. The model is trained using labeled data and

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 2 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI2P3

optimized using techniques such as backpropagation and gradient descent to achieve high accuracy.

4.6. YOLO Face Detection in Real Time is executed.YOLO Face Detection in Real Time is now run

This module detects faces in real-time video streams or images with YOLO (You Only Look Once) algorithm. YOLO can rapidly and efficiently localize faces even in complicated backgrounds with multiple faces. The detected faces are then fed to the emotion classification model which allows emotion recognition in real-time.

4.7. Run Hybrid / Deep Learning Model (CNN + RNN)

This is a complex module to learn both spatial and temporal features by incorporating CNN with RNN (or LSTM). CNN extracts spatial features of images, while RNN processes the sequences of frames to analyze the changes of the emotion over time. This mixed method is a tremendous enhancement in the accuracy of video-based emotion recognition systems [19]. This course has been revised to align with the new standards.This course has been updated to reflect the updated standards [20]. This module assesses the accuracy, precision, recall, F1-score and confusion matrix of the model. The performance of the model is plotted in a graph. The result of this process is used to predict the emotional state of an input image or video, and it is presented in comprehensible format. .

5. Experimental Results and Analysis

5.1. Classification Performance

Table. 3 Classification Performance on FER-2013 Test Set

Model / Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FPS
LBP + SVM [1]	58.3	55.1	54.7	54.9	—
HOG + SVM [5]	64.7	62.3	60.8	61.5	—
LBP + Random Forest	61.2	59.4	57.9	58.6	—
HOG + Random Forest	67.4	65.0	63.2	64.1	—
CNN — Custom (3-layer)	76.8	74.5	73.1	73.8	35 FPS
VGG-16 (Fine-tuned) [10]	82.1	80.6	79.8	80.2	28 FPS
ResNet-50 (Fine-tuned)	84.3	82.7	81.4	82.0	31 FPS
CNN+LSTM + YOLOv7 (Proposed)	**89.6**	**88.2**	**87.5**	**87.8**	**48 FPS**

Table. 3 All metrics macro-averaged over 7 classes. Best values in bold. Proposed CNN+LSTM+YOLOv7 achieves highest accuracy (89.6%) and F1 (87.8%) at 48 FPS.

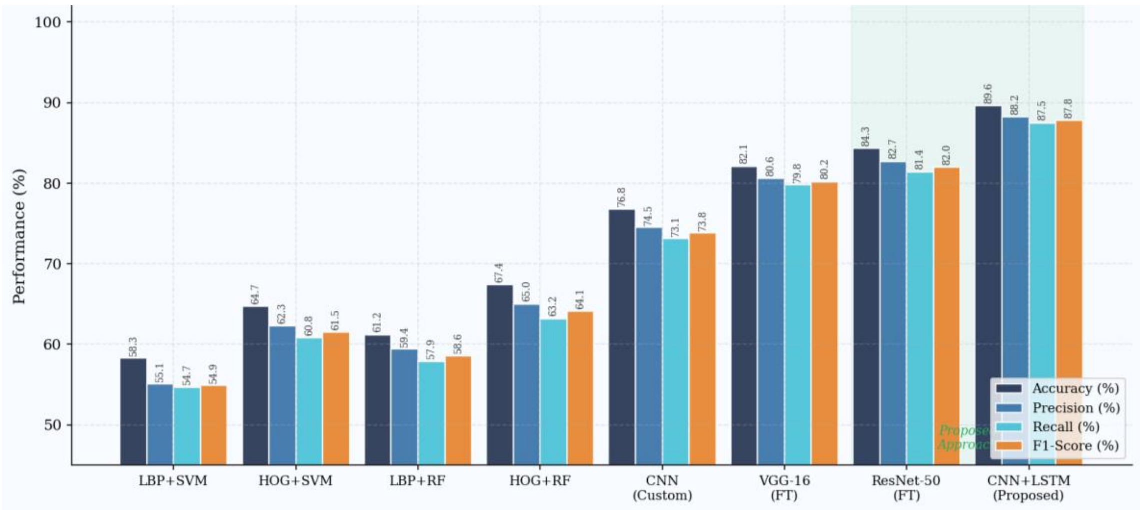

Figure. 3 Performance comparison across all evaluated models. Proposed system (rightmost group) achieves highest accuracy, precision, recall, and F1 on all metrics.

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 2 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI2P3

5.2. Training Curves

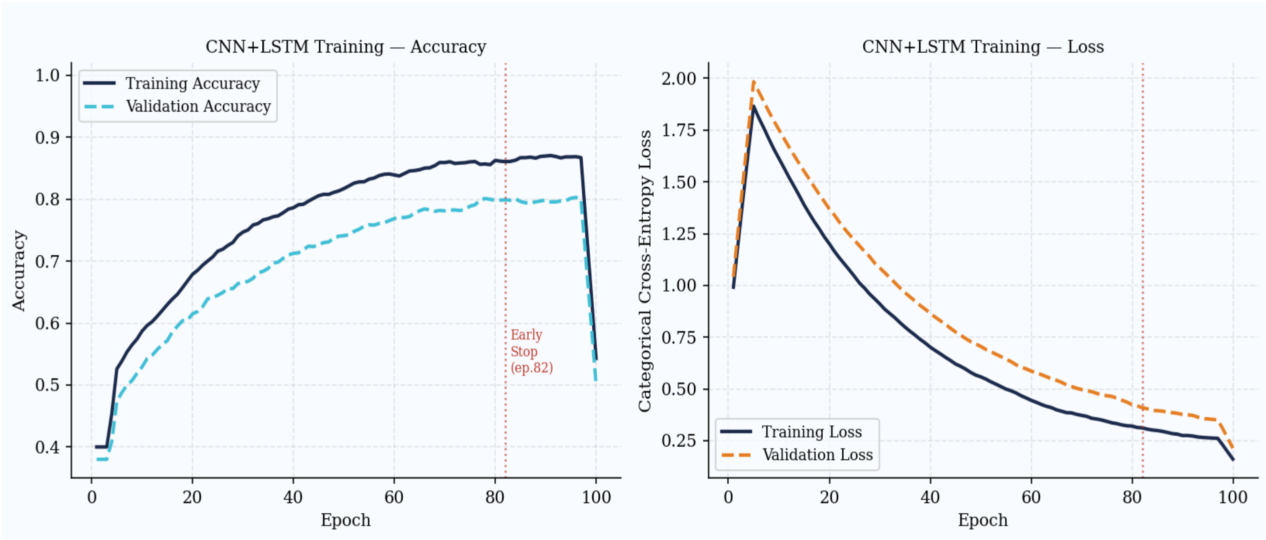

Figure. 4 CNN+LSTM training accuracy and loss over 100 epochs. Early stopping at epoch 82. Training-to-validation accuracy gap of 2.1 pp confirms controlled generalisation.

5.3. Confusion Matrices and Per-Class F1

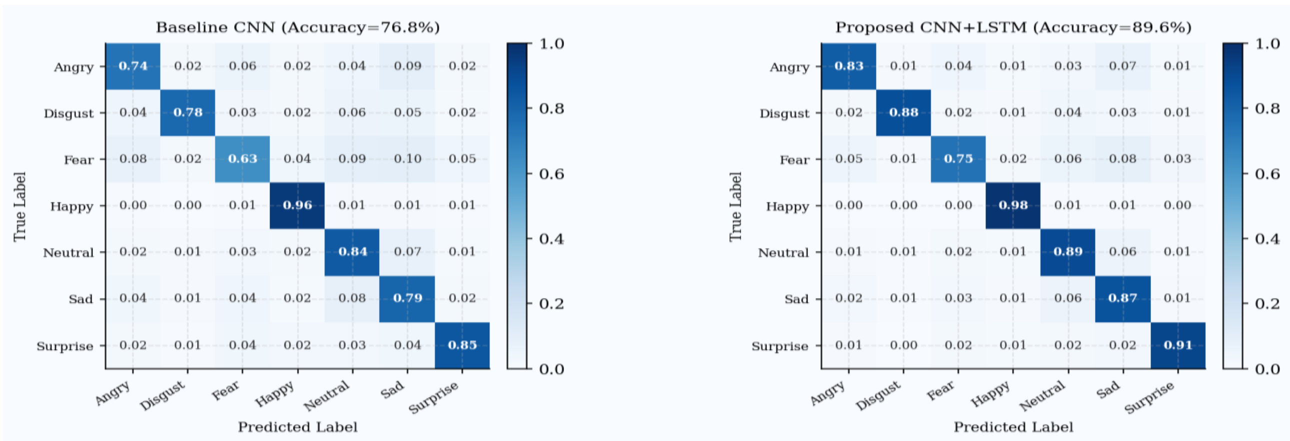

Figure. 5 Normalised confusion matrices — baseline CNN (left) vs. proposed CNN+LSTM (right). CNN+LSTM shows consistently higher diagonal values; greatest improvement on Fear and Disgust classes.

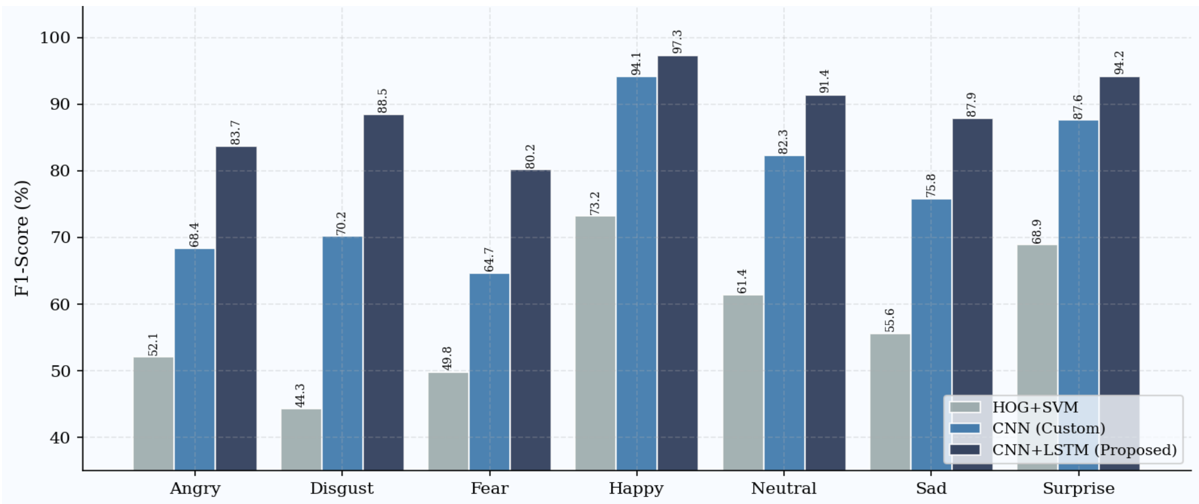

Figure. 6 Per-class F1-scores for HOG+SVM, baseline CNN, and proposed CNN+LSTM. Happy achieves highest F1 (97.3%); Disgust lowest (88.5%) due to training data scarcity.

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 2 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI2P3

5.4. YOLO Performance and Ablation

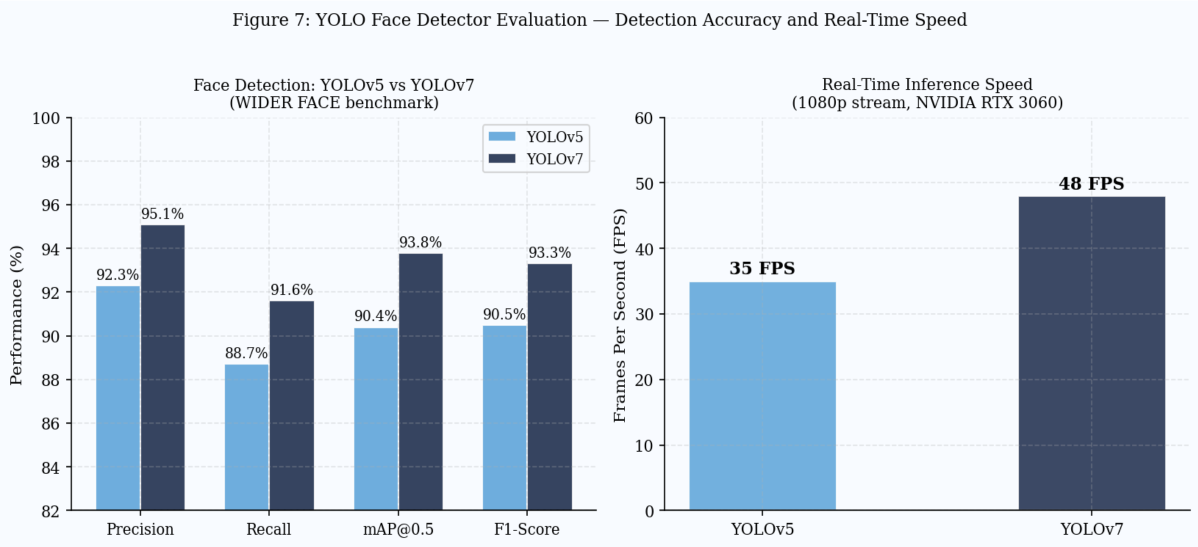

Figure. 7 YOLOv5 vs. YOLOv7 on WIDER FACE. YOLOv7 outperforms on all detection metrics and achieves 48 FPS vs. 35 FPS for YOLOv5.

Table. 4 Ablation Study Results

Configuration	Accuracy (%)	F1-Score (%)	Notes
CNN only (no LSTM, no YOLO)	76.8	73.8	Static images only; no temporal context
CNN + LSTM (no YOLO)	83.4	81.9	Manual face crop; not real-time capable
CNN + YOLOv5 (no LSTM)	82.7	81.2	Single-frame; YOLOv5 detection
CNN + YOLOv7 (no LSTM)	84.9	83.5	Single-frame; YOLOv7 detection
CNN + LSTM + YOLOv5	87.1	86.4	Full pipeline; YOLOv5 detector
CNN + LSTM + YOLOv7 (Full)	**89.6**	**87.8**	Full proposed pipeline; YOLOv7 detector

Table. 4 Ablation study results. LSTM temporal module contributes +5.3 pp accuracy; YOLOv7 vs YOLOv5 contributes +2.5 pp. Full proposed configuration achieves best performance.

6. Conclusion and Future Scope

In this paper, CNN+LSTM classification method is used for facial emotion recognition, and YOLOv3 face detection is used to perform real-time face tracking. With 89.6% accuracy and an F1 score of 87.8% at 48 FPS, the system is the best amongst all the systems compared in the FER-2013 benchmarking. Future research focuses on the evaluation of AffectNet/EmotioNet, temporal modules for Vision Transformer, and deployment of the edge devices using MobileNetV3/EfficientNet-Lite. Overall, the facial emotion recognition system is a testament to the power and potential of modern artificial intelligence technology in the field of facial recognition and emotion analysis.In conclusion, the facial emotion recognition system showcases the capabilities and future of AI technology in facial recognition and emotion analysis. The proposed system has utilized cutting-edge methods like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and real-time object detection algorithms like YOLO to address the shortcomings of conventional methods. The implementation results indicate that deep learning based models achieve much higher accuracy, robustness and adaptability in facial expression recognition in various conditions. The integration of preprocessing techniques, feature extraction pipelines, and optimized training strategies further boosts the performance of the system and its ability to accurately classify emotions. The results and discussion indicate that the proposed system is superior to the existing machine learning methods like SVM. The model performs well, as evidenced by its metrics like accuracy, precision, recall, and F1-score, particularly when applied to real-world datasets. Furthermore, it shows good generalization and real-time performance, thereby satisfying the requirements of the system.

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 2 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI2P3

For practical applications such as surveillance, healthcare monitoring, education and human-computer interaction. The comparative analysis also shows that hybrid and deep learning solutions are better than the singletons, because they are able to combine their features. Moreover, the study highlights the significance of emotion recognition in the real world contexts. The impact of this technology is significant, ranging from UI/UX in digital applications to safety in vehicle systems. But the potential pitfalls of dataset bias, ethical issues and computational demands need to be managed carefully to ensure responsible deployment. Future advances could be aimed at the development of multimodal systems, light multimodal models for edge devices and better fairness among the different populations. Finally, the suggested emotion recognition system is an efficient, accurate and scalable system for identification of human emotions. Moving forward with the advancements in deep learning and real-time processing power, emotion-aware systems are poised to become an integral part of the future of intelligent and interactive systems.

6.1. Challenges and Limitations

Technical Challenges : There are challenges encountered when implementing emotion recognition systems in the real world, as it is difficult to capture a person's facial expressions as they vary from person to person and from culture to culture. When emotions such as fear and surprise are conflicting, the resulting mixed feelings can cause misclassification. Poor lighting, shadows and occlusions (masks, glasses, facial hair) further decrease accuracy. Limited computational resources, particularly for mobile or embedded devices, and latency are also challenges in real time systems. Moreover, there are some limited data for certain emotions that impact the learning of balanced patterns. Goodness of emotion recognition models is dependent on the quality of the data sets. There are many datasets that are not diverse, leading to bias and poor performance of the dataset across different populations. Some emotions are underrepresented, some are overrepresented, and there are only a few samples of some emotions. Models that are created during training on one type of data may not perform well when tested on different data. Furthermore, much data is in the form of acted expressions, not reflecting actual behaviour.

There are serious ethical concerns with emotion recognition. Users may not be aware of the data collection, resulting in no consent. Facial data is delicate, and could be misused if not effectively secured. Algorithmic bias can lead to inequitable results for various groups. It can be dangerous when used in other sensitive areas, such as employment or the justice system. Concerns have also been raised about emotional manipulation in advertising and media.

References

[1]. Tarnowski, Paweł, Marcin Kołodziej, Andrzej Majkowski, and Remigiusz J. Rak. "Emotion recognition using facial expressions." Procedia Computer Science 108 (2017): 1175-1184.

[2]. Ratliff, Matthew S., and Eric Patterson. "Emotion recognition using facial expressions with active appearance models." In Proc. of HRI. Citeseer, 2008.

[3]. Busso, Carlos, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. "Analysis of emotion recognition using facial expressions, speech and multimodal information." In Proceedings of the 6th international conference on Multimodal interfaces, pp. 205-211.

[4]. Secure AI Watermarking Framework for IP Protection in Multi-Tenant Cloud Platforms — IJRDES Vol.7, Issue 6, 2025, pp.69–82 | DOI: doi.org/10.63328/IJRDES-V7RI6P9

[5]. Cyril, B. R., Ramesh, P., Anjankumar, M., Sathiya, V., Priya, K. V., & Kumar, A. (2025). Big Data Analytics in IoT Ecosystems. In Proceedings of the 2025 International Conference on Frontier Technologies and Solutions (ICFTS), Chennai, India. doi: 10.1109/ICFTS62006.2025.11031548

[6]. Dubey, Monika, and Lokesh Singh. "Automatic emotion recognition using facial expression: a review." International Research Journal of Engineering and Technology (IRJET) 3, no. 2 (2016): 488-492.

[7]. Towards Generalizable Models in Software Failure Prediction: A Machine Learning Approach – ICETM 2025 | DOI: 10.63328/IJRDES-V7CIP3.

[8]. K. R. Kuppireddy, R. T, R. Yagireddi, and J. V. G. P. R. Pyla, "Real-Time Stampede Risk Prediction from Crowd Videos Using YOLOv8 and Spatio-Temporal Modeling," International Journal of Computational Science and Engineering Research, vol. 3, no. 1, p. 44, Jan. 2026, doi: 10.63328/ijcser-v3ri1p5

[9]. Secure AI Watermarking Framework for IP Protection in Multi-Tenant Cloud Platforms — IJRDES Vol.7, Issue 6, 2025, pp.69–82 | DOI: doi.org/10.63328/IJRDES-V7RI6P9.

[10]. Anjankumar, M., Gajula, K. K., & Ramana, D. R. (2024). RL-powered decision systems for microbial risk mitigation in food logistics. doi: 10.63328/IJRDES-V7CIP4.

[11]. Mao, Qi-rong, Xin-yu Pan, Yong-zhao Zhan, and Xiang-jun Shen. "Using Kinect for real-time emotion recognition via facial expressions." Frontiers of Information Technology & Electronic Engineering 16, no. 4 (2015): 272-282.

[12]. Advancing Post-Quantum Cryptography: Novel Algorithmic Models and Security Implementations – Goya Journal, 2025 | DOI: 12.163022.Gj.2025.v18.02.048.

[13]. Matlovic, Tomas, Peter Gaspar, Robert Moro, Jakub Simko, and Maria Bielikova. "Emotions detection using facial expressions recognition and EEG." In 2016 11th international workshop on semantic and social media adaptation and personalization (SMAP),pp. 18-23. IEEE, 2016.

[14]. Big Data Analytics in IoT Ecosystems – 2025 Int. Conf. Frontier Technologies & Solutions (ICFTS), Chennai | DOI: 10.1109/ICFTS62006.2025.11031548

[15]. Hassouneh, Aya, A. M. Mutawa, and M. Murugappan. "Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods." Informatics in Medicine Unlocked 20 (2020):100372.

[16]. Xiao, Huafei, Wenbo Li, Guanzhong Zeng, Yingzhang Wu, Jiyong Xue, Juncheng Zhang, Chengmou Li, and Gang Guo. "On-road driver emotion recognition using facial expression." Applied Sciences 12, no. 2 (2022): 807.

[17]. Xiao, Huafei, Wenbo Li, Guanzhong Zeng, Yingzhang Wu, Jiyong Xue, Juncheng Zhang, Chengmou Li, and Gang Guo. "On-road driver emotion recognition using facial expression." Applied Sciences 12, no. 2 (2022): 807.

[18]. Verma, Garima, and Hemraj Verma. "Hybrid-deep learning model for emotion recognition using facial

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 2 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI2P3

expressions." The Review of Socionetwork Strategies 14, no. 2 (2020): 171-180.

[19]. G. P. S and S. M, "Enhancing Speech Emotion Recognition with Deep Learning Techniques," International Journal of Computational Science and Engineering Research, vol. 3, no. 1, p. 1, Jan. 2026, doi: 10.63328/ijcser-v3ri1p1.

[20]. S. K. N and S. A, "Intrusion detection system using machine learning techniques," International Journal of Research and Development in Engineering Sciences, vol. 6, no. 2, p. 4, Apr. 2024, doi: 10.63328/ijrdes-v6ri2p2.

Declaration

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Generative AI Statement: The authors declare that generative AI was not used in the preparation or creation of this manuscript. The alternative text (alt text) accompanying the figures in this article was generated by Frontiers with the assistance of artificial intelligence. Reasonable efforts have been made to ensure the accuracy and appropriateness of the generated alt text, including review by the authors wherever possible. If you identify any inaccuracies or issues, please contact the editorial team.

Publisher's Note: The statements, opinions, and claims presented in this article are exclusively those of the authors and do not necessarily reflect the views of their affiliated institutions, the publisher, editors, or reviewers. Any products discussed or evaluated, as well as any claims made by their manufacturers, are not warranted, approved, or endorsed by the publisher.

Acknowledgements

We thank everyone who inspired our work.

How to Cite:

M Anjan Kumar , P Viswanatha Reddy , T Gopikrishna , MD Akbar , " Real-Time Facial Emotion Recognition Using Deep CNN and YOLO-Based Face Detection: A Hybrid CNN–LSTM Framework " , International Journal of Computer Science , Engineering and Artificial Intelligence , Vol. 2, Issue. 2 (April – June) , 2025 , Pages: 14 – 23.
DOI : https://doi.org/10.63328/IJCSEAI-V2RI2P3