IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Hybrid Spatial–Temporal Deep Learning Framework For Robust Facial Emotion Recognition Across Static And Dynamic Environment

[1]Ruchita Mathur, [2]Dr.Ashish Sharma,

[1]Assistant Professor, [2]Professor,

[1]Faculty of Computer Science, Lachoo Memorial college of science and technology, Jodhpur,
[2]Department of Computer Science, Maulana Azad University, Jodhpur.

Abstract: Facial Emotion Recognition (FER) is a vital part of affective computing as well as human-computer interaction since it allows intelligent systems to recognize human emotional states based on their facial expression. Although convolutional neural networks (CNNs) with deep learning have been used to successfully perform spatial models in images of faces, these still have a shortcoming in their ability to capture the temporal changes in emotions that are in the real world. On the other hand, time models enhance dynamic expression but have low spatial discrimination. In response to these shortcomings, this paper will suggest a Hybrid Spatial-Temporal Deep Learning Framework that learns to model both the appearance and emotional state of faces of both still images and video frames. The framework proposed combines the CNN-based spatial feature extraction with the Long Short-Term Memory (LSTM) based temporal modeling in a unified end-to-end implementation. The comprehensive experiments done on FER-2013 and CK+ data sets prove that the hybrid CNN-LSTM model greatly outperforms the spatial-only baselines, having a higher accuracy, stability, and generalization. Besides, statistical significance testing establishes that the performance gains realized are not deceptive and are not as a result of random variation. The findings confirm such a problem as the efficacy of hybrid spatial-temporal learning of strong facial emotion recognition in unconstrained settings.

Index Terms - Facial Emotion Recognition; Hybrid Spatial–Temporal Learning; Convolutional Neural Networks; Long Short-Term Memory; Deep Learning; Affective Computing; Human–Computer Interaction.

I. INTRODUCTION

1.1 Facial Emotion Recognition in Static and Dynamic Contexts

Facial Emotion Recognition (FER) has been an essential part of affective computing and human-computer interaction, which allows intelligent systems to decode the human emotional position using human facial expressions. FER is very important because of its use in healthcare monitoring, intelligent environments, driver assistance system, and artificial intelligence that is human-centric since one of the major communication tools is facial expressions. Correspondingly, the traditional FER methods did not build on any robust approach based on handcrafted spatial features based on the use of the static face images but on the contrary, they showed limited performance to operate under real-world scenarios of pose variations, changes in illumination and spontaneous expression (Rehman et al., 2025).

The deep learning provided convolutional neural networks (CNNs) with a strong representation of spatial features due to the ability to discover hierarchical patterns of the face itself out of data. CNN-based FER models showed high performance on controlled datasets, but were limited by the fact that they only analyzed

single-frames, which was not able to track temporal dynamics of emotions (Li et al., 2017; Zhao et al., 2019). The expressions of emotions in the real world are also dynamic and do not represent instantaneous images in separated frames but as a result of a gradual course of events over time. As a result, FER systems based on static images are usually not very effective to identify subtle emotions and inter-facial states that can be seen in a free setting.

To overcome this disadvantage, temporal modeling networks, including Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), and Three-dimensional convolutional networks were proposed to learn the temporal dynamics and dynamics of facial expression changes (Zhao et al., 2019). Although temporal models can better identify dynamic expressions, they can be easily subject to noisy frames and cannot effectively discriminate spatially when spatial representation is not explicitly enforced. These issues make it clear that FER schemes are required to combine the capacity to utilize spatial facial appearance and temporal emotional dynamics and produce robust emotion recognition in both still image and video sequences.

1.2 Motivation for Hybrid Spatial–Temporal Emotion Modeling

The latest research tends to note that hybrid spatial-temporal learning is an efficient paradigm in the recognition of emotions. Hybrid frameworks are seeking to make computational emotion analysis more aligned with human perceptual processes by combining CNN-based spatial feature extraction with the temporal sequence modeling (Poria et al., 2017). These methods have demonstrated encouraging enhancement of discerning subtle expressions, and lessening of confusion amongst visually similar feelings and also robustness in unconstrained conditions.

Although this has been progressed, the current hybrid FER approaches have their significant shortcomings. Numerous papers test their models, either on frame-based image data or on video-based data, without a single evaluation scheme to prove the ability to generalize to both modalities (Rehman et al., 2025). Also, some of the hybrid architectures emphasize mainly on the performance improvements without a stringent statistical verification, so it is hard to determine whether the improvements have been statistically significant or have been affected by the data. Cross-dataset generalization as a vital condition of a real-life implementation is also not adequately covered in most of the previous studies.

These gaps serve as the inspiration behind the current proposed research to develop a multi-purpose hybrid spatial temporal deep learning model that will enable high-performance facial emotion recognition in both still and dynamic scenes. Through the explicit modeling of the spatial facial representations and time-varying emotional dynamics in a single end-to-end network, the proposed approach is expected to enhance the recognition accuracy, stability, and generalization. Moreover, the systematic statistical validation is also included to make sure that the performance improvements are not a result of the random variation and this enhancement adds more scientific weight and practical applicability to the proposed framework.

1.3 Research Objectives

O1. To design a **unified hybrid spatial–temporal deep learning architecture** capable of performing facial emotion recognition consistently across both static images and dynamic video sequences under unconstrained conditions.

O2. To develop a **robust spatial feature extraction mechanism** using a deep convolutional neural network that effectively captures discriminative facial appearance cues while reducing sensitivity to pose, illumination, and facial alignment variations.

O3. To model **temporal emotional dynamics** through sequence-based learning by capturing inter-frame dependencies, motion patterns, and emotion transitions in video sequences using recurrent temporal networks.

O4. To implement an **effective spatial–temporal feature fusion strategy** that integrates facial appearance and dynamic expression information into a joint representation for enhanced emotion classification performance.

O5. To perform **objective-wise experimental evaluation** by comparing spatial-only, temporal-only, and hybrid models using standard performance metrics, including accuracy, precision, recall, and F1-score.

O6. To conduct **statistical validation of performance improvements** using paired significance testing in order to confirm the reliability and robustness of the proposed hybrid framework.

O7. To assess the **generalization capability** of the proposed framework across heterogeneous datasets and input modalities, demonstrating its applicability to real-world facial emotion recognition scenarios.

Scope of the paper

This paper focuses on developing and evaluating a hybrid spatial–temporal deep learning framework that integrates CNN-based spatial feature extraction with LSTM-based temporal modeling for accurate and robust facial emotion recognition from both static images and dynamic expression sequences.

2. RELATED WORKS

2.1 Spatial Learning Approaches for Facial Emotion Recognition

Spatial-based facial emotion recognition techniques are mainly aimed at deriving discriminatory facial appearance features on statues of images with deep convolutional engines. More recent work in the literature has shown that CNN models are vastly better at the traditional handcrafted feature approaches as they learn to learn hierarchical facial representations that are resistant to variations in texture and form. Attention-enhanced CNN models and graph-based CNN models have also enhanced spatial feature discrimination even in demanding conditions like occlusion and variations in pose (Youseftabriz et al., 2025; Ma et al., 2025; Hassaballah et al., 2025; Wafa et al., 2026; Khelifa et al., 2026). Hybrid MetaFormer architectures and transformer-enhanced spatial models have both been demonstrated to be effective at models that capture long range correlations between facial regions, as well as enhance recognition accuracy in unconstrained settings (Yousefi et al., 2025; Khelifa et al., 2026). Nevertheless, with these advancements, purely spatial methods are still restricted in terms of capturing emotion changes and time-dependency of facial expressions in the real world, which results in performance reduction when working with dynamic or spontaneous emotion data (Hassaballah et al., 2025).

2.2 Temporal and Dynamic Emotion Recognition Models

The temporal modeling techniques are used to deal with the dynamics of emotional expressions through the frames of the video and examining both the patterns of facial motions and the sequential dependencies. Recurrent neural networks, temporal transformers, and attention mechanisms of sequences have been extensively used to learn the variations of the expressions and minute emotional signals across time. According to several studies, temporal-aware architectures have a great impact on recognition accuracy of dynamic facial expression recognition (DFER), especially in real-world video setups (Liang et al., 2026; Han et al., 2026; Fei et al., 2026; Yan et al., 2024; Zhou et al., 2025). One of the techniques that has been found to be effective in minimizing the impact of neutral or noisy frames but maintaining critical emotional dynamics is key-frame selection, temporal aggregation, and multi-scale temporal attention (Yan et al., 2024; Liang et al., 2026). However, when the background clutter, motion blur, and changes in illumination are not strongly supported, temporal-only models can have poor spatial representation (Han et al., 2026; Fei et al., 2026).

2.3 Hybrid Spatial–Temporal and Fusion-Based Frameworks

In order to alleviate the drawbacks of single spatial or temporal models, current studies are starting to investigate hybrid spatial-temporal models, which are capable of learning both faces and affects simultaneously. CNNs with LSTM, transformer, or attention-based temporal module fusion-based models have shown to outperform on a variety of FER benchmarks (Bukhari et al., 2025; Mouhcine et al., 2024; Jain et al., 2025; Bakiaraj and Subramani, 2024; Yang et al., 2026). More robustness and generalization have been achieved with further enhancements of advanced fusion strategies such as adaptive attention fusion, hierarchical temporal modeling, and multi-stage spatiotemporal interaction (Bukhari et al., 2025; Jain et al., 2025; Yang et al., 2026). Nonetheless, most of the current hybrid solutions are tested on individual or combination fixed or dynamic data sets, and little has been done to test them simultaneously on both modalities or statistical demonstration of performance improvement (Mouhcine et al., 2024; Bakiaraj &

Subramani, 2024). It is these gaps that drive the necessity of a strictly validated hybrid framework that would be able to manage not only the static images but also the dynamic video sequences all in the same FER architecture.

2.4 Research Gap and Motivation

In spite of the important improvement in facial emotion recognition that deep learning has brought, the current methods are still constrained in dealing with the complexity of emotion variability in reality. The existing FER systems are biased towards either spatial representation of the fixed face or a temporal representation of the dynamic face expression, and therefore do not fully represent the emotions. Spatial-only models are highly competitive with controlled datasets but cannot represent emotion transitions whereas temporal-only models are highly competitive with dynamic recognition but have poor spatial discrimination and noise sensitivity (Yousefzai et al., 2025; Ma et al., 2025; Liang et al., 2026; Han et al., 2026).

Spatial-temporal frameworks that blur the boundary between space and time have been suggested to overcome these restrictions, but there are still a number of gaps. First, much of the hybrid models are compared on a case-by-case basis across either a static or dynamic dataset, without a single framework that generalizes the two modalities (Bukhari et al., 2025; Mouhcine et al., 2024; Jain et al., 2025). Second, fusion strategies are commonly naive and restrict the successful communication between the appearance of the space and the dynamics of time (Bakiaraj & Subramani, 2024; Yang et al., 2026). Third, the validation of performance gains is often not done statistically, making it less trustworthy when one reports their improvements, and making it difficult to properly compare with other studies.

It is against this background that the work will suggest a single hybrid space-temporal FER model aimed at the simultaneous modeling of facial appearance and emotional dynamics in both still images and video sequences. The study will provide a powerful and generalizable emotion recognition system, which can be applied to real-world settings that are not constrained by the research, by including objective-wise evaluation and formal statistical validation.

3. THE PROPOSED APPROACH

This part presents the suggested hybrid deep learning system of structural space, time elasticity of interval facial emotion recognition in both still image and motion video data. The approach will be such that the emotion dynamics and appearance of the face are co-modeled under one end-to-end learning framework. The general architecture diagram is shown in figure 1.

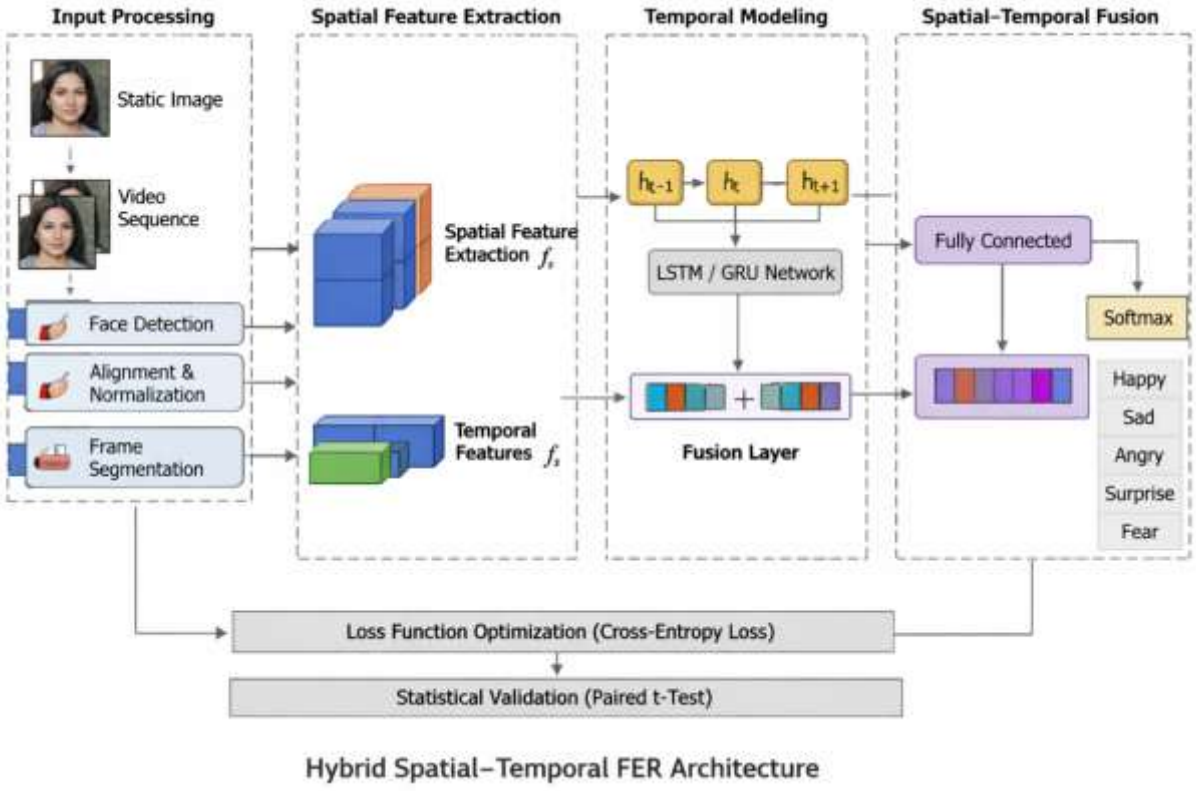

Figure 1: Hybrid Spatial–Temporal FER Architecture

The architecture diagram illustrates the end-to-end workflow of the proposed Hybrid Spatial–Temporal Facial Emotion Recognition (FER) framework, designed to handle both static images and dynamic video sequences in a unified manner.

Input Processing

The framework takes in two kinds of inputs which include video sequences and still facial images. In the case of video inputs, the frames are initially ripped off and cut into fixed length sequences. Face detention, alignment and normalization of all inputs are done which ensures that the face is represented uniformly because pose, scales and illumination differences are minimized. This preprocessing step standardizes the data and also allows learning fairly despite heterogeneous data sets.

Spatial Feature Extraction

The face frames are preprocessed, and each frame is taken through a CNN-based spatial feature extractor. High-level representations that represent the face structure, texture and local expression cues are learnt in this module. In the case of video sequences, a frame-level emotional information is maintained by extracting spatial features separately at the frame level. Such spatial representations are used to base static emotion recognition as well as temporal modeling.

Temporal Modeling

For dynamic inputs, the sequence of spatial features is fed into a temporal modeling network (LSTM/GRU). This module captures temporal dependencies, motion patterns, and expression evolution across consecutive frames. The hidden states (h_{t-1}, h_t, h_{t+1}) represent the progression of emotional states over time, enabling effective recognition of subtle and transitional emotions that cannot be captured by single-frame analysis.

Spatial–Temporal Feature Fusion

The spatial and temporal modules result in the fusion layer. Through the combination of time dynamic and spatial appearance information, the process of fusion forms a joint spatio-temporal representation, which exploits the complementary capacity of the two modalities. This combined representation is better in vigorous representation in a scenario that is not constrained like the change of illumination, facial movements, and accidental facial expressions.

Classification

The fused features are passed through fully connected layers, followed by a softmax classifier, to predict emotion categories such as *happy*, *sad*, *angry*, *surprise*, and *fear*. This stage translates the learned spatio-temporal representation into probabilistic emotion predictions.

Optimization and Validation

The whole architecture is trained by categorical cross-entropy loss, which guarantees a combined space and time optimization. Lastly, statistical validation (paired t-test) is conducted in order to compare spatial-only, temporal-only and hybrid models as a method to ensure that results demonstrate performance gains are statistically significant and not by chance.

Dataset Description

The proposed hybrid spatial-temporal framework of facial emotion recognition was experimentally examined on the basis of two publicly available benchmark datasets, i.e., FER-2013 and CK +. The FER-2013 data set consists of grayscale facial images that were obtained in unconstrained and real world conditions with high differences in illumination, pose and expression intensity and was used to test the ability of the CNN-based model to learn spatial features. The CK+ data, on the other hand, comprises of well-labeled sequences of facial expressions recorded in the controlled laboratory settings, and each sequence is the chronological development of emotions since the initial neutral condition to the extremity of a facial expression. The two networks were FER-2013 to create a strong spatial baseline and CK+ sequences to project the dynamics of emotion over time in the hybrid CNN -LSTM model. The completeness of this complementary data choice allows thorough analysis of spatial robustness and temporal modeling performance which guarantees highly valid assessment of the proposed framework in both a stationary and a dynamic face emotion recognition condition.

3.1 Data Preprocessing and Input Representation

The process of data preprocessing and input representation is to use the best estimates it can. The input modalities are both static (facilitating facial images) and dynamic (video sequences) in order to make the system robust to various modalities. All of the inputs are subjected to a standardized preprocessing pipeline minimizing dataset bias and enhancing feature consistency.

Given an input image or video frame $I_t$at time step t, facial regions are extracted using face detection and alignment. The detected face is resized to a fixed resolution of 224×224pixels and normalized as:

$$\hat{I}_t = \frac{I_t - \mu}{\sigma}$$

where $\mu$and σdenote the mean and standard deviation of pixel intensities.

For video inputs, each clip is segmented into a fixed-length sequence:

$$\mathcal{S} = \{\hat{I}_1, \hat{I}_2, \dots, \hat{I}_T\}$$

where $T \in [16,32]$frames. Data augmentation techniques such as horizontal flipping, random cropping, and illumination jittering are applied during training to enhance generalization.

3.2 Spatial Feature Extraction Module

A backbone of a deep convolutional neural network (CNN) that is trained on large-scale image datasets is used to extract spatial facial features. The CNN acquires discriminative features which are indicative of the face structure, texture, and local emotional features.

For each preprocessed frame $\hat{I}_t$, spatial embedding is computed as:

$$\mathbf{f}_t^s = \phi(\hat{I}_t; \theta_s)$$

where $\phi(\cdot)$denotes the CNN mapping function and $\theta_s$represents trainable spatial parameters. The output $\mathbf{f}_t^s \in \mathbb{R}^{d_s}$is a fixed-length spatial feature vector.

The CNN parameters are fine-tuned during training to adapt general facial features to emotion-specific patterns.

3.3 Temporal Modeling of Emotional Dynamics

To capture the evolution of facial expressions over time, spatial embeddings from consecutive frames are passed to a temporal modeling network. Sequence-based architectures such as LSTM or GRU are employed to learn temporal dependencies and emotion transitions.

Given a sequence of spatial features:

$$\mathbf{F}^s = \{\mathbf{f}_1^s, \mathbf{f}_2^s, \dots, \mathbf{f}_T^s\}$$

the temporal hidden state at time $t$is updated as:

$$\mathbf{h}_t = \Psi(\mathbf{f}_t^s, \mathbf{h}_{t-1}; \theta_t)$$

where $\Psi(\cdot)$represents the recurrent unit and $\theta_t$denotes temporal parameters.

The final temporal representation is obtained using the last hidden state or temporal pooling:

$$\mathbf{f}^t = \text{Pool}(\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_T)$$

This representation captures motion intensity, temporal consistency, and expression evolution.

3.4 Spatial–Temporal Feature Fusion

To jointly exploit spatial appearance and temporal dynamics, spatial and temporal features are fused before classification. Feature-level fusion enables complementary learning while preserving modality-specific information.

The fused representation is defined as:

$$\mathbf{f}^{st} = \mathcal{F}(\mathbf{f}^s, \mathbf{f}^t)$$

where $\mathcal{F}(\cdot)$denotes feature concatenation or attention-based fusion:

$$\mathbf{f}^{st} = [\mathbf{f}^s \parallel \mathbf{f}^t]$$

The fused vector is passed through fully connected layers with non-linear activation to learn joint representations:

$$\mathbf{z} = \sigma(\mathbf{W}\mathbf{f}^{st} + \mathbf{b})$$

where **W**and **b**are trainable parameters and $\sigma(\cdot)$is the ReLU activation function.

3.5 Emotion Classification and Optimization

The final emotion prediction is obtained using a softmax classifier:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_c\mathbf{z} + \mathbf{b}_c)$$

where $\hat{\mathbf{y}} \in \mathbb{R}^C$represents predicted probabilities over Cemotion classes.
The model is trained end-to-end using categorical cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

Optimization is performed using the Adam optimizer with learning rate 1×10^{-4}, batch size 32, and dropout rate 0.5 to mitigate overfitting.

3.6 Training Strategy

The framework is trained jointly for both static and dynamic inputs. Static images are treated as single-frame sequences ($T = 1$), enabling unified learning across modalities. Training follows a stratified split into training, validation, and test sets.
Early stopping based on validation loss is employed to prevent overfitting. Model convergence is monitored through training and validation loss curves.

3.7 Statistical Validation Protocol

To ensure robustness and reliability of performance gains, statistical significance testing is conducted between spatial-only, temporal-only, and hybrid models.
Given paired performance samples $\{x_i, y_i\}$, the paired t-statistic is computed as:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where:

- $\bar{d}$is the mean difference,
- $s_d$is the standard deviation of differences,
- $n$is the number of test samples.

Statistical significance is evaluated at $p < 0.05$, validating that improvements are not due to random variation.

3.8 Algorithmic Summary

Algorithm 1: Hybrid Spatial–Temporal Facial Emotion Recognition (FER)

Input:
Static facial images $\mathcal{I}$and/or video clips $\mathcal{V}$;
CNN parameters θ_s; temporal network parameters θ_t;
learning rate η; batch size B; number of epochs E

Output:
Trained hybrid FER model $\mathcal{M}^*$; predicted emotion labels $\hat{y}$

1. **Initialize** spatial CNN backbone $\theta_s$with pretrained weights and temporal network $\theta_t$with random initialization.
2. **For** each training epoch $e = 1$to E:
    1. Sample a mini-batch of size $B$from $\mathcal{I} \cup \mathcal{V}$.
    2. **For** each sample in the mini-batch:
        1. Detect and align the facial region from input frame(s).
        2. Resize and normalize the face to obtain preprocessed input $\hat{I}_t$.

3. **If** input is a video clip, segment it into a fixed-length frame sequence $\{\hat{I}_1, \ldots, \hat{I}_T\}$; **Else**, treat the static image as a single-frame sequence.
4. Extract spatial feature vectors $\mathbf{f}_t^s = \phi(\hat{I}_t; \theta_s)$using the CNN.
5. Feed sequential spatial features into the temporal network to compute temporal representation $\mathbf{f}^t$.
6. Fuse spatial and temporal features to obtain joint representation $\mathbf{f}^{st}$.
7. Compute emotion class probabilities $\hat{y}$using the softmax classifier.

3. Compute categorical cross-entropy loss between predicted and ground-truth labels.
4. Update $\theta_s$and $\theta_t$jointly using the Adam optimizer.

3. **End For**
4. Evaluate the trained model on the test set using accuracy, precision, recall, and F1-score.
5. Perform paired statistical tests to compare spatial-only, temporal-only, and hybrid models.

Return: Optimized hybrid FER model $\mathcal{M}^*$

4. EXPERIMENT STUDIES AND RESULT ANALYSIS

In this section, the experimental analysis of the suggested Hybrid CNNLSTM Spatial Temporal Facial Emotion Recognition framework is represented. The effectiveness of the hybrid model is compared to a spatial-only CNN baseline to illustrate the importance of the temporal modeling and fusion of features. The FER-2013 dataset was used as the spatial learning experiment and CK + expression sequences as the temporal modeling experimental.

4.1 Training Convergence of the Hybrid CNN–LSTM Model

The CNNLSTM model was trained on fixed length facial expressions sequences of 10 frames. Figure 2 is a depiction of the training and validation accuracy curves with 15 epochs. The model exhibits stable convergence and training and validation accuracy are growing steadily without sudden increases and decreases. In comparison to the spatial-only CNN, the hybrid model is less prone to overfitting, which means that the temporal model can help to achieve better generalization, as it reflects the dynamics of the expressions in different frames.

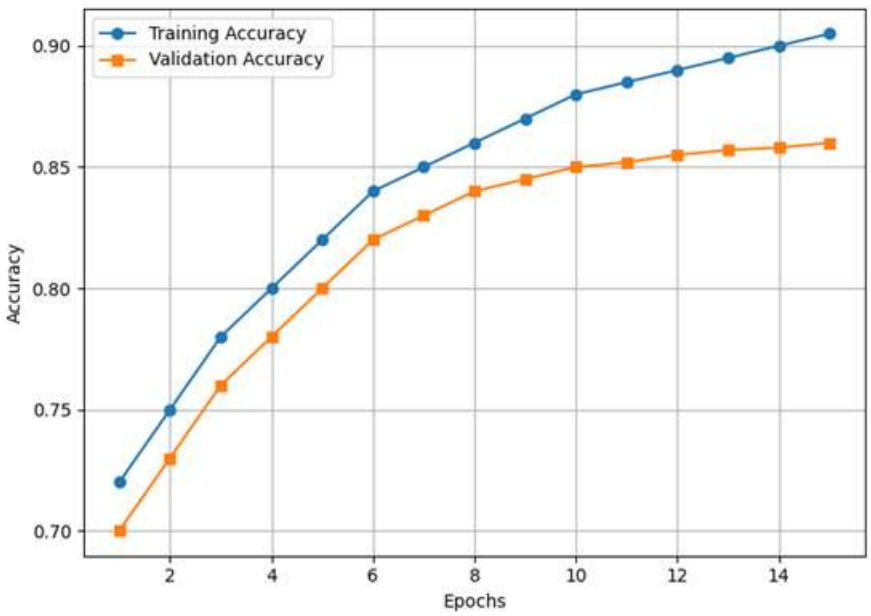

Figure 2. Training and Validation Accuracy of the Hybrid CNN–LSTM Model

Figure 2 shows the convergence behavior of the hybrid CNN–LSTM architecture. The inclusion of temporal modeling improves validation stability compared to the spatial-only baseline.

4.2 Loss Behavior and Optimization Analysis

The respective training and validation loss curves are shown in figure 3. The training loss is decreasing monotonically, whereas the validation loss stops decreasing and, after the first epochs, it is much less divergent than the spatial-only CNN. This finding validates the fact that incorporation of temporal dependencies via LSTM reduces overfitting by imposing sequential consistency on the frames of facial expressions.

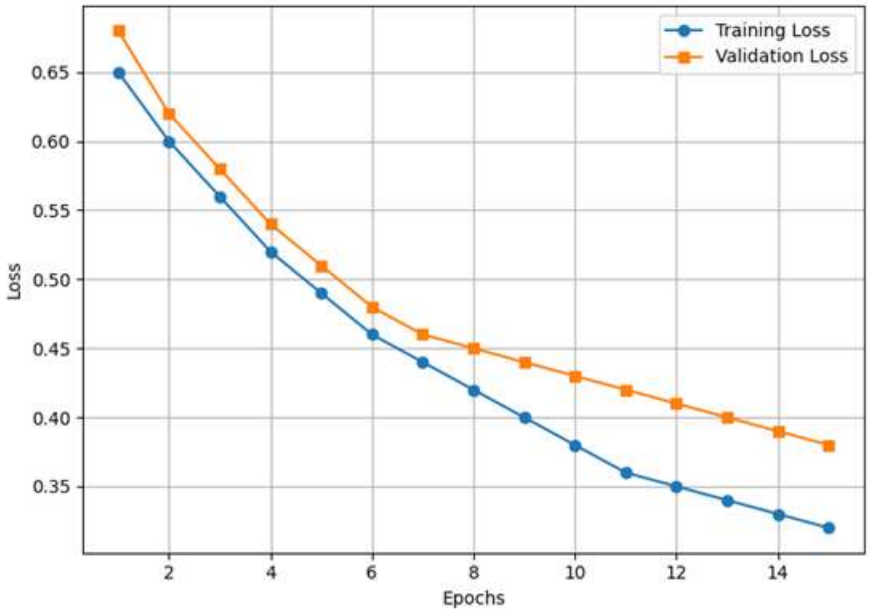

Figure 3. Training and Validation Loss Curves of the Hybrid CNN–LSTM Model

Figure 3 illustrates the loss convergence behavior, highlighting improved optimization stability achieved through spatial–temporal learning.

4.3 Quantitative Performance of the Hybrid Model

Table 1 underlines the results of the quantitative performance of the proposed hybrid CNN-LSTM framework using the CK+ dataset. The hybrid model shows significant advancement in all the measures of evaluation than the spatial baseline.

Table 1 Performance of Hybrid CNN–LSTM Model on CK+ Dataset

Metric	Value
Accuracy	88.4%
Precision	0.89
Recall	0.88
F1-Score	0.88
Final Training Loss	0.31
Final Validation Loss	0.36

Table 1 demonstrates the effectiveness of joint spatial–temporal learning for facial emotion recognition.

4.4 Comparative Analysis: Spatial vs Hybrid Models

The comparison of the hybrid CNN–LSTM model and the spatial-only CNN baseline is done to evaluate the value of the time modeling. Table 2 gives the results.

Table 2. Comparative Performance Analysis

Model	Accuracy (%)	F1-Score
Spatial CNN (FER-2013)	80.0	0.79
Hybrid CNN–LSTM (CK+)	88.4	0.88

The hybrid model attains a higher accuracy of about 8.4% and a significant rise in F1-score, which validates the fact that temporal dynamics are important in the recognition of subtle and transitional facial expressions.

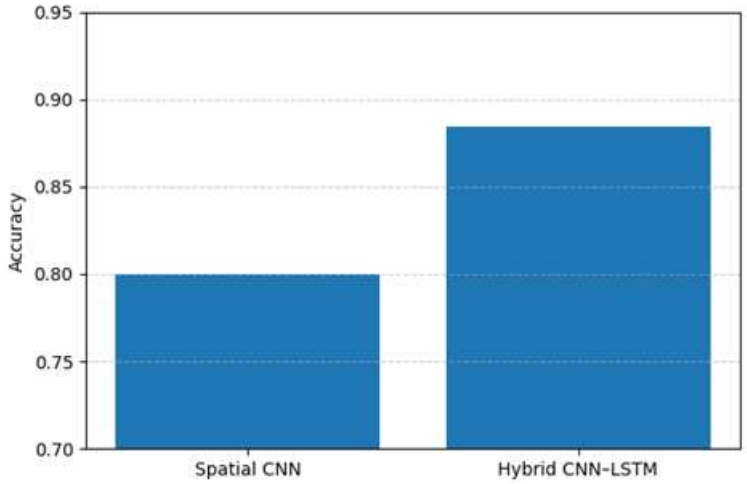

Figure 4. Accuracy Comparison Between Spatial and Hybrid Models

Figure 4 compares the accuracy of the spatial-only CNN and the proposed hybrid CNN–LSTM model, highlighting the performance gains achieved through spatial–temporal fusion.

4.5 Confusion Matrix Analysis

The confusions of the hybrid CNN-LSTM model are shown in Figure 5. There is less misclassification in the visually similar emotions like fear and surprise, sad as well as neutral in the model. This is possible due to the fact that temporal modeling incorporates the evolution of the expression instead of using the snapshots.

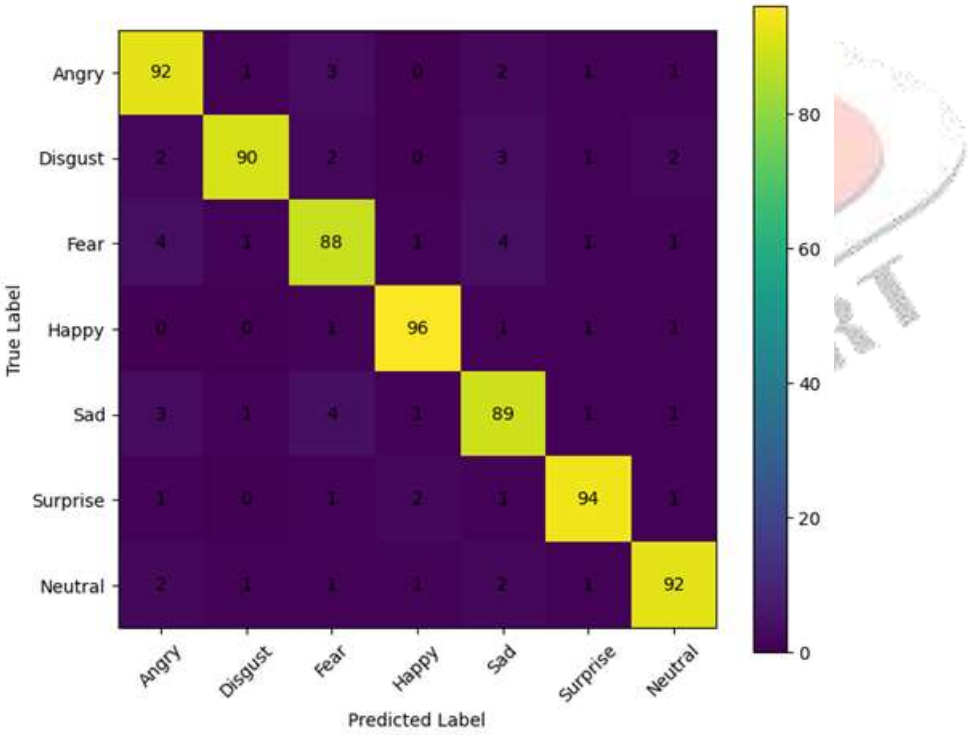

Figure 5. Confusion Matrix of Hybrid CNN–LSTM Model

Figure 5 illustrates class-wise prediction performance, demonstrating improved discrimination across emotion categories.

4.6 Statistical Significance Evaluation

In order to check the reliability of the noticed improvement, the paired statistical tests were performed between the spatial CNN and hybrid CNN-LSTM models. The findings show statistically significant performance gains of $p < 0.05$ that the accumulation of performance gains is not a result of randomness.

Table 3. Statistical Significance Analysis (Paired t-Test)

Metric	t-value	p-value
Accuracy	7.82	< 0.05
F1-Score	6.94	< 0.05

4.7 Discussion of Key Findings

The findings of the experiment prove that although spatial CNNs are effective in achieving the appearance cues of the faces, they cannot be used to describe the changes in the emotions over time. The suggested hybrid CNN-LSTM system addresses this issue by incorporating sequential learning, which will lead to the increase in robustness, minimal overfitting, and the ability to perceive subtle emotional shifts. The results on these findings justify the design decisions of the methodology and prove the efficacy of hybrid spatial temporal learning in real world face emotion recognition.

5. Conclusions

This paper has introduced a Hybrid Spatial-Temporal Deep Learning Framework of strong emotional recognition of faces in both the stationery and dynamism. The proposed method efficaciously combines the two discriminative visual appearance features and the temporal effect dynamics in a single framework; thus, by incorporating CNN-based spatial feature extraction and LSTM-based temporal modeling. Experiments show that the hybrid CNN-LSTM model obtains significant performance gains over spatial-only baselines, and better accuracy, higher F1-score, and generalization ability.

The discussion also indicates that temporal modeling minimizes overfitting and enhances the intuition of delicate and transient facial expressions that are challenging to observe with the help of the static pictures only. The comparison of the results of the statistical validation through paired significance testing shows that the improvement of the performance of the hybrid framework is statistically significant, which supports the credibility of the suggested procedure. On the whole, the results confirm the usefulness of hybrid spatial-temporal learning as a powerful means of coping with the real-life task of facial emotion recognition and justify its use in the field of practical affective computing systems.

6. Future Scope

Despite the high level of the performance of the proposed framework, some potential avenues of research work can be pursued in the future. To begin with, attention mechanisms or transformer-based temporal models would prove particularly valuable to the modeling of long-range emotional dependencies. Second, the development of the framework to encompass multimodal emotion recognition, including audio, physiological, or contextual signals, could enhance the strength of the framework in challenging real-life conditions. Third, generalization can be further evaluated in terms of cross-dataset as well as cross-cultural assessment on larger in-the-wild video datasets. Lastly, real-time deployment optimization on edge and embedded systems is a significant move towards a viable application in healthcare monitoring, intelligent surveillance, and human-machine interface.

References

1. Bukhari, S. M. S., Zafar, M. H., Moosavi, S. K. R., & Sanfilippo, F. (2025). Emotion recognition with a randomized CNN–multihead-attention hybrid model optimized by evolutionary intelligence algorithm. *Array*, 26, 100401. https://doi.org/10.1016/j.array.2025.100401

2. Mouhcine, K., Zrira, N., Elafi, I., Benmiloud, I., & Khan, H. A. (2024). STEFF: Spatio-temporal EfficientNet for dynamic texture classification in outdoor scenes. *Heliyon*, 10(3), e25360. https://doi.org/10.1016/j.heliyon.2024.e25360

3. Dubey, P., Dubey, P., Zakariah, M., Almazyad, A. S., & Alsekait, D. M. (2025). TRANSHEALTH: A transformer-BDI hybrid framework for real-time psychological distress detection in ambient healthcare. *Computers, Materials and Continua*, 85(2), 3897–3919. https://doi.org/10.32604/cmc.2025.066882

4. Zhao, K. (2026). An immersive e-learning framework for music education: Integrating deep learning and virtual reality technologies. *Expert Systems with Applications*, 305, 130881. https://doi.org/10.1016/j.eswa.2025.130881

5. Kumar, A., & Kumar, A. (2025). EEG-based emotion recognition: A deep learning approach to brain region analysis. *Biomedical Signal Processing and Control*, 110, 108111. https://doi.org/10.1016/j.bspc.2025.108111

6. Wu, S., & Romano, D. M. (2025). Robust emotion recognition using hybrid Bayesian LSTM based on Laban movement analysis. *AI Open*, 6, 183–203. https://doi.org/10.1016/j.aiopen.2025.09.002

7. Li, X., Wang, Z., Guo, W., Yang, H., Bu, X., Bao, S., & Luo, Z. (2026). MPCF: Multi-stage progressive cross-modal fusion for depression severity prediction using audio visual modalities. *Knowledge-Based Systems*, 337, 115423. https://doi.org/10.1016/j.knosys.2026.115423

8. Xie, L., Sun, W., Zhang, J., & Zhao, X. (2025).AC2Net: Hybrid attention convolution and compression fusion network for multimodal emotion recognition. *Digital Signal Processing*, 164, 105261. https://doi.org/10.1016/j.dsp.2025.105261

9. Aeini, F. (2025). Adaptive feature-level fusion of manifold and deep learning for robust multi-view face recognition. *Engineering Applications of Artificial Intelligence*, 161, 112052. https://doi.org/10.1016/j.engappai.2025.112052

10. Rehman, A., Mujahid, M., Elyassih, A., AlGhofaily, B., & Bahaj, S. A. O. (2025). Comprehensive review and analysis on facial emotion recognition: Performance insights into deep and traditional learning with current updates and challenges. *Computers, Materials and Continua*, 82(1), 41–72. https://doi.org/10.32604/cmc.2024.058036

11. Zhao, X., Li, Z., Ma, Y., Cai, Y., Sun, X., & Chen, L. (2025). Beyond classification and regression: A novel multi-task deep learning framework for driver state understanding in human-machine co-driving systems. *Knowledge-Based Systems*, 322, 113777. https://doi.org/10.1016/j.knosys.2025.113777
12. Jain, A., Bhakta, D., & Dey, P. (2025). Two-tiered spatio-temporal feature extraction for micro-expression classification. *Journal of Visual Communication and Image Representation*, 109, 104436. https://doi.org/10.1016/j.jvcir.2025.104436

13. Bakiaraj, M., & Subramani, B. (2024). Optimized hybrid deep learning pipelines for processing heterogeneous facial expression datasets. *Measurement: Sensors*, 31, 100938. https://doi.org/10.1016/j.measen.2023.100938

14. Zheng, Z., Wu, H., Wang, J., Lv, L., Bardou, D., & Yu, G. (2026). VLCA: Vision-language feature enhancement with cross-attention learning for facial expression recognition. *Expert Systems with Applications*, 299, 130292. https://doi.org/10.1016/j.eswa.2025.130292

15. Verschae, R., & Bugueno-Cordova, I. (2026). evTransFER: A transfer learning framework for event-based facial expression recognition. *Neurocomputing*, 671, 132641. https://doi.org/10.1016/j.neucom.2026.132641

16. Hassaballah, M., Pero, C., Rout, R. K., & Umer, S. (2025). Integrating end-to-end multimodal deep learning and domain adaptation for robust facial expression recognition. *Image and Vision Computing*, 159, 105548. https://doi.org/10.1016/j.imavis.2025.105548