International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)

ISSN: 2456-236X

Vol. 10 Issue 01 | 2025

Capturing Subtle Emotions: A Novel Parallel CNN-LSTM Architecture for Micro-expression Recognition

Vibha

Assistant professor, Computer Science and Engineering Department,
Padm. Dr. V. B. Kolte College of Engineering, Malkapur , India

DOI: 10.5281/zenodo.15751631

ABSTRACT

This research introduces a new parallel CNN-LSTM model that is intended to improve micro-expression (MiE) recognition by simultaneously learning spatial and temporal features. It employs three benchmark databases—SMIC, CASME II, and SAMM—to alleviate difficulties from subtle, short facial signals through combining convolutional feature extraction with temporal sequence learning. Preprocessing normalizes input data via resizing, normalization, and padding, and a leave-one-subject-out cross-validation (CV) is used to ensure robustness and generalizability. Experimental outcomes show improved accuracy on all datasets over state-of-the-art (SoA) tactics, confirming the parallel approach's effectiveness in detecting fast, low-amplitude facial expressions. The structure improves automated emotion detection and has applications for security, healthcare, and human-computer interaction, revealing the versatility of deep learning in subtle affective computing.

***Keyword:* Parallel CNN and LSTM, microexpression recognition, neural network**

1. INTRODUCTION

Facial gestures play a vital role in everyday social interaction. Through facial expressions, people convey their own emotions and interpret the feelings of those around them.

Facial expressions are generally categorised into two types: macro-expressions (MaEs) and MiEs. MaEs are more visible and last longer, whereas MiEs are fleeting, usually lasting less than half a second, and involve subtle facial muscle movements.

The concept of MiEs was first introduced by Ekman and Friesen[1] and has since gained widespread acceptance in the research community. As shown in Figure 1.1, there are seven universal MiEs, each linked to specific action units (AUs) that represent distinct facial muscle movements. MiEs typically appear when a person attempts to hide or control their emotions.

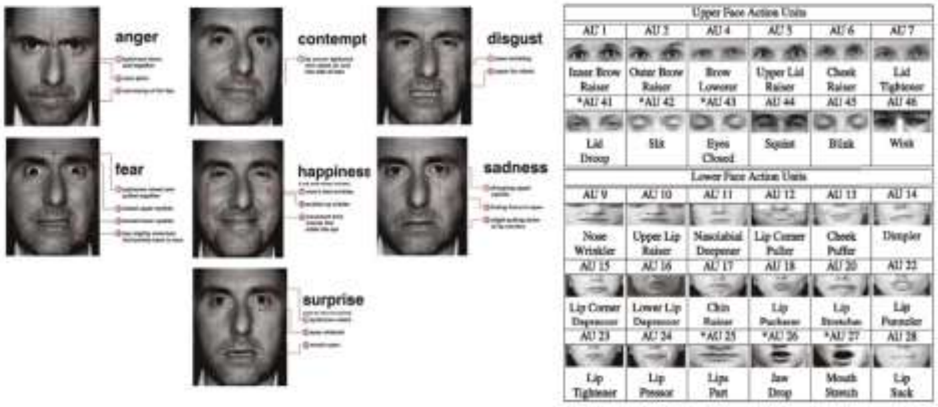
According to [1], there are three main techniques people use to mask their facial expressions **[1].**

(1) Simulated expressions: facial expressions are intentionally produced to convey an emotion, even when the person isn't actually feeling it.

(2) Neutralised expressions: a person maintains a neutral facial appearance despite experiencing a specific emotion.

(3) Concealed expressions: genuine emotions are hidden behind a different expression to mask the true feeling.

A MiE results from subtle disruptions in facial muscles, often occurring when an individual suppresses, conceals, or masks an emotional reaction **[2].**



Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink
Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

Fig-1.1: Seven universal facial MiEs

Some individuals lie primarily to enhance social interactions and gain respect or affection from others. However, detecting lies with harmful intent is particularly important. Liars often fail to fully suppress facial cues, and brief involuntary facial expressions—known as MiEs—can reveal deception. These expressions have promising applications in (1) high-threat settings, counting criminal investigations, airport along with transit checkpoints, and counter-terrorism operations; (2) professional environments like marketing, education, consulting, and administration, and leadership, and trade negotiations; and (3) healthcare services, including physician–patient consultations [3]. Some professionals have learned to detect MiEs with the naked eye through dedicated training tools. Due to the difficulties of manual detection, the demand for automated MiE recognition has increased significantly. While computer vision has progressed considerably, recognising MiEs remains a challenge because of their subtlety and fleeting nature. Research in this field aims to detect and classify these expressions, and numerous approaches have been explored over the past decade [4].

MiEs occur within the natural flow of facial expressions, particularly when a person attempts to obscure their accurate emotions. Lin et al. (2022) noted that various factors can influence how effectively these fleeting expressions are recognized.

1.1 Emotional Context

Previous research has utilised neutral facial expressions both preceding and following emotional ones. Findings indicate that MiEs might either be embedded within these neutral states or occur alongside emotions such as happiness or sadness. Based on emotional regulation theory, longer exposure to emotional primes during a priming task may enhance the priming effect. Additionally, emotional cues have been shown to significantly impact attentional processes [6].

1.2 Duration of Expression

The main difference between MiEs and MaEs is how long they last. Various studies have tried to define exactly how long a MiE lasts, but there's no clear consensus on the precise timeframe. Although the time difference may be subtle, it's important to consider when studying MiEs. To test how duration affects the ability to recognise MiEs, researchers ran two experiments. In the first experiment, participants viewed images of expressions displayed for 40, and 120, and 200, or even 300 milliseconds and completed the Brief Affect Recognition Test (BART). Inside the second, participants underwent training using the MiE Training Tool (METT), which notably improved their recognition skills. Results showed that without training, participants could identify MiEs shown for 200 milliseconds, but after training, this improved to 160 milliseconds. This suggests that MiEs generally last about 200 milliseconds or less, and recognition accuracy depends on how long the expression is shown.

1.3 Challenges

Environmental factor

One of the most difficult challenges in studying MiEs is dealing with environmental variability, which includes fluctuations in both lighting and head position. Changes in illumination primarily affect the pixel intensity values, causing shifts that complicate feature extraction. Dynamic lighting conditions can lead to incorrect evaluation of features, and head movements or adjustments may be mistakenly identified as MiEs. Even slight motions of the head can significantly influence the way expressions are perceived, which ultimately reduces the accuracy of detection methods.

1.4 Spontaneous and subtle motion of the facial movement

A major challenge in recognizing MiEs lies in their faint, subtle, and involuntary nature, which results in facial movements that are difficult to detect with the naked eye. In some cases, the classifier may even mistake these subtle motions for a neutral expression. Therefore, implementing methods to amplify and enhance these slight emotional cues is crucial during the pre-processing stage.

Datasets such as SMIC, and CASME, and CASME II, along with CAS(ME) [7] represent a few publicly available collections dedicated to MiE recognition. Although these datasets are often recommended for evaluating MiE recognition systems, their uneven distribution of samples across different expression classes can introduce biases affecting the performance results. Moreover, the data within these datasets is usually gathered under controlled environments, featuring consistent lighting and fixed camera setups. As a consequence, SoA algorithms tested on these datasets may not perform optimally in real-world scenarios, highlighting the need for more diverse and realistic datasets that better reflect practical conditions

2. LITERATURE REVIEW

Table -1: Literature Review

Author(s)	Method	Objective	Finding
Kim et al. (2022) [8]	Combined CNN along with LSTM in sequence: CNN extracts spatial features; LSTM captures temporal info	To improve MiE recognition by modeling spatial temporal features	Improved accuracy compared to LBP-based methods; temporal modeling enhanced recognition.
Peng et al. (2017) [9]	Dual Temporal Scale CNN processing optical flow sequences at different frame rates	To capture MiE dynamics at multiple temporal scales	Effective feature extraction but limited by insufficient capture of fine-grained micro-scale features.
Liang et al. (2020) [10]	Deep convolutional BiLSTM fusion network combining two CNNs and a BiLSTM	To learn discriminative spatial features and temporal correlations	BiLSTM improved long-term temporal context modeling; outperformed traditional CNNs.
Hashmi et al. (2021) [11]	LARNet with lossless attention residual network for real-time MiE detection	To achieve high accuracy in realtime facial MiE detection	High accuracy under constrained conditions; performance sensitive to face quality and distance.
Wu & Guo (2021) [12]	Three-stream CNN combining 2D and 3D CNNs for MiE feature extraction	To capture spatial and spatiotemporal features from facial videos	Multi-stream approach enhanced feature richness and classification performance.
Ma et al. (2019) [13]	Multi-layer feature fusion CNN with fusion module to combine features of different scales	To improve feature integration across CNN layers	Fusion of multi-scale features improved robustness and recognition accuracy.

3. OBJECTIVE

To implement Parallel CNN and LSTM network with SGD optimizer to improved performance and accuracy of the model.

4. PROBLEM IDENTIFICATION

Although deep learning has improved MiE recognition, most models deal with spatial and temporal features sequentially and hence cannot completely identify subtle, fleeting facial cues. Parallel structures that concurrently extract spatial and temporal knowledge are unexplored. Moreover, available datasets are small and imbalanced, thus limiting model generalizability. Current fusion mechanisms can be inefficient, making them unpractical for real-time usage, particularly under diverse recording conditions. As such, there is an evident requirement for a powerful parallel CNN-LSTM architecture that effectively combines spatial and temporal features in order to enhance the accuracy, computational power, and generalizability across varied datasets and real-world applications.

5. METHODOLOGY

The current research proposes to create a strong MiE recognition framework grounded in the merits of deep learning architectures, in particular, a parallel Convolutional Neural Network (CNN) along with Long Short-Term Memory (LSTM) network, to efficiently extract both spatial along with temporal characteristics inherent in subtle facial MiEs. The methodology is built upon three popularly accepted datasets and includes an extensive preprocessing pipeline, a new hybrid model architecture, and stringent evaluation with a leave-one-subject-out validation procedure.

5.1 Dataset Accumulation

Any machine learning model's success depends substantially on the representativeness and quality of training data. For guaranteeing a wide and representative sample of MiEs, three benchmark datasets are used in the research:

- **SMIC (Spontaneous MiE Corpus)**: SMIC consists of videos from 20 subjects recorded simultaneously through three cameras: High-Speed (HS) PixeLINK PL-B774U, Near Infrared (NIR), and Normal Vision (VIS). Footage is recorded by each camera at 640 × 480 pixels and 25 frames per second. There are 164 HS samples and 71 samples of both VIS and NIR. MiEs are labeled broadly into positive, negative, and surprise classes, recording various emotional responses.
- **CASME II (Chinese Academy of Sciences MiE II):** CASME II is a high-fidelity dataset containing 247 MiE samples of 18 subjects collected with a Point Grey GRAS-03K2C camera at 640 × 480 RAW 8-bit resolution and 200 frames per second. Participants were requested to keep neutral expressions or resist emotions stimulated by video clips. Emotions were categorized into seven classes, counting disgust, and happiness, and surprise, and repression, along with others, which represent an extensive affective spectrum.
- **SAMM (Spontaneous Actions and Micro-Movements)**: There are 159 micro-movements from 32 subjects in this dataset, captured with a Basler Ace acA2000-340km camera with a greyscale sensor at resolution 2040 × 1088 and 200 frames per second. The dataset is meticulously lit to minimize flicker artifacts with arrays of LED lights. Emotional classes cover seven categories: anger, sadness, contempt, disgust, and fear, and happiness, along with surprise.

These datasets together offer an inclusive and diverse base, recording spontaneous MiEs with varying recording conditions and emotional classes, thus making it possible to conduct rigorous model training and testing.

5.2 Data Pre-processing

Raw image data needs to be very carefully preprocessed in order to maintain uniformity and improve performance for deep learning models. The preprocessing operations that are used are:

- **Image Reading and Resizing:** Video frames are all read and standardized to 48 × 48 pixels. The size is a good trade-off between having enough spatial information and being computationally efficient.
- **Data Organization:** Images, their respective emotion labels, and subject IDs are stored methodically in different lists, ensuring correspondence for supervised training.
- **Label Encoding:** Emotion labels, initially in string form, are mapped to integer indices with encoding methods for compatibility with classification problems.
- **One-Hot Encoding:** Integer labels are encoded as one-hot vectors with categorical encoding functions for support of multi-class classification with neural networks.
- **Normalization:** Pixel intensity values are normalized to the interval [0,1] by dividing by 255, normalizing the input and speeding up model convergence.
- **Sequence Padding:** To balance different video lengths, sequences are padded or truncated to a constant length of 48 frames with zero-padding to have equal input dimensions for the LSTM component.

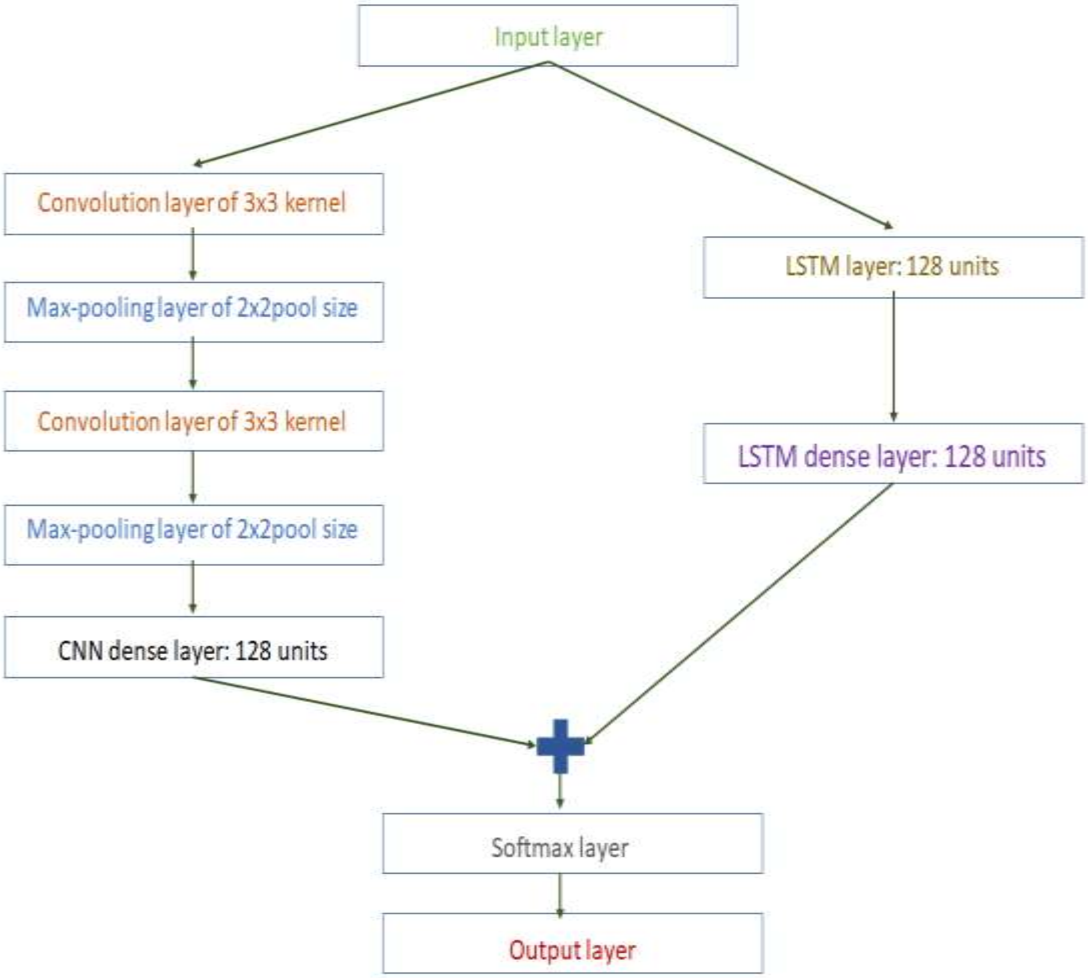

Input layer
Convolution layer of 3x3 kernel
Max-pooling layer of 2x2pool size
Convolution layer of 3x3 kernel
Max-pooling layer of 2x2pool size
CNN dense layer: 128 units
LSTM layer: 128 units
LSTM dense layer: 128 units
Softmax layer
Output layer

Fig-2: CNN+LSTM Architecture

5.3 Implementation

The model is compiled utilising the Stochastic Gradient Descent (SGD) optimizer using learning rate - 0.01 and momentum - 0.9, optimized against the categorical cross-entropy loss function, appropriate for multi-class classification. Training is performed over 100 epochs having batch size - 25.

5.4 Validation Strategy: Leave-One-Subject-Out (LOSO)

To prevent data leakage and ensure robust evaluation, LOSO CV is employed. In each iteration, the model trains on all subjects except one, which is reserved for testing. This is repeated for each subject, and metrics are averaged, ensuring model generalizability to unseen subjects and mitigating bias from subject-specific features.

5.5 Performance Metrics and Evaluation

The model's working is assessed utilising several metrics derived from the confusion matrix framework:

- **True Positives (TP):** Accurate predictions of positive cases.
- **True Negatives (TN):** Accurate predictions of negative cases.

- **False Positives (FP):** Incorrect positive predictions.
- **False Negatives (FN):** Incorrect negative predictions.

From these, key metrics are calculated:

- Accuracy: Proportion of total accurate predictions.

$$\text{Accuracy } = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision: The proportion of accurately predicted positive cases out of all instances predicted as positive, reflecting the model's effectiveness in minimizing false positives.

$$\text{Precision } = \frac{TP}{TP + FP}$$

- Recall: Ratio of accurately detected positives to all actual positives, reflecting the model's sensitivity.

$$\text{Recall } = \frac{TP}{TP + FN}$$

- F1-Score: Harmonic mean of precision alongside recall, balancing both false positives and negatives.

$$F1 = 2 \times \frac{\text{Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}}$$

These metrics provide a comprehensive assessment of model working, especially valuable for imbalanced datasets such as MiEs.

6. RESULTS

The experimental findings of our suggested methodology for MiE recognition will be detailed in this chapter. Using the datasets with different features, various experiments are undertaken to test the performance and validation of the generated model. Performance measures are used to estimate the results as well as to compare the accuracy of the algorithms.

Comparison of accuracy with three class datasets are analysed in Table 2 with other stste-of-the-art work.

Table 2: Comparison of accuracy with three class

Method (Bold: Best performance of each method at each column)		Training and Testing Set		
		CASME II (145)	SAMM (133)	SMIC (164)
		Accuracy	Accuracy	Accuracy
LBP-TOP (on temporal frames)	$\{3,3,3,4\}_{\text{TIM}} = 10$	0.448	0.519	0.274
	{3,3,3,8} T|M=10	0.441	0.511	0.262
	(3,3,3,4) TIM=64	0.414	0.466	0.299
	(3,3,3,8) TIM=64	0.421	0.444	0.287
HOOF (on temporal frames)	$\mathrm{N} = 4, \mathrm{T} \mid \mathrm{M} = 10$	0.386	0.451	0.36
	$\mathrm{N} = 8, \mathrm{T} \mid \mathrm{M} = 10$	0.434	0.436	0.421

	$N = 4, T \mid M = 64$	0.497	0.586	0.366
	$N = 8, T \mid M = 64$	0.559	0.602	0.299
Micro-Attention [19] (on apex frames)		0.703	0.677	0.482
ATNet	Spatial Stream	0.669	0.684	0.457
	Temporal Stream	0.724	0.632	0.53
	Fusion	0.834	0.701	0.561
Proposed Model	CNN+LSTM	0.981	0.829	0.914

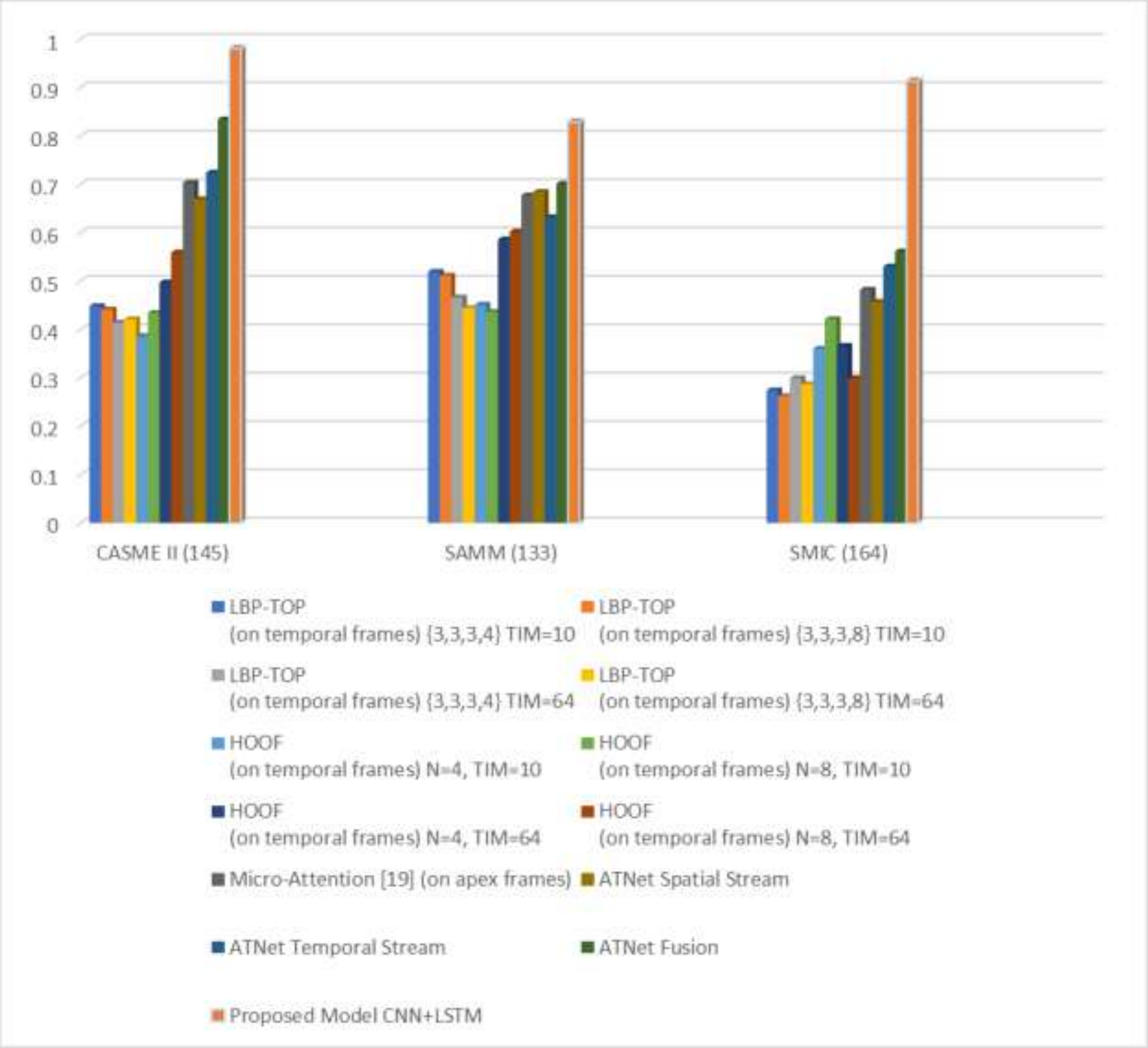


Fig-3: CNN+LSTM Architecture

7. CONCLUSIONS

The proposed parallel CNN-LSTM framework significantly improves MiE recognition accuracy by concurrently leveraging spatial and temporal information, overcoming limitations of sequential feature processing. The model's superior performance across diverse datasets underlines its robustness in handling the subtlety and brevity inherent in MiEs. Employing LOSO validation strengthens its generalizability to unseen subjects, crucial for real-world application. Despite advancements, challenges such as environmental variability and dataset imbalance remain, motivating future research toward more adaptive models and extensive real-world datasets. This work contributes a scalable, computationally efficient approach, enhancing automated detection capabilities essential for domains requiring high precision in emotional inference, including security and medical diagnostics.

6. REFERENCES

[1] B. Vukovic, "Verbal, Nonverbal, and Physiological Cues to Deception," 2024.

[2] M. L. Patterson, A. J. Fridlund, and C. Crivelli, "Four misconceptions about nonverbal communication," Perspectives on Psychological Science, vol. 18, no. 6, pp. 1388–1411, 2023.

[3] S. Cen, Y. Yu, G. Yan, M. Yu, and Y. Guo, "Multi-task facial activity patterns learning for MiE recognition using joint temporal local cube binary pattern," Signal Processing: Image Communication, vol. 103, p. 116616, 2022.

[4] H. X. Xie, L. Lo, H. H. Shuai, and W. H. Cheng, "An overview of facial MiE analysis: Data, methodology and challenge," IEEE Transactions on Affective Computing, vol. 14, no. 3, pp. 1857–1875, 2022.

[5] Q. Lin, Z. Dong, Q. Zheng, and S. J. Wang, "The effect of facial attractiveness on MiE recognition," Frontiers in Psychology, vol. 13, p. 959124, 2022.

[6] J. Li et al., "CAS (ME) 3: A third generation facial spontaneous MiE database with depth information and high ecological validity," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 3, pp. 2782–2800, 2022.

[7] S. Agarwal and D. P. Mukherjee, "Facial expression recognition through adaptive learning of local motion descriptor," Multimedia Tools and Applications, vol. 76, pp. 1073–1099, 2017.

[8] J. C. Kim, M. H. Kim, H. E. Suh, M. T. Naseem, and C. S. Lee, "Hybrid approach for facial expression recognition using convolutional neural networks and SVM," Applied Sciences, vol. 12, no. 11, p. 5493, 2022.

[9] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual temporal scale convolutional neural network for MiE recognition," Frontiers in Psychology, vol. 8, p. 1745, 2017.

[10] D. Liang, H. Liang, Z. Yu, and Y. Zhang, "Deep convolutional BiLSTM fusion network for facial expression recognition," The Visual Computer, vol. 36, pp. 499–508, 2020.

[11] M. F. Hashmi et al., "LARNet: Real-time detection of facial micro expression using lossless attention residual network," Sensors, vol. 21, no. 4, p. 1098, 2021.

[12] C. Wu and F. Guo, "TSNN: three-stream combining 2D and 3D convolutional neural network for micro-expression recognition," IEEJ Transactions on Electrical and Electronic Engineering, vol. 16, no. 1, pp. 98–107, 2021.

[13] C. Ma, X. Mu, and D. Sha, "Multi-layers feature fusion of convolutional neural network for scene classification of remote sensing," IEEE Access, vol. 7, pp. 121685–121694, 2019.