SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Advanced Facial Expression Recognition using Unified Multimodal LSTM Framework with Position Attention and Hybrid Grey Wolf–Ant Lion Optimization

Jyoti[1], Vishwanath P[2], Sangamesh H[3]

[1]Department of Electronics and Communication Engineering, Sir M Visvesvaraya College of Engineering, Raichur, Affiliated to VTU, Belagavi, Karnataka, India. Email: jo.jyotimailme@gmail.com[1]
[2]Department of Electronics and Communication Engineering, Sir M Visvesvaraya College of Engineering, Raichur, Affiliated to VTU, Belagavi, Karnataka, India. Email: vishalpetli73@gmail.com
[3]Department of Electronics and Communication Engineering, Sir M Visvesvaraya College of Engineering, Raichur, Affiliated to VTU, Belagavi, Karnataka, India. Email: sangamesh.hosgurmath@gmail.com

Abstract
Facial Expression Recognition (FER) is a cornerstone capability in human-computer interaction, affective computing, and intelligent systems. Despite significant progress, two persistent challenges remain unresolved: (i) effective integration of multimodal data streams encompassing audio, video, and text, and (ii) robust feature selection and classification under high-dimensional conditions. This paper presents a unified investigation of two complementary FER frameworks. The first framework proposes Long Short-Term Memory with a Position Attention Mechanism (LSTM-PAM) for multimodal FER, leveraging the Surrey Audio-Visual Expressed Emotion (SAVEE) and CMU Multimodal Opinion Sentiment Intensity (CMU-MOSI) datasets. Preprocessing employs Wiener filtering for audio, Contrast-Limited Adaptive Histogram Equalization (CLAHE) for video, and lemmatization for text. Feature extraction utilizes Mel-Frequency Cepstral Coefficients (MFCC), Gray-Level Co-occurrence Matrix (GLCM), and Word2Vector representations, fused into a unified feature vector and classified with LSTM-PAM. The second framework introduces a hybrid Grey Wolf Optimization and Ant Lion Optimization algorithm (GWALO) combined with a Recurrent Neural Network (RNN) for single-modality FER on the JAFFE dataset, employing Viola-Jones face detection, Gabor filtering, and Affine-Scale-Invariant Feature Transform (ASIFT). Comprehensive experiments demonstrate that LSTM-PAM achieves 98.25% and 95.78% accuracy on SAVEE and CMU-MOSI datasets, while GWALO-RNN achieves 99.7% accuracy on JAFFE, outperforming all baseline and state-of-the-art methods. This synthesis provides researchers a holistic perspective on advancing FER through complementary deep learning and hybrid optimization strategies.

Keywords: Facial expression recognition; Long Short-Term Memory; Position Attention Mechanism; Grey Wolf Optimization; Ant Lion Optimization; multimodal learning; deep learning; hybrid optimization.

1. Introduction

Facial Expression Recognition (FER) is one of the most actively researched domains in computer vision and affective computing, having evolved substantially over the past two decades [1]. Its applications span healthcare diagnostics, human-computer interaction (HCI), autonomous driving, mental health monitoring, and security surveillance [2]. The human face conveys a rich spectrum of emotional states through coordinated muscular movements, and the automated recognition of these states is fundamental to building emotionally intelligent machines [3].

Two major paradigms have emerged in FER research. The first paradigm addresses unimodal FER recognizing expressions from a single signal such as still images or video frames using classical machine learning and deep learning models [4][5]. Traditional approaches employ hand-crafted descriptors combined with classifiers such as Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors (kNN) [6][7]. However, these techniques are limited by their dependence on manually engineered features and are susceptible to occlusion, illumination variation, and pose changes [8][9]. Deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated superior feature extraction and classification capabilities, though they introduce computational overhead and risk losing local spatial information [9][10].

The second paradigm multimodal FER integrates complementary information from audio, video, and text streams to provide richer, more robust emotional representations [11]. Research has consistently shown that multimodal data captures emotion more comprehensively than unimodal approaches [3]. Nevertheless, multimodal fusion introduces its own challenges, including modality alignment, missing modality handling, fine-grained temporal dependency modeling, and sensitivity to noise across different signal types [12][13].

To address these challenges, this paper presents a unified study of two complementary frameworks. The first, LSTM-PAM, is a multimodal FER framework that leverages Long Short-Term Memory (LSTM) networks enhanced with a Position Attention Mechanism (PAM) to capture both temporal and spatial dependencies across audio, video, and text modalities [16]. The second, GWALO-RNN, is a unimodal FER framework that employs a novel hybrid Grey Wolf Optimization and Ant Lion Optimization algorithm (GWALO) for intelligent feature selection, coupled with an RNN classifier [15]. By studying these two frameworks together, we offer a broad and complementary perspective on the current frontiers of FER research.

The principal contributions of this work are as follows:

• A comprehensive multimodal preprocessing pipeline employing Wiener filtering [16], CLAHE, and lemmatization for audio, video, and text modalities, respectively, effectively reducing noise and standardizing representations across modalities.

• A multimodal feature extraction strategy integrating MFCC [16], GLCM, and Word2Vector, fused into a unified feature vector to capture spectral, textural, and semantic characteristics simultaneously.

• An LSTM architecture enhanced with PAM that dynamically allocates attention to critical spatial positions in facial feature maps, improving sensitivity to expression-relevant regions and enabling accurate spatiotemporal modeling.

• A novel hybrid GWALO algorithm that synergizes the exploration efficiency of Ant Lion Optimization (ALO) with the exploitation precision of Grey Wolf Optimization (GWO), enabling robust feature selection in high-dimensional spaces for unimodal FER [15].

• Empirical validation across three benchmark datasets SAVEE [14], CMU-MOSI [15], and JAFFE demonstrating state-of-the-art performance across both unimodal and multimodal FER settings.

The remainder of this paper is organized as follows: Section 2 reviews the related literature. Section 3 describes the proposed methodologies. Section 4 details experimental setup and datasets. Section 5 presents results and comparative analysis. Section 6 provides discussion. Section 7 concludes the paper with future directions.

2. Literature Review

A substantial body of literature has investigated FER using deep learning, optimization, and multimodal fusion. This section surveys the most relevant recent contributions, identifying key methodological trends and persistent limitations.

Jagadeesh and Bharanidharan [11] developed a DL model for real-time FER of online learners by combining the Coyote-Deer Hunting Optimization (C-DHO) algorithm with a Heuristically Modified RNN (HM-RNN). The Local Directional Texture Pattern (LDTP) was used for feature extraction, and a CNN performed deep feature learning. While the model achieved competitive results, it lacked the scalability provided by ensemble learning approaches and required extension to incorporate intelligent augmentation mechanisms.

Xu et al. [12] proposed the Joint Hierarchical Cross-Attention Graph Convolutional Network (JHGCN) for multimodal FER, integrating patch-embedding-based visual graphs, graph convolution for intra-modality relationships, and cross-attention mechanisms for inter-modality complementarity. Despite strong performance, the model faced challenges with ambiguous and noisy data points and required broader multimodal integration of video, audio, text, and EEG signals.

Sumalakshmi and Vasuki [13] introduced a fused DL approach using Feasible Weighted Squirrel Search Optimization (FW-SSO) for feature selection and LSTM with Attention Mechanism (ALSTM) for classification. The model performed well in capturing long-term landmark dependencies, but the assumption of neutral initial frames constrained its generalizability to real-time image collections.

Albraikan et al. [14] presented an Intelligent FER and classification system via optimal deep transfer learning (IFER-DTFL) using Mask R-CNN for detection, DenseNet121 for feature extraction, and Weighted Kernel Extreme Learning Machine (WKELM) for classification. Although highly accurate, the system's performance was limited by the absence of metaheuristic optimization methodologies for further refinement.

Ashok Kumar et al. [15] enhanced FER through optimal descriptor selection using the hybrid MV-WOA algorithm to reduce the high-dimensional ASIFT feature space, combined with neural network classification. While effective, this model demonstrated limited capability in solving real-time FER problems with varied emotional expressions.

In the multimodal domain, Nguyen et al. [9] proposed a deep cross-domain transfer with joint learning for emotion recognition, achieving strong cross-dataset generalization but struggling with fine-grained temporal dependencies in multimodal settings. He et al. [10] introduced the Mixture of Attention Variants for Modal Fusion (MAVMF), leveraging Bidirectional GRU and four multimodal attention variants, though overfitting remained a challenge with combined spatial and temporal data.

Wang et al. [12'] applied Tucker decomposition and self-supervised multi-tasking for multimodal FER, effectively reducing overfitting via parameter compression but remaining sensitive to speech pattern variations. Sun et al. [13'] embedded Multimodal Privileged Information (MPI) into FER, enabling inference

from face images alone at test time but suffering from spatial information loss during frequency-to-time domain transformation.

From this comprehensive analysis, two primary research gaps emerge: (i) existing multimodal methods struggle to simultaneously model fine-grained temporal dependencies and filter modality-specific noise effectively, and (ii) unimodal FER methods relying on standard optimization algorithms lack the adaptive search efficiency required for high-dimensional feature spaces. This paper directly addresses both gaps through LSTM-PAM and GWALO-RNN respectively.

3. Proposed Methodologies

This section presents both proposed frameworks in detail. Framework I (LSTM-PAM) addresses multimodal FER across audio, video, and text modalities. Framework II (GWALO-RNN) addresses unimodal FER with intelligent hybrid optimization-based feature selection.

3.1 Framework I: Multimodal LSTM with Position Attention Mechanism (LSTM-PAM)

The LSTM-PAM framework processes multimodal inputs through a sequential pipeline encompassing dataset acquisition, modality-specific preprocessing, feature extraction, feature fusion, and attention-enhanced sequential classification. Figure 1 presents the complete block diagram of the LSTM-PAM framework.

Fig. 1. Block diagram of the multimodal LSTM-PAM FER framework

3.1.1 Datasets

Two multimodal datasets are employed in Framework I:

SAVEE Dataset [14]: The Surrey Audio-Visual Expressed Emotion (SAVEE) database was collected at the University of Surrey and includes recordings from four native male British speakers aged 27–31 years. Each speaker expressed seven distinct emotional categories: anger, disgust, fear, happiness, sadness, surprise, and neutral, with 120 utterances per speaker. The dataset provides synchronized audio, video, and text modalities, making it suitable for multimodal FER research.

CMU-MOSI Dataset [15]: The CMU Multimodal Opinion Sentiment Intensity (CMU-MOSI) corpus comprises 2,199 video clips sourced from YouTube opinion recordings. Each clip is segmented into opinion units, with sentiment polarity scored by five human evaluators on a continuous scale from −3 (strongly negative) to +3 (strongly positive). The training and test sets contain 1,284 and 915 opinion segments, respectively.

3.1.2 Multimodal Preprocessing

Preprocessing is applied independently to each modality to enhance signal quality and reduce noise prior to feature extraction.

Audio Preprocessing: Wiener filtering [16] is applied to the raw audio signals to attenuate background noise. The filter estimates clean speech by exploiting the statistical properties of the observed noisy signal and its noise characteristics, improving signal-to-noise ratio and enhancing downstream feature extraction reliability.

Video Preprocessing: Raw video frames are extracted and processed as individual images. Contrast-Limited Adaptive Histogram Equalization (CLAHE) is applied to enhance local contrast. Unlike standard Adaptive Histogram Equalization (AHE), which amplifies noise by boosting local contrast without restriction, CLAHE incorporates a contrast-limiting threshold parameter that constrains the enhancement per image region. This approach preserves the hue and saturation of value components consistent with human color perception, delivering cleaner and more discriminative facial image representations.

Text Preprocessing: Lemmatization is applied to normalize the raw textual content. This process maps words to their morphological base forms using dictionary and context information for example, converting

"cars" to "car" while linking semantically related terms such as "automobile." Lemmatization reduces feature dimensionality at the textual level while preserving semantic richness, facilitating more effective training.

3.1.3 Feature Extraction

Following preprocessing, modality-specific feature extraction methods are applied to capture representative characteristics.

Audio Features MFCC: Mel-Frequency Cepstral Coefficients (MFCC) [16] are extracted from the preprocessed audio to mimic the human auditory system's frequency sensitivity. The pipeline applies discrete wavelet denoising, hamming windowing, Fast Fourier Transform (FFT) for time-to-frequency conversion, Mel filter bank application for spectral energy distribution, Discrete Cosine Transform (DCT) for coefficient decorrelation, and Cepstral Mean Normalization (CMN) for robustness against additive noise and speaking style variation.

Video Features GLCM: The Gray-Level Co-occurrence Matrix (GLCM) is employed for video frame texture analysis, capturing second-order statistical relationships among pixel intensity values. GLCM computes co-occurrence frequencies of pixel pairs across spatial neighborhoods, yielding features including correlation and homogeneity that characterize the textural properties critical for facial expression discrimination.

Text Features Word2Vector: Word2Vector generates dense, high-dimensional vector representations of text tokens that preserve semantic and contextual word relationships. Training is performed on the textual corpus with a word vector dimensionality of 100 and a context window size of 5, ensuring that semantically similar words occupy proximate positions in the embedding space, thereby enriching textual feature quality for FER.

3.1.4 Feature Fusion

After independent extraction, MFCC, GLCM, and Word2Vector feature vectors are concatenated into a unified multimodal feature vector [16]. This fusion strategy integrates spectral, textural, and semantic representations into a comprehensive descriptor, enabling the classification model to exploit cross-modal complementarity and improving recognition performance under diverse expression conditions.

3.1.5 LSTM-PAM Classification

The fused feature vector is fed into the LSTM-PAM classifier, which combines the temporal modeling strength of LSTM with the spatial focusing capability of the Position Attention Mechanism.

LSTM Architecture: LSTM [16] is an advanced variant of Recurrent Neural Networks (RNN) designed to overcome the vanishing and exploding gradient problems in standard RNNs. Each LSTM cell comprises three gates input, forget, and output and a memory cell that selectively retains or discards sequential information across time steps.

For input sequence $X = (x_1, x_2, \ldots, x_t)$ and hidden state $H = (h_1, h_2, \ldots, h_t)$, the LSTM computations are defined as:

$$h_t = H(W_{h\gamma}\, x_t + W_{hh}\, h_{t-1} + b_h) \quad (1)$$

$$y_t = W_{h\gamma}\, h_t + b_\gamma \quad (2)$$

The input gate i_t, forget gate f_t, cell state C_t, output gate o_t, and hidden state h_t are computed as:

$$i_t = \sigma(W^x_i\, x_t + W_{hi}\, h_{t-1} + W^c_i\, C_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W^{xi}\, x_t + W_h{}^i\, h_{t-1} + W^{ci}\, C_{t-1} + b^i) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * \tanh(W^{xc}\, x_t + W_h{}^c\, h_{t-1} + b^c) \quad (5)$$

$$o_t = \sigma(W^{xo}\, x_t + W_h{}^o\, h_{t-1} + W^{co}\, C_t + b^o) \quad (6)$$

$$h_t = o_t * \tanh(C_t) \quad (7)$$

where σ and tanh are the sigmoid and hyperbolic tangent activation functions, respectively:

$$\sigma(x) = 1 / (1 + e^{-x}) \quad (8)$$

$$\tanh(x) = (e^x - e^{-x}) / (e^x + e^{-x}) \quad (9)$$

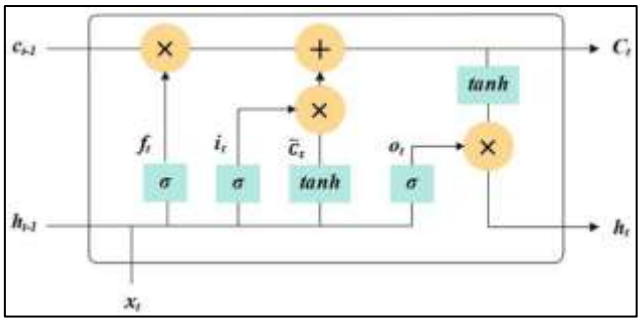

Fig. 2. LSTM architecture with memory cell structure

Position Attention Mechanism (PAM): PAM captures global spatial dependencies among feature map positions, enabling the model to selectively emphasize expression-critical facial regions. Given an input feature map $F \in \mathbb{R}^{cxHxW}$, three intermediate feature maps F_1, F_2, and F_3 are generated via convolution, batch

normalization, and ReLU activation. Reshape and matrix multiplication operations compute positional relationships, and the final attention-enhanced output F'' is:

$$F'' = \eth \cdot \text{reshape}(V \cdot M_p) + F' \quad (10)$$

where $\eth$ is a learnable scaling parameter and M_p encodes the positional affinity matrix. By integrating PAM into LSTM, the model simultaneously captures spatial structure and temporal dynamics, yielding more accurate expression recognition over time.

3.2 Framework II: Hybrid GWALO-RNN for Unimodal FER

The GWALO-RNN framework performs FER from facial image data through a pipeline of face detection, preprocessing, feature extraction, hybrid optimization-based feature selection, and RNN classification. Figure 3 presents the block diagram of this framework.

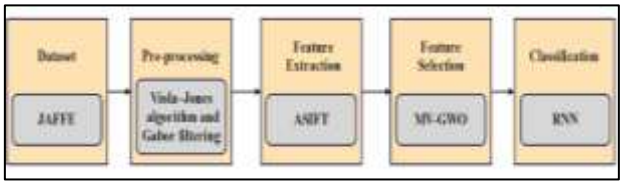

Fig. 3. Block diagram of the GWALO-RNN FER framework

3.2.1 Dataset

The Japanese Female Facial Expression (JAFFE) database is utilized in Framework II. It comprises 213 grayscale facial images from 10 Japanese female subjects, each expressing 2–4 of seven emotional categories: anger, disgust, fear, happiness, neutral, sadness, and surprise. All images have dimensions of 213 × 213 pixels and are normalized to 28 × 28 pixels during preprocessing.

3.2.2 Preprocessing

Viola-Jones Algorithm: The Viola-Jones algorithm performs face detection from input images by identifying key facial landmarks (eyes, nose, mouth, ears) essential for subsequent expression analysis. The algorithm rotates detection windows 45° around both vertical and horizontal axes, enabling robust frontal face localization through a cascade of boosted classifiers applied across multiple image scales.

Gabor Filtering: Following face detection, Gabor filtering is applied for texture analysis and orientation-sensitive noise reduction. The Gabor filter acts as a bandpass filter that selectively attenuates non-relevant frequency components while preserving expression-discriminative texture patterns, facilitating more precise feature point and landmark localization.

3.2.3 Feature Extraction via ASIFT

The Affine-Scale-Invariant Feature Transform (ASIFT) is applied to extract robust keypoint descriptors from preprocessed images [15]. ASIFT extends the standard SIFT algorithm by systematically simulating the full range of affine camera perspective distortions, encompassing spin (γ), longitude (φ), zoom (μ), and latitude (α) parameters. This enables invariant feature matching across images captured from varying viewpoints, providing highly discriminative and geometrically robust expression descriptors.

3.2.4 Hybrid GWALO Feature Selection

A novel hybrid optimization algorithm, GWALO, is proposed by integrating Grey Wolf Optimization (GWO) and Ant Lion Optimization (ALO) to intelligently select the most discriminative features from the high-dimensional ASIFT feature space.

Grey Wolf Optimization (GWO): GWO mimics the hierarchical hunting behavior of grey wolf packs, employing four agent roles alpha (α), beta (β), delta (δ), and omega (ω) where α, β, and δ guide the search process. The positional update equations are:

$$D = |C \cdot x_p(t) - x(t)| \quad (11)$$

$$x(t+1) = x_p(t) - A \cdot D \quad (12)$$

where $A = 2a \cdot r_1 - a$ and $C = 2r_2$, with a decreasing linearly from 2 to 0 and $r_1, r_2 \in [0,1]$. The final wolf position updates from three best solutions (α, β, δ) are:

$$X_1 = X\alpha - A_1 \cdot D\alpha;\ X_2 = X\beta - A_2 \cdot D\beta;\ X_3 = X\delta - A_3 \cdot D\delta \quad (13\text{-}14)$$

$$x(t+1) = (X_1 + X_2 + X_3) / 3 \quad (15)$$

Ant Lion Optimization (ALO): ALO replicates the trap-construction hunting behavior of antlions. Ants perform random walks across the search space and fall into funnel-shaped traps built by antlions. The random walk is governed by:

$$x(t) = [0, \text{cumsum}(2\gamma(t_1)-1), \text{cumsum}(2\gamma(t_2)-1), \ldots, \text{cumsum}(2\gamma(t_n)-1)] \quad (16)$$

Ant boundary constraints, trap falling, sliding dynamics, capture, and antlion trap rebuilding are implemented through Equations (11–18). The elite antlion tracks the best-found solution and guides ant movement across iterations, enabling progressive refinement of the feature subset.

Proposed GWALO Integration: GWO excels at exploitation in low-dimensional spaces, while ALO compensates with superior exploration in high-dimensional spaces. GWALO synergizes these complementary strengths: GWO provides precise position refinement around the best-known solutions,

while ALO supplies diverse candidate solutions through roulette-wheel-selected random walks. GWALO selects n features from the initial candidate set, iteratively refining the feature subset through position updates informed by both GWO's hierarchical ranking and ALO's trapping dynamics, yielding a compact and highly discriminative feature subset.

3.2.5 RNN Classification

A Recurrent Neural Network (RNN) classifies facial expressions from the optimized feature subset. RNN processes sequential input through a feedback loop at each time step, updating hidden states based on current input and prior context. The hidden and output state computations are:

$h_t = f(W_h \cdot h_{t-1} + V_h \cdot x_t + b_h)$ (17)

$o_t = f(W_o \cdot h_t + b_o)$ (18)

where W and V are learnable weight matrices, b denotes bias vectors, and f is the nonlinear activation function. Figure 4 shows the RNN structural representation.

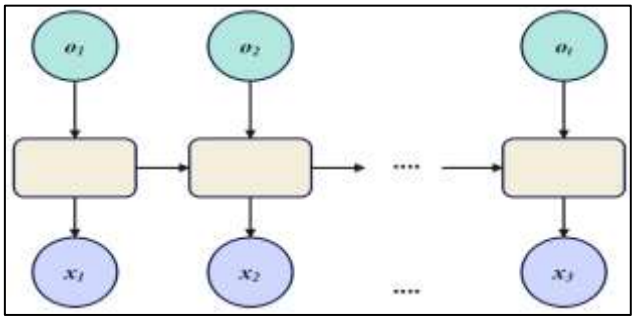

Fig. 4. Recurrent Neural Network (RNN) architecture

4. EXPERIMENTAL SETUP

Both frameworks are implemented in MATLAB R2020b. Framework I (LSTM-PAM) uses a system configuration of Windows 10, Intel i5 processor, and 8 GB RAM. Framework II (GWALO-RNN) utilizes Windows 10, Intel i7 processor, and 16 GB RAM. Performance evaluation employs four standard metrics for both frameworks: accuracy, precision, recall, and F1-score, as defined by Equations (19-23):

Accuracy = (TP + TN) / (TP + TN + FP + FN) (19)

Precision = TP / (TP + FP) (20)

Recall = TP / (TP + FN) (21)

F1-Score = 2 × (Precision × Recall) / (Precision + Recall) (22)

Specificity = TN / (FP + TN) × 100 (23)

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

5. Results And Comparative Analysis

5.1 Framework I: LSTM-PAM Evaluation

Table I presents a comparative evaluation of different classification models for multimodal FER on the SAVEE and CMU-MOSI datasets. Results for ANN, CNN, and RNN baselines are reported alongside the proposed LSTM model. LSTM achieves 98.25%, 97.00%, 97.74%, and 96.28% of accuracy, F1-score, precision, and recall on SAVEE, and 95.78%, 94.01%, 94.53%, and 93.49% on CMU-MOSI, consistently outperforming all baseline classifiers across both datasets.

Table I. Performance Comparison of Different Classification Models for FER

Dataset	Method	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
SAVEE	ANN	91.25	90.48	90.74	90.22
	CNN	94.52	92.90	93.27	92.53
	RNN	96.19	93.60	94.09	93.11
	LSTM (Proposed)	98.25	97.00	97.74	96.28
CMU-MOSI	ANN	90.15	90.19	90.26	90.12
	CNN	92.78	91.88	93.00	90.78
	RNN	93.00	92.38	93.27	91.50

Dataset	Method	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
	LSTM (Proposed)	95.78	94.01	94.53	93.49

Table II compares different attention mechanisms integrated within the LSTM framework. Channel Attention Mechanism (CAM), Global Attention Mechanism (GAM), and Self-Attention Mechanism (SAM) are evaluated against the proposed PAM. PAM achieves the highest performance across all metrics on both datasets, confirming the advantage of spatially-aware positional attention in capturing expression-critical facial regions.

Table II. Performance Comparison of Different Attention Mechanisms for FER

Dataset	Method	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
SAVEE	CAM	93.13	93.51	94.47	92.56
	GAM	95.27	93.65	94.71	92.62
	SAM	96.58	95.07	96.12	94.02
	PAM (Proposed)	98.25	97.00	97.74	96.28
CMU-MOSI	CAM	91.87	91.26	91.18	90.36
	GAM	92.79	90.63	91.57	89.74
	SAM	94.23	92.15	93.08	91.23
	PAM (Proposed)	95.78	94.01	94.53	93.49

Table III presents the state-of-the-art comparison for Framework I. LSTM-PAM is compared with Joint Learning [9], MAVMF [10], MMER-TD [12], and MPI-FER [13]. On SAVEE, LSTM-PAM achieves 98.25% accuracy versus 97.7% for the next best (Joint Learning [9]). On CMU-MOSI, LSTM-PAM significantly outperforms MAVMF [10] (82.31%), MMER-TD [12] (85.52%), and MPI-FER [13] (79.30%) by substantial margins, demonstrating its superiority in multimodal expression recognition.

Table III. State-of-the-Art Comparative Analysis for Multimodal FER (FRAMEWORK I)

Dataset	Method	Accuracy (%)	F1-Score (%)
SAVEE	Joint Learning [9]	97.70	N/A
	LSTM-PAM (Proposed)	98.25	97.00
CMU-MOSI	MAVMF [10]	82.31	82.20
	MMER-TD [12]	85.52	85.47
	MPI-FER [13]	79.30	80.10
	LSTM-PAM (Proposed)	95.78	94.01

5.2 Framework II: GWALO-RNN Evaluation

Table IV presents the quantitative performance of the GWALO-RNN framework against four baseline optimization algorithms Grasshopper Optimization Algorithm (GOA), Whale Optimization Algorithm (WOA), GWO, and ALO on the JAFFE dataset. GWALO-RNN achieves 99.7% accuracy, 98.5% precision, 97.8% recall, 99.6% specificity, and 96.7% F1-score, outperforming all baseline methods across every metric. Notably, GWO individually achieves 95.1% accuracy, while ALO achieves 97.7%, confirming that the hybrid GWALO integration yields performance superior to either component algorithm in isolation.

Table IV. Performance Comparison of Optimization Algorithms for FER (FRAMEWORK II)

Method	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)
GOA	91.3	93.4	92.1	92.8	90.8
WOA	94.7	94.7	94.7	95.1	91.4
GWO	95.1	95.3	96.0	96.4	92.4
ALO	97.7	97.6	96.5	96.9	94.9
GWALO-RNN (Proposed)	99.7	98.5	97.8	99.6	96.7

Table V presents a comparative analysis of GWALO-RNN against published state-of-the-art methods C-DHO-HM-RNN [11], IFER-DTFL [14], and MV-WOA [15]. GWALO-RNN achieves the highest accuracy (99.7%), precision (98.5%), recall (97.8%), specificity (99.6%), and F1-score (96.7%). Notably, while IFER-DTFL [14] achieves 99.1% accuracy with strong specificity (99.5%), GWALO-RNN surpasses it in accuracy and F1-score while maintaining competitive precision and recall. MV-WOA [15] shows particularly low F1-score (69.5%), indicating limited discriminative ability despite competitive specificity.

Table V. State-of-the-Art Comparative Analysis for Unimodal FER (FRAMEWORK II)

Method	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)
C-DHO-HM-RNN [11]	91.5	98.0		51.3	95.4
IFER-DTFL [14]	99.1	97.4	97.1	99.5	97.1
MV-WOA [15]	92.9	90.7		99.0	69.5
GWALO-RNN (Proposed)	99.7	98.5	97.8	99.6	96.7

6. Discussion

The experimental results across both frameworks provide strong empirical evidence of the complementary strengths and practical utility of the proposed approaches.

In Framework I, the LSTM-PAM achieves state-of-the-art multimodal FER by effectively addressing the two primary challenges identified in the literature: fine-grained temporal dependency modeling and modality-specific noise sensitivity. The modality-specific preprocessing pipeline Wiener filtering [16], CLAHE, and lemmatization substantially improves feature quality at the input stage. The combination of MFCC [16], GLCM, and Word2Vector features provides complementary spectral, textural, and semantic representations that individually capture distinct aspects of emotional expression, while feature-level concatenation fusion ensures comprehensive cross-modal information integration. The PAM component enables the LSTM to dynamically focus on expression-critical spatial positions within facial feature maps, effectively filtering irrelevant and noisy spatial regions that could confuse classification. This spatial awareness, combined with LSTM's ability to model long-range sequential dependencies, results in a model that is simultaneously temporally deep and spatially precise, explaining the significant accuracy gains over standard LSTM baselines and existing multimodal FER methods [9][10][12][13].

In Framework II, the GWALO-RNN demonstrates that intelligent hybrid optimization is a decisive factor in unimodal FER with high-dimensional feature spaces. GWO's hierarchical social structure provides strong exploitation near the current best solution, but its performance degrades on high-dimensional feature spaces due to limited diversity. ALO's roulette-wheel-based trap selection provides the necessary diversity for effective high-dimensional exploration. By combining both algorithms, GWALO achieves a dynamically balanced exploration-exploitation trade-off that neither algorithm achieves in isolation evidenced by GWALO-RNN's accuracy of 99.7% versus GWO's 95.1% and ALO's 97.7% individually. The ASIFT feature extractor [15] provides viewpoint-invariant descriptors that enable robust recognition across pose variations in the JAFFE dataset, while the RNN classifier effectively leverages the optimized feature subset for sequential classification of facial muscle dynamics.

Comparing the two frameworks reveals an important complementarity: LSTM-PAM excels in multimodal, temporally rich settings with heterogeneous data streams, while GWALO-RNN provides near-perfect accuracy in controlled unimodal settings through intelligent feature selection. The broader implication is that optimal FER system design requires careful alignment between the data modality structure, the feature representation strategy, and the classification paradigm. Future work should explore the integration of GWALO-based feature selection as a preprocessing stage for LSTM-PAM, potentially yielding a unified

framework that benefits from both hybrid optimization and attention-enhanced sequential modeling across multimodal inputs.

7. Conclusion

This paper presented a comprehensive investigation of two complementary FER frameworks: LSTM-PAM for multimodal FER and GWALO-RNN for unimodal FER. Framework I employs Wiener filtering [16], CLAHE, and lemmatization for preprocessing, MFCC [16], GLCM, and Word2Vector for feature extraction, and an LSTM enhanced with PAM for spatiotemporal classification, achieving 98.25% accuracy on SAVEE and 95.78% on CMU-MOSI. Framework II employs Viola-Jones detection, Gabor filtering, ASIFT feature extraction [15], and a novel GWALO hybrid optimizer for feature selection with RNN classification, achieving 99.7% accuracy on JAFFE. Both frameworks individually surpass state-of-the-art baselines across all evaluated metrics.

The findings collectively demonstrate that: (i) modality-specific preprocessing and targeted feature extraction are essential prerequisites for effective multimodal FER; (ii) attention mechanisms that encode spatial position information significantly enhance the discriminability of sequential models; and (iii) hybrid optimization integrating complementary algorithmic strengths achieves superior feature selection compared to any single optimization method. Future research will focus on developing a unified architecture that integrates hybrid optimization with multimodal attention mechanisms, extending evaluation to larger and more diverse benchmark datasets, exploring model compression for real-time deployment, and investigating FER under occlusion, low-light, and cross-cultural expression variations.

References

[1] A. Anand and S. Babu, “Predicting the Facial Expression Recognition Using Novel Enhanced Gated Recurrent Unit-based Kookaburra Optimization Algorithm,” Int. J. Intell. Eng. Syst., vol. 17, no. 3, 2024.

[2] N. Perveen, D. Singh, and C. K. Mohan, “Spontaneous facial expression recognition: A part based approach,” in Proc. IEEE ICMLA, 2016, pp. 819–824.

[3] C. Cheng, W. Liu, Z. Fan, L. Feng, and Z. Jia, “A novel transformer autoencoder for multi-modal emotion recognition with incomplete data,” Neural Netw., vol. 172, p. 106111, 2024.

[4] P. Ganesan, G. P. Ramesh, C. Puttamdappa, and Y. Anuradha, “A Modified Bio-Inspired Optimizer with Capsule Network for Diagnosis of Alzheimer Disease,” Appl. Sci., vol. 14, no. 15, p. 6798, 2024.

[5] S. Liang, “Enhancing Multimodal Emotional Information Extraction in Film and Television through Adaptive Feature Fusion with DenseNet, Transformer, and 3D CNN Models,” Appl. Artif. Intell., vol. 38, no. 1, p. 2419609, 2024.

[6] N. Gahlan and D. Sethia, “AFLEMP: Attention-based federated learning for emotion recognition using multi-modal physiological data,” Biomed. Signal Process. Control, vol. 94, p. 106353, 2024.

[7] Z. Du, X. Ye, and P. Zhao, “A novel signal channel attention network for multi-modal emotion recognition,” Front. Neurorobotics, vol. 18, p. 1442080, 2024.

[8] B. Yin et al., “Dynamic facial expression recognition with pseudo-label guided multi-modal pre-training,” IET Comput. Vis., vol. 18, no. 1, pp. 33–45, 2024.

[9] D. Nguyen et al., “Deep cross-domain transfer for emotion recognition via joint learning,” Multimedia Tools Appl., vol. 83, no. 8, pp. 22455–22472, 2024.

[10] C. He et al., “Mixture of attention variants for modal fusion in multi-modal sentiment analysis,” Big Data Cogn. Comput., vol. 8, no. 2, p. 14, 2024.

[11] M. Jagadeesh and B. Bharanidharan, “Facial expression recognition of online learners from real-time videos utilizing a novel deep learning model,” Multimedia Syst., vol. 28, no. 6, pp. 2285–2305, 2022.

[12] C. Xu, Y. Du, J. Wang, W. Zheng, T. Li, and Z. Yuan, “A joint hierarchical cross-attention graph convolutional network for multi-modal facial expression recognition,” Comput. Intell., vol. 40, no. 1, p. e12607, 2024.

[13] C. H. Sumalakshmi and P. Vasuki, “Fused deep learning based Facial Expression Recognition of students in online learning mode,” Concurr. Comput. Pract. Exp., vol. 34, no. 21, p. e7137, 2022.

[14] A. A. Albraikan et al., “Intelligent facial expression recognition and classification utilizing optimal deep transfer learning model,” Image Vis. Comput., vol. 128, p. 104583, 2022.

[15] P. M. Ashok Kumar, J. B. Maddala, and K. Martin Sagayam, “Enhanced facial emotion recognition by optimal descriptor selection with neural network,” IETE J. Res., vol. 69, no. 5, pp. 2595–2614, 2023.

[16] K. R. Babu, A. Suneetha, and K. K. Kumar, “Multimodal Face Expression Recognition Using Parametric Exponential Linear Unit-Long Short-Term Memory,” J. Comput. Sci., vol. 20, no. 10, pp. 1339–1348, 2024.

[17] R. Wang, J. Zhu, S. Wang, T. Wang, J. Huang, and X. Zhu, “Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking,” Int. J. Multimedia Inf. Retr., vol. 13, no. 4, p. 39, 2024.

[18] N. Sun et al., "Multimodal Sentimental Privileged Information Embedding for Improving Facial Expression Recognition," IEEE Trans. Affect. Comput., 2024.
[19] SAVEE Dataset. [Online]. Available: https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee (Accessed: March 2025).
[20] CMU-MOSI Dataset. [Online]. Available: https://www.kaggle.com/datasets/mathurinache/cmu-mosi (Accessed: March 2025).