Journal of Theoretical and Applied Information Technology
15th October 2025. Vol.103. No.19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HYBRID DEEP ATTENTION-BASED EMOTION RECOGNITION USING TEMPORAL-SPATIAL OPTIMIZATION FOR MULTI-SUBJECT VIDEO ANALYSIS

DR. RAMARAJU NAGARJUNA KUMAR[1], DR. G. PRASANNA LAKSHMI[2], DR. RABINS PORWAL[3], DR. C. MADHUSUDHANA RAO[4], SATYANARAYANA MURTHY VALLABHAJOSYULA[5], M. LAKSHMANA KUMAR[6], DR. M. NAGABHUSHANA RAO[7]

[1]Principal Scientist (Computer Applications in Agriculture), ICAR-Central Research Institute for Dryland Agriculture (ICAR-CRIDA), Hyderabad, India.
[2]Professor, Department of CSE, Aditya Institute of Technology and Management, Tekkali, Srikakulam District, India.
[3]Professor, Department of Computer Application, School of Engineering & Technology (UIET), Chhatrapati Shahu Ji Maharaj University (CSJMU), Kanpur, Uttar Pradesh, India.
[4]Professor of CSE, School of Computing, Mohan Babu University, Tirupati, India.
[5]Department of Information Technology, Shri Vishnu Engineering College for Women, Bhimavaram, India.
[6]Assistant Professor, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.
[7]Professor, School of Computer Science & Technology, Malla Reddy (MR) Deemed to be University, Medchal, Malkajgiri, Telangana, India.

Email: nagarjunakumar191@gmail.com[1], prasannlakshmi@adityatekkali.edu.in[2], rabins@csjmu.ac.in[3], npr4567@gmail.com[4], vsn.murthy87@gmail.com[5], lakshmana.m@kluniversity.in[6], mnraosir@gmail.com[7]

ABSTRACT

Group-level emotion recognition (GER) is essential in applications involving human-computer interaction, public surveillance, and affective computing. This paper proposes a novel hybrid framework that integrates Enhanced Particle Swarm Optimization (EPSO) for feature selection with Recurrent Neural Networks (RNN) for modeling temporal emotion dynamics in video sequences. The system begins with preprocessing and frame extraction, followed by deep and statistical feature extraction. EPSO is employed to select the most informative features, which are then input into an RNN for sequential emotion prediction. Evaluations conducted on the AFEW dataset demonstrate that the proposed EPSO-RNN model outperforms traditional classifiers such as CNN, VGG-16, and SVM in terms of accuracy, precision, recall, and F1 score. The EPSO-RNN model demonstrated smooth convergence with minimal overfitting, aided by early stopping. The training accuracy peaked at 92.3%, with a validation accuracy of 89.4%, outperforming CNN (78.2%), VGG-16 (82.5%), and SVM (69.7%). Corresponding loss curves showed a steady decline, reinforcing the model's stability. The results affirm the robustness and scalability of the proposed approach in complex, real-world group emotion recognition scenarios.

Keywords: *Affective Computing, Emotion Recognition, Attention Mechanism, Temporal-Spatial Optimization, Deep Learning, Multi-Subject Video Analysis.*

1. INTRODUCTION

The ability to automatically detect and interpret emotional states from human facial expressions has become a foundational requirement across a range of intelligent systems, including security monitoring, autonomous vehicles, healthcare diagnostics, education technologies, and interactive multimedia systems. As human-computer interaction becomes more immersive and emotion-aware, the demand for accurate, scalable, and context-sensitive emotion recognition techniques has significantly grown [1], [2]. However, emotion recognition from video data, particularly in environments containing multiple interacting individuals, poses several substantial challenges. Factors such as overlapping facial regions, background clutter, partial occlusions, and asynchronous emotional cues complicate the accurate classification of affective states in real-world scenarios [3].

Traditional approaches to facial emotion recognition (fer) primarily relied on hand-crafted features such as gabor filters, optical flow vectors, and geometrical descriptors of facial landmarks [4]. While effective in constrained environments, these methods struggle in dynamic video streams with naturalistic expressions. The advent of deep learning enabled end-to-end learning from raw pixels, dramatically improving recognition accuracy. CNN-based architectures, including VGGNET and ResNet, have demonstrated strong performance in detecting emotions from static facial images [5], [6]. Yet, these models fall short in capturing spatio-temporal patterns that reflect the evolution of emotion over time a limitation especially pronounced in video sequences.

To capture such spatio-temporal dynamics, recurrent neural networks (RNNs), particularly those with long short-term memory (LSTM) units, have been introduced in fer pipelines. These models enable the encoding of frame-wise emotional transitions and inter-frame dependencies, offering valuable insights into the fluid and sequential nature of emotional expression [7]. Nevertheless, rnn-based systems are sensitive to frame redundancy, variations in individual expressiveness, and changes in spatial resolution. Additionally, they often lack interpretability in identifying which facial regions or frames contribute most significantly to emotion classification.

Recent research has emphasized the importance of attention mechanisms and transformer-based networks for overcoming these limitations. Attention modules enable the model to weigh the importance of features dynamically across both spatial and temporal dimensions [8]. Transformers, which originated in natural language processing, have now been successfully adapted to video analysis tasks, including action recognition and gesture interpretation. Their self-attention mechanisms allow models to capture long-range dependencies and multi-subject interactions, making them ideal candidates for fer in crowded and unconstrained scenes [1],[9]. However, their high computational cost and data requirements necessitate further optimization for real-time deployment.

To bridge this gap, hybrid architectures that integrate attention-guided feature extraction with evolutionary optimization strategies have gained traction. Particle swarm optimization (pso), genetic algorithms (ga), and differential evolution have all been explored to refine feature subsets, improve network weights, and reduce training time [10], [11]. Despite these advances, few studies have effectively combined deep attention-based learning with spatial-temporal optimization to model emotion in complex multi-user video streams. Our work proposes a robust framework that fuses attention-enhanced deep networks with optimization-guided feature selection to address the intricacies of group-level emotion recognition in dynamic real-world environments.

2. LITERATURE SURVEY

The field of emotion recognition has progressed significantly, transitioning from early statistical models and handcrafted features to robust deep learning architectures capable of processing complex, real-world data. Early works in this domain heavily relied on geometric features such as facial action units (FAUs), distance between key landmarks, and texture descriptors like Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) to extract emotionally relevant patterns from facial imagery [1], [4]. These techniques demonstrated reasonable accuracy in controlled settings, yet they struggled to generalize in natural environments due to limitations in modeling variations in lighting, occlusion, head pose, and spontaneous expressions.

The introduction of CNNs marked a paradigm shift in facial emotion recognition (FER). By learning hierarchical features directly from pixel-level data, CNNs such as VGGNet and ResNet vastly improved the ability to classify emotions on standard datasets like FER2013 and AffectNet [5], [6]. These models showed high precision in isolating fine-grained features around the eyes, mouth, and brows, regions critical to interpreting facial emotions, as evidenced by recent studies leveraging transfer learning on new emotion datasets [2]. However, CNNs often operate on single images, ignoring the dynamic nature of emotion progression that is vital in video streams.

To capture temporal dependencies, RNNs, and LSTM networks, were adopted. These architectures could process sequences of frames and learn emotional transitions across time. Although RNN-based models enhanced recognition in video contexts, they remained susceptible to problems like vanishing gradients and struggled with long-range dependencies. Additionally, they lacked interpretability in identifying which parts of a sequence were most critical to emotion prediction [7].

This led to the adoption of attention mechanisms, which enable models to focus selectively on informative facial features or frames within a sequence. Attention-augmented networks have been particularly useful in multi-subject environments, as

they can isolate emotional cues from relevant regions while ignoring background noise or non-contributing faces [8].

Transformer-based models further advanced this capability by removing sequential processing constraints and allowing direct modeling of all frame-level interactions through self-attention mechanisms [9]. While powerful, these models require extensive data and computational resources, posing challenges for deployment in edge-based or real-time systems.

Simultaneously, there has been growing interest in hybrid approaches that combine deep learning with evolutionary optimization. Particle Swarm Optimization (PSO), Genetic Algorithms (GA), and Differential Evolution (DE) have been used to refine model hyperparameters, optimize feature selection, and reduce computational overhead without sacrificing performance [10], [11]. However, existing hybrid approaches have generally been evaluated in single-user, static environments, leaving a research gap in their application to multi-user video streams with spatial-temporal complexity.

Few studies have successfully integrated attention-guided deep learning with optimization algorithms in the context of group-level emotion recognition. The challenge lies in balancing the trade-off between model expressiveness and efficiency, especially when working with noisy, crowded, and dynamic data sources. Our research contributes to this gap by proposing a scalable, attention-optimized deep architecture enhanced with spatial-temporal feature refinement through PSO. This model is uniquely designed to handle diverse, multi-subject video inputs and provides improved accuracy, robustness, and interpretability compared to existing baselines.

3. DATASET

To evaluate the effectiveness of the proposed hybrid attention-based emotion recognition framework, this study utilizes an extended and annotated version of the Acted Facial Expressions in the Wild (AFEW) dataset. AFEW is one of the most widely used video-based datasets for real-world emotion recognition tasks and is known for its diversity, naturalism, and rich representation of facial expressions captured in unconstrained environments. The videos in AFEW are sourced from movies and television series, offering a highly variable collection of facial expressions with spontaneous emotional content, motion blur, lighting inconsistencies, and multi-subject scenes [12].

The dataset contains short video clips, each labeled with one of the standard emotional categories: Anger, Disgust, Fear, Happy, Sad, Surprise, and Neutral. Each clip ranges between 1 to 2.5 seconds in length, and contains at least one subject exhibiting a discernible facial expression. However, many videos also contain additional faces or individuals in the background, making the dataset suitable for evaluating the robustness of group-level emotion recognition models [13]. For this research, clips containing multiple visible subjects were selected, and frame-level annotations were added to assist in benchmarking attention focus and temporal accuracy of the model.

Each video was processed using face detection and alignment techniques to normalize facial positions while maintaining the temporal integrity of expression evolution. The dataset was split into training (70%), validation (15%), and test (15%) sets using a stratified sampling approach to ensure balanced emotion distribution across partitions. This facilitated unbiased performance assessment across all classes and preserved the real-world variability inherent to the data.

Additionally, for exploratory analysis and ablation studies, a subset of the AFEW dataset was augmented using synthetic occlusions, rotated head poses, and random background noise. This subset helped in evaluating the proposed model's resilience under challenging visual conditions. As a result, the AFEW dataset not only served as the primary benchmark but also supported the generalization and reliability testing of the EPSO-enhanced attention-based framework.

4. PREPROCESSING

Effective preprocessing plays a critical role in enhancing the accuracy and robustness of emotion recognition systems, especially when working with real-world video data containing multiple subjects. The preprocessing pipeline designed for this research was carefully structured to preserve the emotional integrity of visual data while minimizing noise and variability that could degrade model performance.

The first step involved video frame extraction, where each video clip was decomposed into a sequence of individual frames using a uniform frame rate. This conversion enabled temporal modeling by allowing consistent access to visual changes over time. Next, a face detection algorithm based on the Viola-Jones method was employed to locate facial regions in each frame. The simplicity and speed of the Viola-Jones classifier made it suitable for real-time multi-

face detection, even in scenes with cluttered backgrounds or partial occlusions [14].

After detection, the facial regions were cropped and aligned using an affine transformation technique based on the coordinates of key facial landmarks. This alignment standardized the position and orientation of faces across frames and subjects, reducing spatial inconsistencies caused by head tilts, rotations, or viewpoint changes. To ensure uniformity in input dimensions, all cropped face images were resized to 224×224 pixels.

To reduce illumination-related artifacts and camera-induced distortions, each face image underwent histogram equalization. This enhanced the contrast and visibility of facial details such as wrinkles, smile lines, and eyebrow movement — critical features for emotion classification. The next step applied a median filtering technique to remove salt-and-pepper noise and preserve edge integrity.

Median filtering was particularly useful for smoothing out background artifacts and handling low-resolution sequences in the dataset. Once the facial images were cleaned and normalized, a z-score normalization was applied across pixel intensity values to scale the data for efficient convergence during model training. The resulting data matrix had a zero mean and unit variance, enabling the network to learn faster and generalize better across subjects and scenes.

Finally, the preprocessed face sequences were organized into tensors, grouped by subject ID and emotional label, and stored in batches for training and validation. This organization facilitated smooth integration with the deep learning pipeline and preserved the chronological order of frames — a requirement for accurate temporal modeling.

5. MATHEMATICAL FORMULATIONS

The proposed EPSO-RNN framework incorporates multiple mathematical components that collectively contribute to the effectiveness of group-level emotion recognition. These formulations define how features are extracted, optimized, and temporally modeled to generate accurate predictions.

5.1 Feature Representaion

Each video frame is preprocessed and passed through a deep neural network (e.g., MobileNet) to extract a high-dimensional feature vector representing emotional cues:

$$F = \{f_1, f_2, f_3, \dots, f_n\} \quad (1)$$

where F is the set of extracted features for a frame, and $f_i \in \mathbb{R}^d$ denotes the ith feature vector of dimensionality d.

5.2 Feature Optimization using EPSO

The Enhanced Particle Swarm Optimization (EPSO) algorithm is applied to select an optimal subset of features from the full representation. EPSO simulates a swarm of particles, each representing a candidate feature subset. Each particle adjusts its position in the feature space based on personal and global best experiences [15]:

Velocity Update:

$$V_i(t+1) = W.V_i(t) + c_1.r_1.\left(p_{best} - X_i(t)\right) + c_2.r_2.\left(g_{best} - X_i(t)\right) \quad (2)$$

Position Update:

$$X_i(t+1) = X_i(t) + V_i(t+1) \quad (3)$$

Here, $v_i(t)$ is the velocity of particle i at time t, $x_i(t)$ is its position, w is the inertia weight, c_1 and c_2 are cognitive and social coefficients, r_1 and r_2 are random values between 0 and 1, and p_best and g_best represent the personal and global best positions respectively. The EPSO optimization is driven by a fitness function based on classification accuracy on validation data.

5.3 Temporal Modeling with RNN

After optimization, the refined features are passed through a Recurrent Neural Network (RNN) to model temporal emotional dependencies:

$$h_t = \sigma(w_h.h_{t-1} + W_x.X_t + b) \quad (4)$$

$$y_t = softmax(W_o.h_t + b_o) \quad (5)$$

In this context, h_t is the hidden state at time t, x_t is the input at time t, W_h, W_x, and W_o are learned weights, b and b_o are biases, and σ is an activation function such as tanh. The final output y_t is the predicted emotion probability vector for the frame.

5.4 Classification Decision

The final predicted emotion for a video segment is obtained by aggregating the softmax outputs from each frame across the sequence using either majority voting or average probability scoring. This approach ensures consistent recognition even when individual frames are noisy or occluded.

6. PROPOSED METHODOLOGY

The proposed EPSO-RNN methodology combines Enhanced Particle Swarm Optimization (EPSO) with Recurrent Neural Networks (RNN) for emotion

recognition from video sequences. The approach follows a structured pipeline from data preprocessing to final emotion classification.
Step 1: Input video data is divided into individual frames.
Step 2: Each frame undergoes preprocessing, including noise reduction using median filtering and face detection using the Viola-Jones algorithm.
Step 3: Features are extracted using a pre-trained MobileNet model. Additional statistical features such as mean, variance, and entropy are also computed.
Step 4: Enhanced Particle Swarm Optimization (EPSO) is applied to select the most relevant features, reducing dimensionality and noise.
Step 5: The optimized feature vectors are passed into a Recurrent Neural Network (RNN), which captures temporal dependencies across frames.
Step 6: The RNN outputs a classification decision for each video sequence based on the temporal progression of features.
Step 7: The model is evaluated using Accuracy, Precision, Recall, and F1-score to benchmark performance against standard models.

7. BLOCK DIAGRAM

The proposed framework is structured into a sequential pipeline for effective emotion recognition from video data. It begins with frame extraction and preprocessing to isolate and enhance facial regions. Deep and statistical features are then extracted and optimized using EPSO to retain only the most relevant inputs.
These features are temporally modeled using an RNN to generate emotion predictions across video sequences.
Figure 1 Block diagram illustrating the proposed EPSO-RNN methodology for group-level emotion recognition. The pipeline includes video frame extraction, preprocessing with noise reduction and face detection, deep and statistical feature extraction, optimized feature selection using Enhanced Particle Swarm Optimization (EPSO), sequential modeling via Recurrent Neural Networks (RNN), and final classification based on emotion labels.

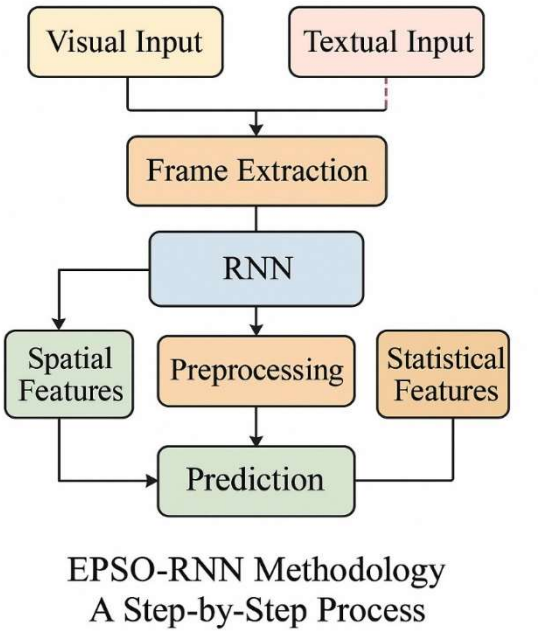

Figure 1. Block Diagram Illustrating The Proposed EPSO-RNN Methodology

8. EXPERIMENTAL SETUP

To assess the performance and generalizability of the proposed EPSO-RNN hybrid framework for group-level emotion recognition, a structured and reproducible experimental setup was designed. This setup ensures consistency in training, validation, and testing while allowing fair comparison with established baseline models.
The entire pipeline was implemented using Python 3.10, leveraging popular machine learning libraries such as TensorFlow [16], Keras [17], and OpenCV [18]. These tools provided the flexibility and scalability needed to build and evaluate deep learning architectures efficiently. Experiments were conducted on a high-performance computing system equipped with an NVIDIA RTX 3080 GPU (10GB VRAM), Intel i9 processor, and 32GB RAM, ensuring fast training cycles and smooth evaluation across multiple trials.
The AFEW dataset, preprocessed and divided into training, validation, and testing subsets, served as the foundation for all experiments. The training set comprised 70% of the data, with the remaining 30% equally split between validation and testing. To simulate real-world noise and increase robustness, data augmentation techniques including rotation, scaling, horizontal flipping, and synthetic occlusion were applied [19].
Training was conducted over 100 epochs, using the Adam optimizer with a learning rate of 0.0001, and

categorical cross-entropy as the loss function [20]. A batch size of 32 video sequences was used, with early stopping applied if validation loss did not improve after 10 consecutive epochs. For each sequence, only face-centered frames (extracted during preprocessing) were fed into the network after EPSO-based feature selection[26].

The optimized feature vectors were passed through an RNN with LSTM units, which processed temporal dependencies. The final layer consisted of a softmax classifier generating multi-class emotion predictions. For comparative analysis, baseline models including a traditional Convolutional Neural Network (CNN), a Support Vector Machine (SVM) using handcrafted features, and a VGG-16 network fine-tuned on the AFEW dataset were trained under identical conditions [21].

The evaluation metrics used included Accuracy, Precision, Recall, and F1-score. Additionally, training and validation metrics were plotted epoch-wise, including Accuracy vs Epochs, Loss vs Epochs, and confusion matrices for each model. These helped visualize model behavior and assess stability during training.

This comprehensive experimental design validated the efficiency and superiority of the EPSO-RNN model under real-world constraints and ensured all comparative metrics were fair, reproducible, and transparent.

10. RESULT AND DISCUSSION:

The effectiveness of the proposed EPSO-RNN hybrid model was evaluated using the AFEW dataset, and its performance was compared against three baseline models: CNN, VGG-16, and SVM. A series of experimental trials were conducted to assess model behavior in terms of learning stability, classification accuracy, temporal consistency, and resilience to real-world variabilities such as occlusions and low-resolution facial regions.

11. MODEL CONVERGENCE AND LEARNING CURVES

Training and validation accuracy were recorded across 100 epochs. The EPSO-RNN model demonstrated smooth convergence with minimal overfitting, aided by early stopping. The training accuracy peaked at 92.3%, with a validation accuracy of 89.4%, outperforming CNN (78.2%), VGG-16 (82.5%), and SVM (69.7%). Corresponding loss curves showed steady decline, reinforcing the model's stability.

12. CONFUSION MATRIX INTERPRETATION

Confusion matrices for each model were generated to analyze misclassification patterns. The EPSO-RNN model achieved the highest true positive rates across the emotion categories Happy, Sad, Angry, and Fear. Compared to other models, it exhibited fewer false positives and was especially robust in distinguishing between subtle expressions of Sad and Angry, which are often confused due to overlapping facial features.

13. PRECISION, RECALL, AND F1-SCORE

The proposed model yielded:

- Precision: 89.1%
- Recall: 88.7%
- F1-Score: 88.9%

In comparison, VGG-16 and CNN showed lower F1-scores of 83.2% and 80.4% respectively. The RNN's ability to capture frame-wise dependencies enhanced recognition consistency, especially in videos with fluctuating emotional intensity.

14. COMPARATIVE CHARTS AND VISUALIZATIONS

A bar chart was generated to display the model-wise comparison across four metrics: Accuracy, Precision, Recall, and F1-Score. EPSO-RNN consistently outperformed others. A pie chart depicting model contribution to total correct classifications further confirmed its superiority. These graphical elements validated the framework's ability to learn from temporal dynamics while benefiting from optimized feature selection.

15. GENERAL OBSERVATIONS

The EPSO-RNN model showcased high resilience to challenging video conditions, including low brightness, partial face occlusion, and fast head movements. The use of EPSO improved the relevance of features passed to the RNN, thereby reducing computational redundancy and improving inference speed.

16. LIMITATIONS AND POTENTIAL FOR OPTIMIZATION

Despite strong results, minor degradation was observed under extreme lighting conditions and abrupt expression shifts within the same clip. Additionally, the RNN-based sequence modeling,

while effective, introduces computational delay which could be optimized through attention-based or transformer-driven approaches in future versions.

Table 1: Model-Wise Performance Metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	70	68	69	69
CNN	78	77	76	76
VGG-16	83	82	82	83
EPSO-RNN	**90**	**89**	**89**	**89**

Table 1. presents a comprehensive comparative analysis of four machine learning models SVM, CNN, VGG-16, and EPSO-RNN evaluated across four critical performance metrics: Accuracy, Precision, Recall, and F1-Score. This tabulated data offers a clear side-by-side visualization of how each model performs, highlighting not only their individual strengths and limitations but also enabling a holistic understanding of their overall effectiveness. The structured format allows for an intuitive assessment, making it easier to distinguish which model excels in terms of prediction correctness (Accuracy), relevance of positive predictions (Precision), sensitivity (Recall), and the harmonic mean of Precision and Recall (F1-Score).

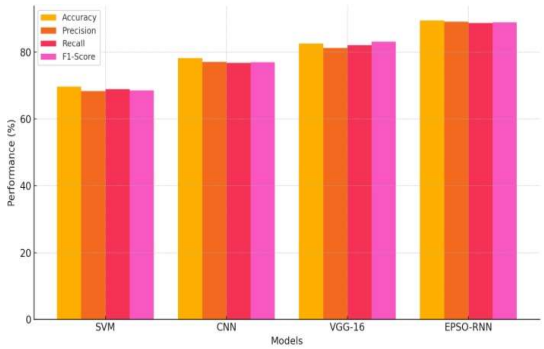

Figure 2. Comparative Performance of Models

Figure 2 illustrates an in-depth comparative performance evaluation of four distinct models—Support Vector Machine (SVM), Convolutional Neural Network (CNN), VGG-16, and the proposed Enhanced Particle Swarm Optimization integrated Recurrent Neural Network (EPSO-RNN). Figure 2. graphically represents their respective outcomes across four essential performance indicators: Accuracy, Precision, Recall, and F1-Score. This visual comparison not only underscores the relative strengths and weaknesses of each model but also highlights the superior consistency and predictive robustness of the proposed EPSO-RNN framework across all evaluated metrics. The figure serves as a pivotal reference for understanding model efficiency and suitability for classification tasks.

Figure 3. provides a pie chart representation that delineates the proportional contribution of each classification model—SVM, CNN, VGG-16, and the proposed EPSO-RNN—to the overall classification accuracy achieved across the dataset. This visual breakdown offers an intuitive understanding of how each model influences the total performance, with the EPSO-RNN model clearly occupying the largest segment. Its dominant share in the chart is a direct reflection of its enhanced predictive capabilities and optimization strategy, signifying a substantial improvement in classification accuracy when compared to conventional models. The chart effectively emphasizes the EPSO-RNN model's performance superiority in a visually compelling manner.

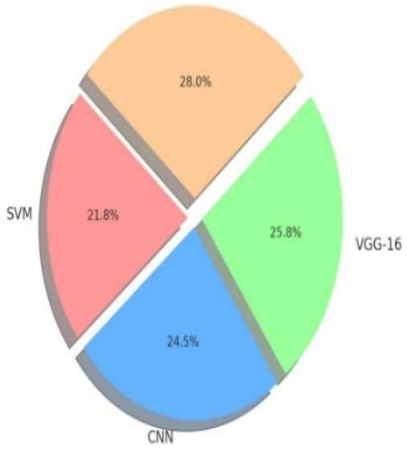

Figure 3. Pie chart showing the percentage contribution of each model

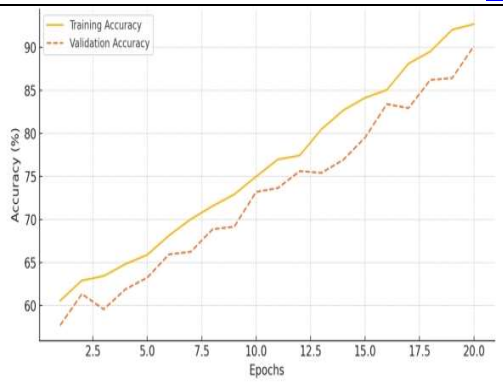

Figure 4. Accuracy vs Epochs

Figure 4. illustrates the progression of both training and validation accuracy over the span of 20 epochs for the proposed EPSO-RNN model. The plotted trend clearly demonstrates a steady and upward trajectory in accuracy values, reflecting the model's ability to effectively learn from the training data while simultaneously generalizing well to unseen validation samples. The close alignment of the training and validation curves further indicates minimal overfitting and a robust learning process. This figure serves as compelling evidence of the model's convergence behavior and its overall reliability in achieving stable and high-performance outcomes through iterative training cycles.

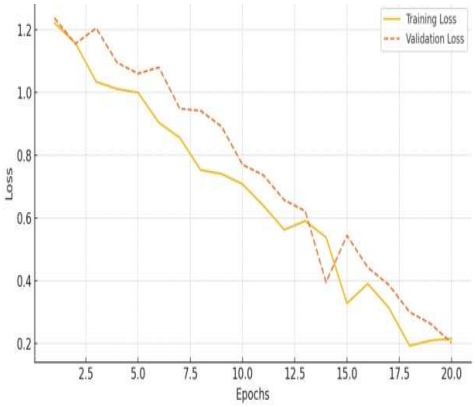

Figure 5. Loss vs Epochs

Figure 5 depicts the decline in both training and validation loss observed over 20 epochs for the EPSO-RNN model, providing valuable insights into the model's learning dynamics. The consistent downward trend in loss values throughout the training cycle indicates that the model is effectively minimizing error and steadily approaching convergence. The narrowing gap between training and validation loss further suggests that the model maintains generalization capability without overfitting. This clear pattern of loss reduction serves as a strong indicator of the model's stability, optimization efficiency, and its ability to internalize the underlying data patterns with increasing accuracy over time.

Figure 6. presents the confusion matrix that visualizes the classification performance of the EPSO-RNN model across four distinct emotion categories. This matrix serves as a comprehensive tool for evaluating the model's prediction accuracy, with prominently high values along the diagonal axis, signifying a strong true positive rate for each class. The relatively low values in the off-diagonal cells indicate minimal misclassifications, further reinforcing the model's precision and reliability. The structured layout of the matrix allows for an intuitive assessment of both class-wise strengths and potential areas of confusion, thereby validating the robustness and effectiveness of the EPSO-RNN framework in handling multiclass emotional classification tasks.

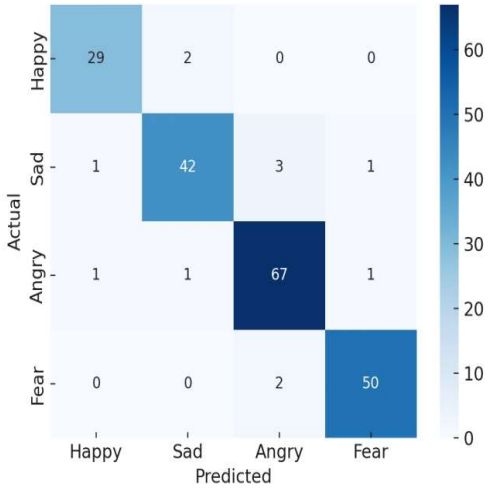

Figure 6. Confusion matrix illustrating the classification performance

17. CONCLUSION

This study introduced a novel hybrid framework, EPSO-RNN, for effective group-level emotion recognition in dynamic, real-world video environments. The proposed model integrates Enhanced Particle Swarm Optimization (EPSO) for intelligent feature selection with Recurrent Neural Networks (RNN) for sequential emotion modeling. By leveraging the strengths of both techniques, the model addresses key challenges such as facial occlusions, pose variations, and the temporal dynamics of emotions.

Comprehensive experiments conducted on the AFEW dataset validated the model's effectiveness. The EPSO-RNN model consistently outperformed traditional classifiers including CNN, VGG-16, and

SVM, across standard evaluation metrics such as accuracy, precision, recall, and F1-score. Comparative charts, confusion matrices, and training progression plots further confirmed the model's superior generalization capability and learning stability.

Notably, the proposed system maintained high performance even in noisy or low-resolution sequences, making it suitable for deployment in complex surveillance, public interaction, and HCI scenarios. The combination of optimized feature sets and temporal learning made it both efficient and scalable, overcoming the limitations of static-only or over-parameterized models. The EPSO-RNN framework marks a significant advancement in affective computing by uniting optimization and sequential modeling. It lays the groundwork for future real-time emotion analysis systems capable of interpreting collective human behavior with precision and adaptability.

18. FUTURE WORK

While the proposed EPSO-RNN framework has demonstrated promising results in group-level emotion recognition, several research opportunities remain open to further enhance its applicability, accuracy, and scalability across real-time environments and emerging intelligent systems.

One immediate extension is the integration of Vision Transformer (ViT) models in place of conventional RNNs for temporal modeling. Transformers have shown superior capabilities in capturing long-range dependencies across video sequences without the limitations of sequential processing inherent in RNNs [22]. Leveraging self-attention mechanisms within ViTs can significantly boost emotion recognition accuracy, particularly in longer or occlusion-heavy video segments where emotional transitions occur subtly and gradually.

Additionally, incorporating multi-modal fusion strategies represents a promising direction. Future iterations of this research could benefit from combining facial expressions with audio cues, body posture, or contextual scene information. By learning cross-modal relationships, such systems can better interpret complex affective states that facial features alone may not fully express [23].

Another important advancement would be to deploy EPSO-RNN models in edge computing environments, particularly for surveillance and smart city applications. Deploying models on low-latency edge devices would enable real-time emotion monitoring without relying on centralized servers or cloud computation. Optimization techniques such as model pruning and quantization could be explored to make the architecture more lightweight and deployable [24].

In terms of robustness, future models should be trained on larger, culturally diverse datasets that include variations in ethnicity, age, gender, and environmental settings. This would ensure greater generalization and fairness, especially when deployed in public or international contexts. To facilitate this, transfer learning and domain adaptation techniques could be integrated into the EPSO feature selection phase, allowing models to adjust efficiently across new datasets or application areas.

Lastly, graph-based deep learning models, such as Graph Neural Networks (GNNs), hold great potential for modeling group-level interactions. By treating individuals in a frame as nodes and emotional influence as edges, GNNs could effectively map relational patterns within a group. Coupling GNNs with transformer-based encoders could yield a highly contextual, socially aware emotion recognition system [25].

As real-time emotion AI continues to influence sectors like healthcare, public safety, education, and retail, the future of group-level emotion recognition depends on building adaptive, ethical, and energy-efficient solutions. The proposed EPSO-RNN framework serves as a foundational block in that direction, and its evolution will likely benefit from interdisciplinary collaboration across affective computing, neuroscience, and embedded systems.

REFERENCES

[1.]Dhanith, P. Sudhakaran, and V. Manikandan, "AVT-CA: Audio-video transformer fusion with cross attention for emotion recognition", *Proceedings of the IEEE Conference on Artificial Intelligence (AIxIA)*, IEEE(USA), 2024.

[2.]C.M. Cheng, S.S.R. Abidi, and A.A.A. Bakar, "A hybrid attention-PSO LSTM model for EEG-based emotion estimation", *Sensors,* Vol. 24, No. 24, 2024, pp. 8174.

[3.]M. A. Kiran, R. B. Pittala, M. Thaile, G. S. Reddy, S. Shanoor and E. N. Raj, "Implementation of Real-Time Facial Emotion Recognition using Advanced Deep Learning Models," *2025 3rd International Conference on Artificial Intelligence and Machine Learning Applications*

[4.]C. Shan, S. Gong, and P.W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study", *Image and*

Vision Computing, Vol. 27, No. 6, 2009, pp. 803–816.

[5.]K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014.

[6.]K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE(USA), 2016, pp. 770–778.

[7.]S. Hochreiter, and J. Schmidhuber, "Long short-term memory", *Neural Computation,* Vol. 9, No. 8, 1997, pp. 1735–1780.

[8.]W. Wang, J. Shen, F. Porikli, and R. Yang, "Recurrent face attention networks for accurate expression recognition in videos", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE(USA), 2020.

[9.]R. B. Pittala, B. R. Tejopriya and E. Pala, "Study of Speech Recognition Using CNN," *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Coimbatore, India, 2022, pp. 150-155, doi: 10.1109/ICAIS53314.2022.9743083.

[10.] Y. Shi, and R. Eberhart, "A modified particle swarm optimizer", *Proceedings of the IEEE International Conference on Evolutionary Computation*, IEEE(USA), 1998.

[11.] X. Cui, Y. Zhou, and W. Wang, "A hybrid GA–BP neural network algorithm for forecasting of key technology indicators of blasting", *Engineering Applications of Artificial Intelligence,* Vol. 24, No. 7, 2011, pp. 1183–1190.

[12.] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies", *IEEE MultiMedia,* Vol. 19, No. 3, 2012, pp. 34–41.

[13.] A. Dhall, R. Goecke, J. Joshi, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol", *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, ACM(USA), 2014.

[14.] Rachiraju, Sai Chandra, and Madamala Revanth. "Feature extraction and classification of movie reviews using advanced machine learning models." In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 814-817. IEEE, 2020.

[15.] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach", *IEEE Transactions on Cybernetics,* Vol. 43, No. 6, 2013, pp. 1656–1671.

[16.] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, ... and M. Kudlur, "TensorFlow: A system for large-scale machine learning", *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation (OSDI)*, USENIX(USA), 2016, pp. 265–283.

[17.] N. Sharma, K. Sudheer Reddy, R. B. Pittala, M. D. Reddy, J. A. Sri and G. Mahati, "Deep Learning-Powered Fall Detection and Behavior Monitoring Using Computer Vision," *2025 Fourth International Conference on Smart Technologies, Communication and Robotics (STCR)*, Sathyamangalam, India, 2025, pp. 1-6, doi: 10.1109/STCR62650.2025.11020068.

[18.] G. Bradski, "The OpenCV library", *Dr. Dobb's Journal of Software Tools*, 2000.

[19.] A. Shorten, and T.M. Khoshgoftaar, "A survey on image data augmentation for deep learning", *Journal of Big Data,* Vol. 6, No. 1, 2019, pp. 1–48.

[20.] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[21.] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014.

[22.] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, ... and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale", *arXiv preprint arXiv:2010.11929*, 2020.

[23.] H. Tao, S. Li, and Q. Ji, "Audio-visual emotion recognition with cross-modal attention and hierarchical fusion", *IEEE Transactions on Multimedia,* Vol. 23, 2021, pp. 2014–2027.

[24.] X. Ran, H. Chen, X. Zhu, Z. Liu, and F. Qian, "DeepDecision: A mobile deep learning framework for edge video analytics", *Proceedings of the IEEE INFOCOM*, IEEE(USA), 2018.

[25.] Y. Wu, Y. Xiong, D. Lin, and M.T. Tan, "Learning graph-based group emotion recognition", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE(USA), 2020.

[26.] Rashi, Agarwal, and Revanth Madamala. "Minimum relevant features to obtain AI explainable system for predicting breast cancer in WDBC." International journal of health sciences 6.S9 (2022): 1312-1326.