Traitement du Signal
Vol. 43, No. 1, February, 2026, pp. 361-373
Journal homepage: http://iieta.org/journals/ts

PsySFA-Net: A Lightweight Psychological Face Analysis Framework Using Spatial-Frequency Attention for Emotion Detection from Facial Images

Thilaga Meena K.[1*], Raja R.[2], Regina S.[3]

[1] Department of ECE, Saveetha Engineering College, Chennai 602105, India
[2] Department of ECE, Pandian Saraswathi Yadav Engineering College, Sivagangai 630561, India
[3] Department of ECE, Velammal College of Engineering and Technology, Madurai 625009, India

Corresponding Author Email: thilagameenakece@gmail.com

https://doi.org/10.18280/ts.430125

Received: 10 November 2025
Revised: 1 January 2026
Accepted: 22 February 2026
Available online: 28 February 2026

Keywords:
facial expression analysis, PsySFA-Net, attention mechanisms, multihead attention mechanism

ABSTRACT

Facial expression analysis plays a major role in understanding human psychological states. Face expression classification is used for numerous applications like mental health assessment, human-computer interface and affective computing. In this work, a novel deep learning architecture called PsySFA-Net is proposed for emotion recognition and psychological face analysis. This model includes dynamic feature fusion with attention mechanisms and emotion suppression enhancement to improve classification performance. For feature extraction, it involves three different blocks: Fractal, Spectral, and Psychological Attention modules. The extracted features are combined using a dynamic feature fusion layer with a multihead attention mechanism to dynamically combine the extracted features. Also, the model uses emotion suppression through an attention-guided mechanism for selective suppression of emotional noise to improve feature clarity and enhance model robustness. Experimental results show that the proposed PsySFA-Net achieves better performance on three different datasets of RAF-DB, FER-2013, and CK+, respectively.

1. INTRODUCTION

Emotion recognition and psychological face analysis are interdisciplinary fields that use facial expressions to understand human emotions and mental states [1]. This analysis is an essential field of research with far-reaching applications across multiple domains like healthcare, human-computer interaction, and mental health monitoring. The ability to accurately detect emotions from facial expressions is used for personalised care and stress detection. However, emotion recognition remains a complex challenge due to the variability of emotional expressions.

Existing emotion recognition systems are based on hand-crafted features [2]. The facial landmarks like texture patterns and geometric features are manually selected and processed to identify emotions [3]. But these methods have failed to reach higher accuracy due to the factors of the complexity of human emotions and individual differences in facial expressions. Also, the difficulties in handling dynamic changes affect the accuracy of detecting emotional states. The manual feature extraction techniques struggled to adapt to new and unseen data.

Recently, deep learning (DL) models [4-6] like convolutional neural networks (CNNs) have significantly advanced emotion recognition by learning hierarchical feature representations from raw data. These models are capable of automatically extracting discriminative features from large datasets. Despite their success, traditional DL models face several limitations. These models still struggle with various factors. It involves noise, differences in lighting, head pose and occlusions etc. In addition, DL models are less effective at capturing subtle or suppressed emotions. This model requires more sophisticated feature extraction techniques.

The DL model is capable of extracting texture and geometric features from facial images. But it failed to capture contextual psychological cues. The p**sy**chological cues refer to subtle and implicit facial patterns. It denotes the emotional and cognitive states beyond obvious expressions. In computer vision, these cues are represented through attention-weighted feature maps and fine-grained texture variations. These cues are used for the model to capture both explicit and suppressed emotions. It mainly focuses on important facial regions and reduces irrelevant information. In expression detection, the contextual psychological cues are very important for classification when emotions are suppressed or hidden. These challenges highlight the need for more sophisticated models capable of combining multiple types of information and maintaining robustness against noise and emotional ambiguity.

To solve the above issues, a DL model is proposed in this work based on multiple feature extractions and attention mechanisms. It involves Fractal, Spectral, and Psychological Attention modules to enhance the model's ability to focus on relevant information and to suppress the unwanted emotional noise. Also, the model parameters are tuned using the Modified Bobcat Optimization Algorithm (MBOA) to achieve higher classification accuracy.

2. RELATED WORK

2.1 CNN-based and traditional feature-based methods

This category includes methods based on CNN and handcrafted feature extraction techniques for facial expression recognition. Chitrapu et al. [7] proposed a face recognition model by combining contrast-limited adaptive histogram equalization with a multi-task cascaded CNN for accurate face detection under varying conditions. It achieves 99.91% accuracy on CASIA3D and 98.77% on 105PinsFace, respectively

To address illumination variation and noise issues in expression detection, Karamizadeh et al. [8] proposed a pre-trained VGG-16 model combined with the Boltzmann machine algorithm. It effectively learns intricate facial features. Similarly, Mukhopadhyay et al. [9] proposed a CNN-based classification method combined with different texture feature extraction techniques. In validation, it achieved accuracy levels of 90.4%, 81.2%, and 71.6% for the CK+, JAFFE, and FER2013 datasets, respectively.

Yao [10] adopted VGG-16 for classification, integrating multiscale features for richer feature extraction. The approach is validated on the CK+ and RAF-DB datasets. Grover and Bansal [11] and Xie et al. [12] proposed a CNN model combined with a Dynamic Region Feature Network for facial expression recognition.

Li et al. [13] proposed a convolutional attention-based mechanism for facial microexpression recognition. In validation, it achieves 99.73% accuracy on the SAMM dataset and a 99.4% F1-score on the CASME-II dataset, respectively. Gautam and Seeja [14] applied different feature extraction techniques, such as HOG and SIFT (Scale-Invariant Feature Transform), for facial expression detection.

A two-stage face recognition system based on SVM was proposed by Chandrakala and Devi [15]. Initially, the Histogram of Oriented Gradients (HOG) is used for feature extraction. Then, the cascaded SVM is applied for final facial expression classification. Similarly, Xu et al. [16] developed a model using Sobel edge detection and a Neural Network classifier. A hyperparameter-tuned multi-class support vector machine (SVM) model is proposed by Dagher et al. [17] for face expression classification.

2.2 Transformer-based and attention-based methods

These approaches utilize transformer architectures and attention mechanisms to capture global dependencies and improve feature representation.

Chaudhari et al. [18] proposed a Vision Transformer-based DL model for elderly expression recognition. A two-stage DL model is proposed by Pei et al. [19] for facial expression recognition. Two modules, named the Local Multi-Patch Attention Module and the Global Self-Attention Channel Module, are used for robust feature extraction.

Guo and Xu [20] developed an attention mechanism-based transformer model for expression detection. Yang et al. [21] proposed a Global–Local Feature Fusion Transformer-based model for expression classification. Xu et al. [22] proposed an improved YOLOv8 model for infant facial expression classification. It uses a Swin Transformer backbone and a feature pyramid network to extract multi-scale features.

2.3 Multimodal and hybrid approaches

This category focuses on methods that combine multiple data modalities such as images, signals, and temporal information for improved emotion recognition.

In their study, Xiang et al. [23] developed a multi-model-based facial expression classification model. It combines features from different sources like RGB, near-infrared, and depth maps from different angles. For emotion classification in video data, the hybrid Recurrent Neural Network (RNN) based model is proposed by Manalu and Rifai [24]. Jain et al. [25] proposed a hybrid model based on RNN and ResNet for mental state classification in video sequences.

2.4 GAN-based and data augmentation methods

These methods address data scarcity and variability using generative models for data augmentation and improved training. To solve data insufficiency, Tang et al. [26] proposed a Generative Adversarial Networks (GAN)-based model for complex emotion estimation.

2.5 Optimization and ensemble-based methods

This group includes methods that improve performance through optimization techniques and ensemble learning strategies. Likewise, Shah et al. [27] propose a novel weighted voting method inspired by the Prisoner's Dilemma for expression detection. In their study, Wang et al. [28] presented a Partitioned Random Forest model to improve facial emotion recognition. Bhagat et al. [29] proposed a pretrained ensemble for emotion recognition. Nan et al. [30] developed a MobileNetV1 model for classification.

2.6 Other specialized and application-based methods

These works focus on specific applications, datasets, or domain-specific improvements in facial expression recognition. Marinova et al. [31] implemented a smart glasses device for facial expression detection.

Lin et al. [32] developed a new dataset for infant facial expressions containing 1,500 annotated images. Tarnowski et al. [33] developed two-stage model for emotional state classification based on facial expressions. Ge et al. [34] proposed an occluded expression categorisation network based on the generated countermeasure network. Gong et al. [35] explored a deep neural network model for human expression detection in virtual reality (VR) applications. Bakariya et al. [36] presented a parameter-tuned DL model for human stress detection combined with music recommendation.

A feature-grafting technique was proposed by Qadir et al. [37] to recognise multicultural facial expressions. Yang et al. [38] presented a DL model based on a multi-threshold for improving feature representation in expression detection.

Despite the significant progress in facial expression recognition, existing methods still exhibit notable limitations. CNN-based approaches mainly focus on spatial features. It lacks the ability to capture multi-domain representations. Multimodal approaches improve performance but require additional data sources. Furthermore, most existing models do not effectively address subtle or suppressed emotional cues. To overcome these limitations, the proposed PsySFA-Net uses a multi-domain feature extraction system with dynamic attention-based fusion and emotion suppression.

3. PROPOSED SYSTEM

In this work, a DL model called PsySFA_Net is proposed to handle complex facial expression recognition tasks. It integrates fractal, spectral, and attention-based features into a dynamic fusion mechanism. To achieve higher accuracy, the model parameter is tuned using MBOA. The overall workflow is shown in Figure 1.

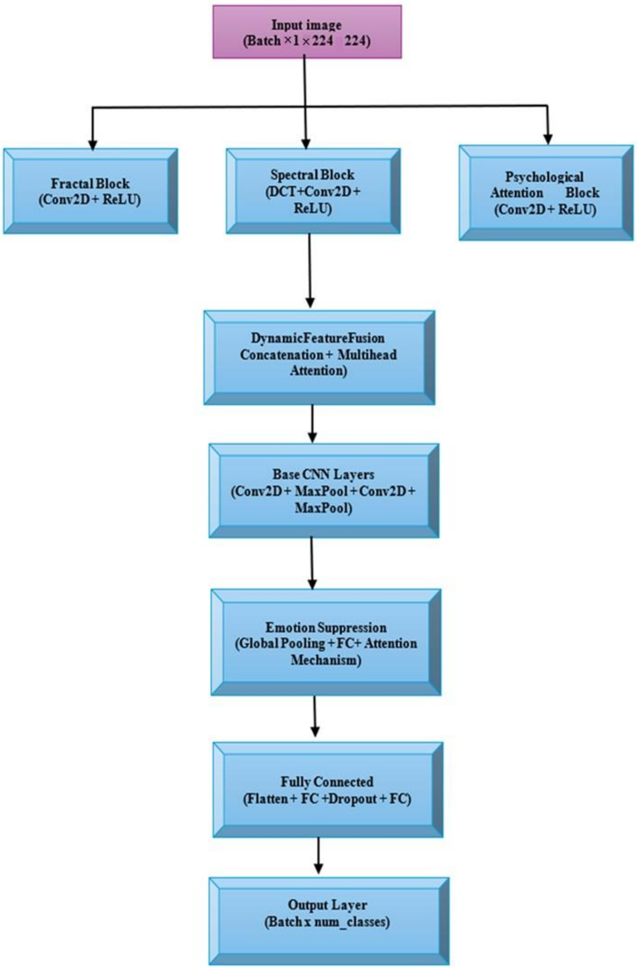

Figure 1. Overall workflow of PsySFA_Net

3.1 Fractal block

The Fractal Block is responsible for capturing the intricate and self-similar patterns of the input facial image. It calculates these texture patterns by performing the absolute difference between diagonally adjacent pixels. This operation highlights the local textural complexity, similar to how fractal analysis identifies patterns that repeat at different scales. The operation can be mathematically represented as:

$$diff(x,y) = |I(x,y) - I(x-1,y-1)| \quad (1)$$

where, *I*(*x*,*y*) is the pixel value at coordinates (x,y).

Then, the module applies a standard convolution and ReLU activation on this difference map to extract higher-level features from the textural information. In this block, the diagonal pixel difference operation is used as a computationally efficient approximation of local self-similarity. It is used to capture directional intensity variations associated with micro-textural changes in facial expressions. The conventional descriptors like Local Binary Patterns (LBP) or Gabor filters are mainly based on predefined patterns and multi-scale filter banks. The proposed operation is learnable and seamlessly integrated within the convolutional system. Also, it has minimal computational overhead and preserves sensitivity to subtle diagonal transitions.

3.2 Spectral block

The Spectral Block analyzes the image in the frequency domain to identify underlying structural information that might be obscured in the spatial domain. It uses the Discrete Cosine Transform (DCT) to convert the image into its spectral coefficients. The 2D DCT of an N×N image I is defined as:

$$C_{uv} = \alpha_u \alpha_v \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} I_{xy}\, cos\left[\frac{\pi u(2x+1)}{2N}\right] \cos\left[\frac{\pi v(2y+1)}{2N}\right] \quad (2)$$

where, $\alpha_u = \frac{1}{\sqrt{N}}\ for\ u = 0\ and \alpha_v = \sqrt{2/N}\ for\ u > 0$

In the Spectral Block, the selection of mid-frequency DCT coefficients is motivated by their ability to capture structural and discriminative facial details. It eliminates redundancy and noise. The low-frequency components primarily represent global illumination and coarse intensity information. These components contribute limited discriminative power for emotion recognition. In contrast, high-frequency components tend to capture fine edges and noise. These components are sensitive to variations such as illumination changes and compression artifacts. The mid-frequency band provides a balanced representation. It preserves essential texture and structural variations relevant to facial expressions.

3.3 Psychological attention

The psychological attention module is a form of spatial attention strategy. This module uses a larger convolutional kernel (7 × 7) to capture a broader contextual understanding of the image. This is based on the idea that human visual perception focuses on salient regions to understand a scene. The larger kernel is used for the network to pay attention to wider regions to identify important features. The convolution operation for this module can be written as:

$$A_{out}(x,y) = ReLU(\sum_{i=-3}^{3} \sum_{j=-3}^{3} W_{ij}.I(x+i,y+j) + b) \quad (3)$$

where, W is the 7 × 7 convolutional kernel and b is the bias. This produces a feature map that highlights the regions deemed most important by the larger receptive field. The choice of a 7 × 7 convolution kernel in the Psychological Attention Module is motivated by the need to capture broader contextual dependencies across facial regions. Compared to smaller kernels, a 7 × 7 kernel provides a larger receptive field and allows the model to integrate information from multiple facial components simultaneously.

The Fractal, Spectral, and Psychological Attention modules interact through a parallel feature extraction and unified fusion mechanism. Each module captures complementary information like spatial textures, frequency-domain structures and context-aware salient regions. Then, these features are mapped into a common feature space and combined using multi-head attention. It learns their interdependencies and assigns adaptive weights.

3.4 Dynamic feature fusion block

The dynamic feature fusion block combines features from multiple sources using a Multi-head Attention Mechanism. This fusion process is used for the model to dynamically focus on the most important features across all modalities. Given feature maps f from the Fractal Block, s from the Spectral Block, and a from the Psychological Attention Block, the features are concatenated:

$$Combined\ features = concat(f.flatten(2), s.flatten(2), a.flatten(2)) \tag{4}$$

The fused feature set is then passed through the Multi-head Attention mechanism:

$$attention\ output = MultiheadAttention(combined\ feature) \tag{5}$$

The resulting output is reshaped and passed forward in the network for further processing.

The attention weights are computed using scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

which captures relationships between fractal, spectral, and psychological features.

In addition, attention map visualization is performed to interpret feature contributions.

3.5 CNN block

The CNN block includes a standard convolutional layer, ReLU activation, and max pooling. This block refines the fused features before passing them into the Fully Connected (FC) layers for classification. The output from the Dynamic Feature Fusion Block is passed through several convolutional layers, each followed by ReLU and Max Pooling:

$$out = ReLU(Cov2D(x)) \tag{7}$$

After passing through these layers, the feature map is downsampled using max pooling and extracts relevant patterns for further processing.

3.6 Emotion suppression block

The emotion suppression block is used to suppress irrelevant or noisy features. In this block, an attention mechanism is applied to focus on the most salient regions. This block uses Global Average Pooling (GAP) followed by an FC layer to produce a suppression factor. The GAP operation computes the average value of each channel:

$$w = GAP(x) \tag{8}$$

The weight w is then passed through an FC layer:

$$w = sigmoid(fc(w)) \tag{9}$$

The suppression factor is applied to the input features using the following attention mechanism:

$$attention\ output = MultiheadAttention(w) \tag{10}$$

The final output is obtained by suppressing less important features:

$$out = x \times (1 - attention\ output) \tag{11}$$

Compared to SE and CBAM attention mechanisms, the proposed block performs *negative attention* by suppressing irrelevant features instead of only enhancing important ones. The SE focuses on channel-wise scaling and CBAM applies sequential channel and spatial attention. The proposed method reduces redundant activations using:

$$x' = x \cdot (1 - \alpha) \tag{12}$$

where, α represents the learned suppression weights. This mechanism improves robustness by attenuating noisy or overlapping emotional features.

3.7 FC layer

The FC layer flattens the output from the previous convolutional block. Then, the flattened output is passed through a series of fully connected layers. These layers are responsible for performing the final classification into one of the target facial expression categories. The flattened features are passed through two Linear layers, with ReLU activation and Dropout for regularisation.

3.8 Modified Bobcat Optimization Algorithm

The Bobcat Optimization Algorithm (BOA) is a novel bio-inspired metaheuristic algorithm designed to solve complex optimization problems. BOA is inspired by the hunting strategy of bobcats for their survival. The optimization of BPA is modelled in two primary phases: exploration and exploitation. The optimization of conventional BOA is based on a single prey (to guide the position updates during the exploration phase. As a result, BOA may become prematurely focused on local optima for high-dimensional problems.

To solve these issues, in this work, a MBOA is proposed. Here, in the exploration phase, the multiple prey for each bobcat is used and a randomised approach is used to update positions.

Stages of MBOA

Initialization

Initialise the population of bobcats in the solution search space. The population represents candidate solutions, each with a position in the solution space. The initial locations of the bobcats are randomly generated within predefined bounds for each decision variable as follows:

$$X \begin{bmatrix} x_1 & \cdots & x_m \\ \vdots & \ddots & \vdots \\ x_N & \cdots & x_m \end{bmatrix} \tag{13}$$

where, X is the population matrix, x_i is the position of the i-th bobcat, and N is the number of bobcats, m is the number of decision variables.

Each bobcat's location is initialised as:

$$x_{i,d} = lbd + r(ubd - lbd) \tag{14}$$

where, *lbd* and *ubd* are the lower and upper bounds for the decision variable d, and *r* is a random number in the interval [0,1].

Exploration Phase (Tracking Prey)

The bobcats move towards the prey (better solutions) in search of optimal areas in the solution space. This phase explores the global search space to identify promising regions. Each bobcat's position is updated by moving towards the best candidate solution (prey) in the population. This mimics the bobcat's behavior of tracking its prey in the wild. The candidate prey set for each bobcat is determined by:

$$CP_i = \{X_k : F_k < F_i \ and \ k \neq i\} \tag{15}$$

where, CP_i is the set of preys for the i-th bobcat, X_k is a bobcat with an improved objective function rate than X_i, and F_k is its objective rate. A new place is identified for each bobcat by simulating its move towards the selected prey:

$$x_{P1_{i,j}} = x_{i,j} + W_{exploration}(t).(\sum_{k=1}^{n} \gamma_k \, (SP_{i,j} - x_{i,j}) \tag{16}$$

where, $x_{P1_{i,j}}$ is the new position of the i-th bobcat in the j-th dimension after the update during the exploration phase. $x_{i,j}$ is the current position of the i-th bobcat in the j-th dimension. $W_{exploration}(t)$ is the exploration weight at iteration t. $SP_{i,j}$ is the j-th dimension of the position of the prey (better solution) for bobcat i. γ_k is the random weight factor between 0 and 1 for each prey, n is the number of prey considered and t is the current iteration of the algorithm.

Each bobcat i considers multiple prey and uses them to update its position in the search space. The weighted sum of the difference between each prey's position and the bobcat's current position is used to compute the new position. The random factor γ_k is used for the bobcat to explore various areas of the search space. The exploration weight $W_{exploration}(t)$ gradually decreases as the algorithm progresses, which shifts the focus towards exploitation in later iterations. If the new position improves the objective function, it replaces the previous position:

$$X_i = X_{P1_i} \ if \ F_{P1i} \leq F_i \tag{17}$$

Exploitation Phase: Chasing the Prey

Once a promising region is found during exploration, the algorithm switches to local search to refine the solution and exploit the potential of the best region. The bobcats simulate a chase behavior and refine their position by performing smaller changes to exploit the local search space. This phase is used to fine-tune the solutions around the best-found area. Each bobcat's new position near the prey is calculated as:

$$x_{P2_{i,j}} = x_{i,j} + (1 - 2.r_i).\frac{1}{1+t}.x_{i,j} \tag{18}$$

where, t is the iteration counter. If the new position improves the objective function, it replaces the previous position:

$$X_i = X_{P2_i} \ if \ F_{P2i} \leq F \tag{19}$$

Repeat the exploration and exploitation phases for several iterations until a termination condition is met. In each iteration, the bobcats update their positions based on the exploration and exploitation strategies. The best solution found so far is updated and stored.

3.9 Hyperparameter tuning using MBOA

In the proposed PsySFA_Net architecture, the hyperparameter contributes major role in classification performance. The important hyperparameters are learning rate, batch size, dropout rate, and number of filters in the convolutional layers. The performance of PsySFA_Net and convergence speed mainly depend on these parameters.

In this work, the MBOA is used to adjust the hyperparameters. The learning rate (LR) controls how quickly the model adjusts its weights in response to errors. The batch size (BS) determines the number of training samples processed before updating the model's parameters. Likewise, the dropout (DR) is used to prevent overfitting by randomly "dropping out" neurons during training. The number of filters (NF) in the convolutional layers controls the capacity of the model to capture complex features from the input image. The fitness function f(LR,BS,DR,NF) is defined as the validation loss of the model trained using these hyperparameters. The objective is to minimise this function since lower validation loss implies better model performance.

$$\begin{aligned} &f(LR, BS, DR, NF) \\ &= ValidationLoss(f(LR, BS, DR, NF)) \end{aligned} \tag{20}$$

where, the validation loss is the average loss computed over the validation set after training the model for a specified number of epochs using a set of hyperparameters.

Pseudocode for MBOA

```
# Step 1: Initialize Population (bobcats)
population_size = 10    # Number of bobcats (solutions)
num_iterations = 5    # Number of iterations for the optimization
lower_bounds = [np.log10(1e-5), 8, 0.1, 8]  # Lower bounds for LR, BS, DR, NF
upper_bounds = [np.log10(1e-2), 64, 0.5, 32]  # Upper bounds for LR, BS, DR, NF
# Step 2: Generate Initial Bobcat Population
Bobcat = initialize_population(population_size, lower_bounds, upper_bounds)
Step 3: Evaluate Fitness of Each Bobcat (using the fitness function)
fitness = []
    lr, batch_size, dropout_rate, num_filters = Bobcat
    # Apply the hyperparameters to train the model and evaluate validation loss
    val_loss = evaluate_model(lr, batch_size, dropout_rate, num_filters)
    fitness.append(val_loss)  # Fitness is the validation loss
# Step 4: Find Best Bobcat (Global Best Solution)
best_ bobcat = find_best_ bobcat (fitness, bobcats) # Find the bobcat with the lowest validation loss
best_fitness = min(fitness) # Best fitness (minimum validation loss)
# Step 5: Iterative Optimization (Echolocation Update)
for iteration in range(num_iterations):
        # Step 5a: Exploration and Exploitation
        explore_or_exploit(bobcat, best_ bobcat)  # Update bobcat's position based on echolocation
            # Step 5b: Update Bobcat Position
```

```
(Hyperparameters)
            update_ bobcat _position(bobcat)
                  # Step 5c: Evaluate fitness again after update
            lr, batch_size, dropout_rate, num_filters = bobcat
            val_loss = evaluate_model(lr, batch_size,
dropout_rate, num_filters)
                   # If new fitness (validation loss) is better, update
the bobcat's position
            if val_loss < best_fitness:
                best_ Bobcat = Bobcat
                best_fitness = val_loss
    # Step 6: Output Best Hyperparameters and Final Fitness
    print("Optimized Hyperparameters:", best_ bobcat)
    print("Minimum Validation Loss:", best_fitness)
```

The MBOA works by initialising a population of bobcat, each representing a potential solution with a random set of hyperparameters within defined bounds. The fitness of each bobcat is evaluated by training the model with its specific hyperparameters and calculating the validation loss; the bobcat with the lowest validation loss is considered the best solution. In each iteration, the bobcat update their positions using a mix of exploration and exploitation. This process is used to improve the overall hyperparameter configuration. After several iterations, the algorithm identifies the optimal set of hyperparameters that minimize validation loss. Finally, the best-performing hyperparameters are output by the optimizer.

4. RESULTS AND DISCUSSION

The proposed PsySFA_Net architecture is coded in Python and implemented using the Collab environment. In this work, three datasets are used for facial expression recognition: RAF-DB, FER-2013, and CK+. Each dataset contains labelled facial images or video sequences categorised into seven basic emotion labels: Happy, Sad, Anger, Neutral, Surprise, Fear, and Disgust. The visualisation of the data set image is shown in Figure 2. Below is Table 1 summarising the number of images or sequences available for each emotion in these datasets.

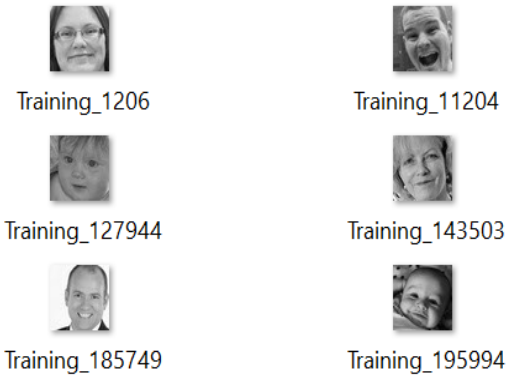

Figure 2. Visualisation of dataset images

The RAF-DB dataset comprises a total of 24,000 images. The FER-2013 dataset includes 35,887 labelled grayscale images. The CK+ dataset contains 1124 Images. The entire dataset is divided into 70% training set and 30% test set. For PsySFA_Net parameter optimization, MBOA is applied before model training. The algorithm was set to run for 60 iterations with a population size of 20 bobcats. Figure 3 shows the optimization process of finding the best hyperparameters for a model. The curve demonstrates a successful hyperparameter tuning process. The initial chaotic behavior shows the exploration phase, and the subsequent stable low-loss region indicates the exploitation phase, where the algorithm has successfully located a set of optimal hyperparameters. The lower and upper bounds of parameters with tuned values are given in Table 2.

Table 1. Dataset statistics

Emotion	RAF-DB (Images)	FER-2013 (Images)	CK+ (Images)
Happy	12,000	7,000	207
Sad	4,000	5,000	112
Anger	4,500	4,000	135
Neutral	3,000	7,000	115
Surprise	3,500	4,000	249
Fear	2,000	4,000	129
Disgust	1,500	2,000	177

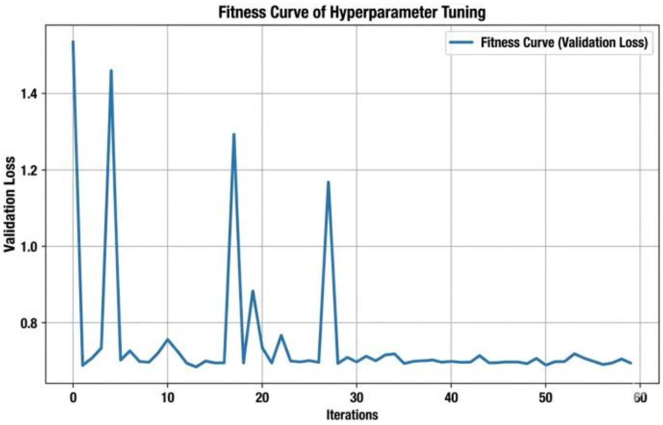

Figure 3. Fitness curve of optimizer

Table 2. MBOA tuned values

Parameter	Lower Bound	Upper Bound	Optimized Value
Learning Rate	1×10−5 (0.00001)	0.01	0.0084
Batch Size	8	64	18
Dropout Rate	0.1	0.5	0.3555
Num Filters	8	32	27

The performance of the model is assessed in terms of accuracy, precision, recall and F-Score rate. It can be computed as follows:

$$Accuray = \frac{T_P+T_N}{T_P+F_P+F_N+T_N} \tag{21}$$

$$Precision = \frac{T_P}{T_P+F_P} \tag{22}$$

$$Recall = \frac{T_P}{T_P+F_N} \tag{23}$$

$$F1\ Score = \frac{2T_P}{2T_P+F_P+F_N} \tag{24}$$

Figure 4 shows the feature maps at various stages of the PsySFA-Net model architecture. The fractal block features highlight the fine-grained texture and self-similar patterns within the facial regions. The spectral block features capture structural and edge information that is robust to illumination changes. The psychological attention features focus on semantically important regions by applying a broader receptive field to capture the overall facial expression.

The final set of feature maps shows the output after the emotion suppression module. Compared to the previous layer, these features appear to have some activations suppressed or "dimmed" in certain regions. It is observed that the module effectively reduces the influence of irrelevant or noisy features.

The model loss and accuracy curves for the datasets given in Figures 5-7. The Loss vs. Epoch graph shows that both the Training Loss and Test Loss (validation loss) decrease rapidly for all epochs. The test loss closely follows the training loss without significant divergence, suggesting that the model is not overfitting to the training data. Both Training Accuracy and Test Accuracy increase sharply for all epochs. The close alignment of the two curves further confirms that the model generalizes well to unseen data.

The ROC curve plots the True Positive Rate against the False Positive Rate. A curve that hugs the top-left corner of the graph indicates a model with high predictive power. For all curves, it is observed that the model can perfectly distinguish between each emotion class and the other classes. The curves for all classes are almost identical and follow the ideal path. The model has perfect classification performance on the validation set for each individual emotion.

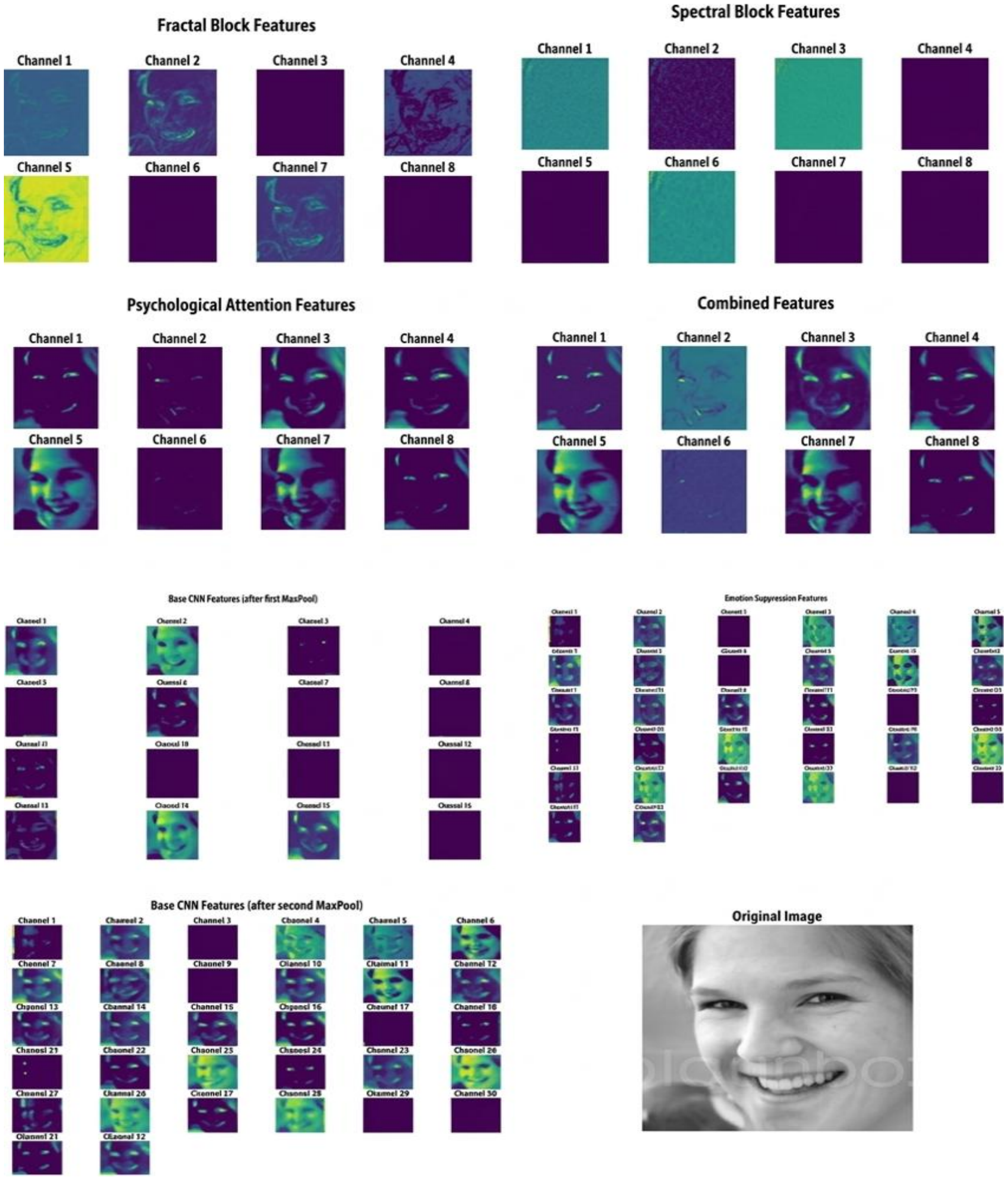

Figure 4. Feature maps of the PsySFA-Net model

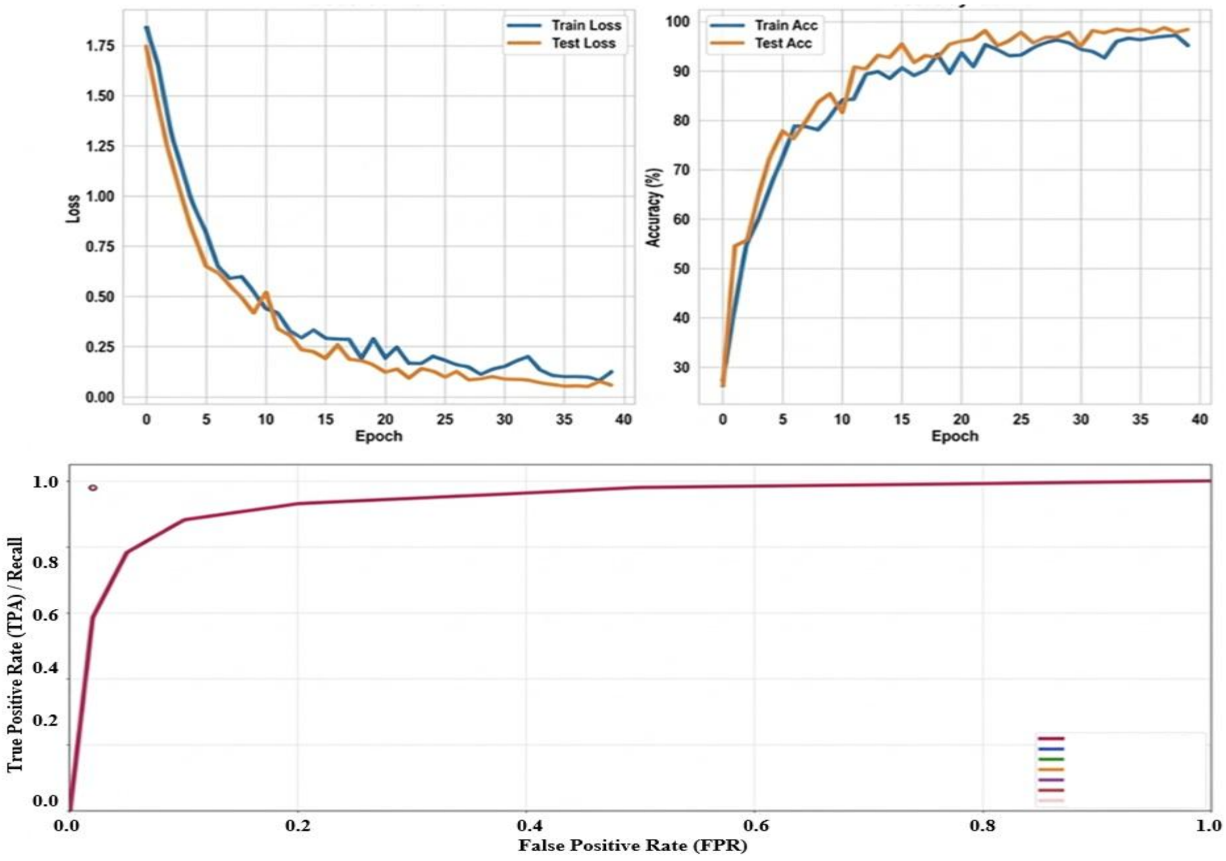

Figure 5. Validation and ROC curve for the RAF-DB dataset

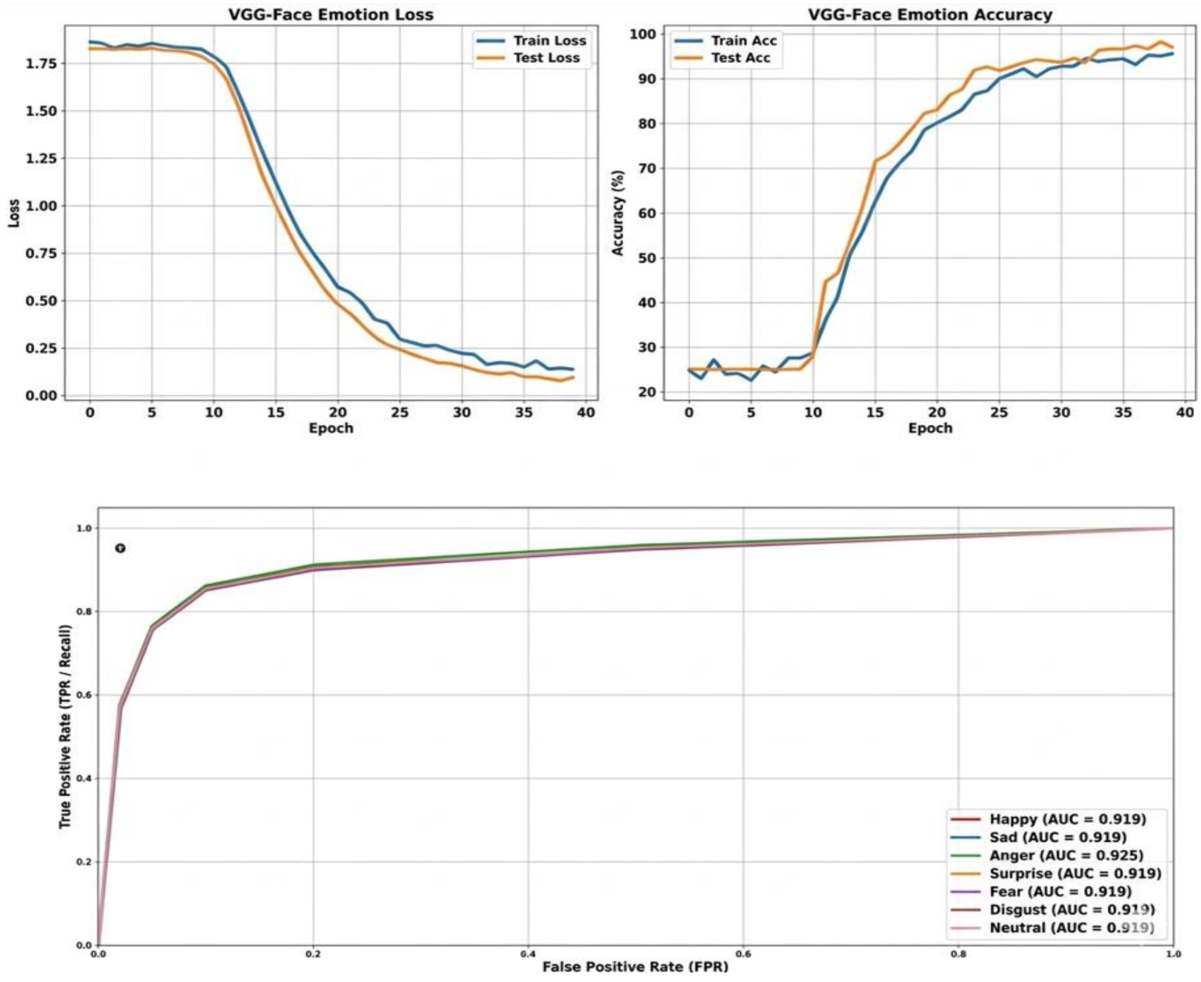

Figure 6. Validation and ROC curve for FER-2013 dataset

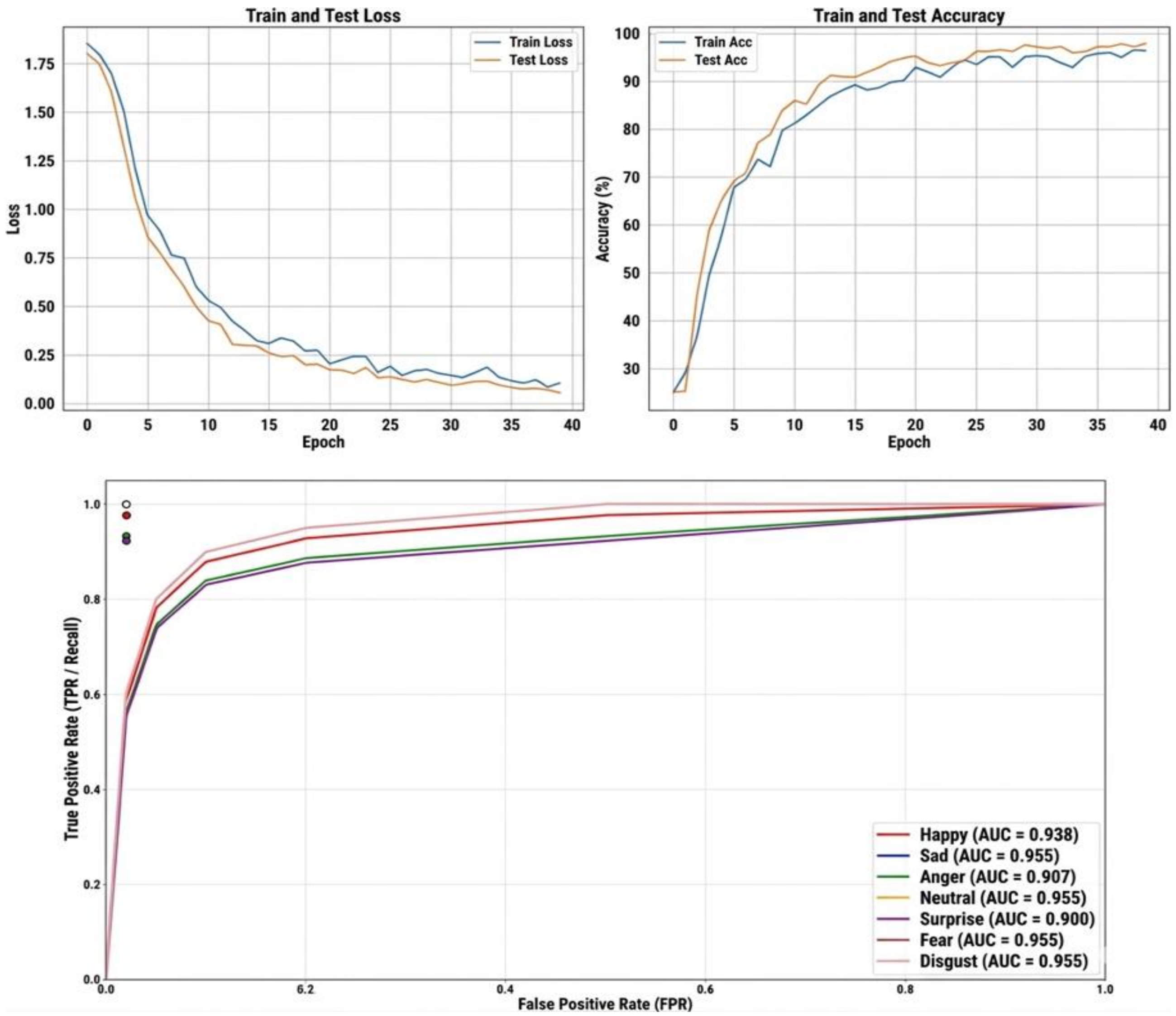

Figure 7. Validation and ROC curve for CK+

Table 3. Performance of PsySFA-Net for RAF-DB dataset

Emotion	Precision	Recall	F1-Score
Happy	0.996	0.980	0.988
Sad	0.981	0.980	0.981
Anger	0.979	0.980	0.979
Neutral	0.968	0.980	0.974
Surprise	0.974	0.980	0.977
Fear	0.959	0.980	0.969
Disgust	0.923	0.980	0.950
Overall accuracy 98.1			

Table 4. Performance of PsySFA-Net for FER-2013 dataset

Emotion	Precision	Recall	F1-Score
Happy	0.971	0.950	0.960
Sad	0.954	0.950	0.952
Anger	0.945	0.958	0.951
Surprise	0.937	0.950	0.943
Fear	0.939	0.950	0.944
Disgust	0.877	0.950	0.912
Neutral	0.966	0.950	0.958
Overall accuracy 94.05			

The results obtained for three different data sets are given in Tables 3-5. In the RAF-DB Dataset, PsySFA-Net achieves an impressive overall accuracy of 98.1% with strong precision and recall across all emotions. In the FER-2013 Dataset, the model reaches an overall accuracy of 94.05% It balances precision and recall well for Neutral and Sad emotions. On the CK+ dataset, PsySFA-Net achieves 97.06% accuracy with better results for emotions like Sad, Fear, and Neutral. The corresponding confusion matrix of the model for datasets given in Figures 8-10.

Table 5. Performance of PsySFA-Net for CK+ dataset

Emotion	Precision	Recall	F1-Score
Happy	0.9836	0.9677	0.9756
Sad	0.9706	0.9706	0.9706
Anger	0.9756	0.9756	0.9756
Neutral	0.9444	0.9714	0.9577
Surprise	0.9733	0.9733	0.9733
Fear	0.9500	0.9744	0.9620
Disgust	0.9808	0.9623	0.9714
Overall accuracy 97.06			

The ablation study of the proposed PsySFA-Net is given in Table 6. The accuracy level is dropped to 97.9% when MBOA is removed. It denotes the contribution of MBOA to tune the parameters. Similarly, the accuracy is reduced to 96.6% when DFF-MHA is removed. The concatenation of features without MHA reduces model performance. The removal of the

Spectral Block (S) and Fractal Block (F) leads to moderate drops in accuracy (97.5% and 97.3%, respectively). Finally, the Backbone-only model shows the lowest performance of 94.7% accuracy. The Psychological Attention module contributes a significant role by guiding the model to focus on contextually important facial regions. The removal of this model reduces the ability to capture subtle emotional cues. The Spectral and Fractal blocks contribute complementary information, and their removal causes moderate degradation due to loss of multi-domain representation.

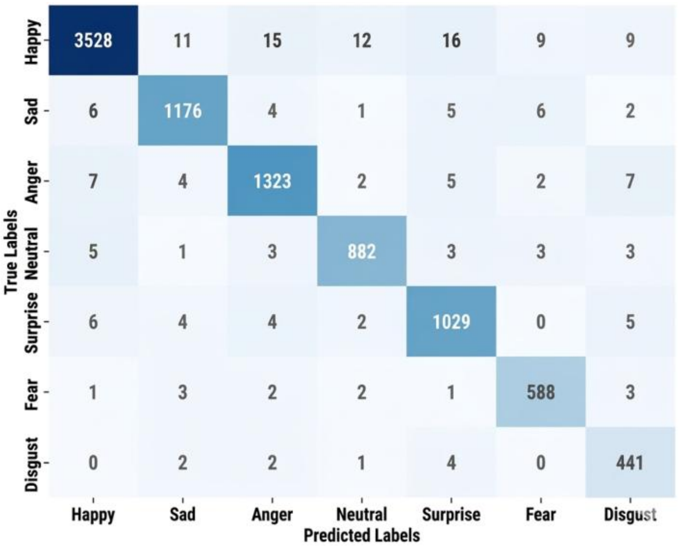

Figure 8. Confusion matrix for RAF-DB dataset

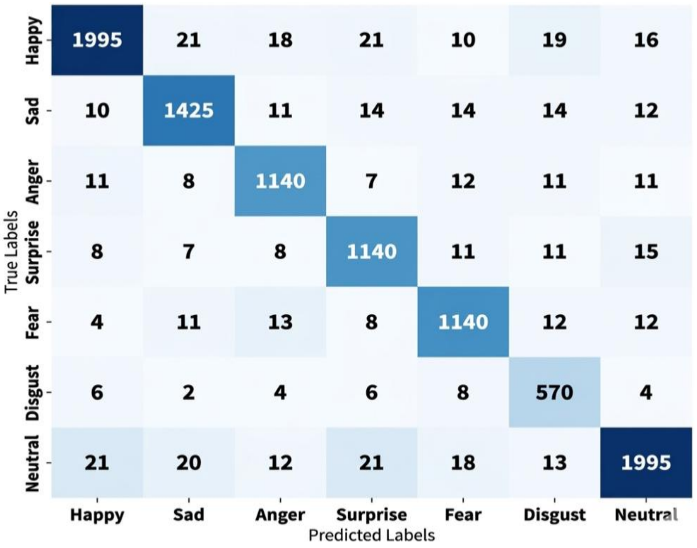

Figure 9. Confusion matrix for FER-2013 dataset

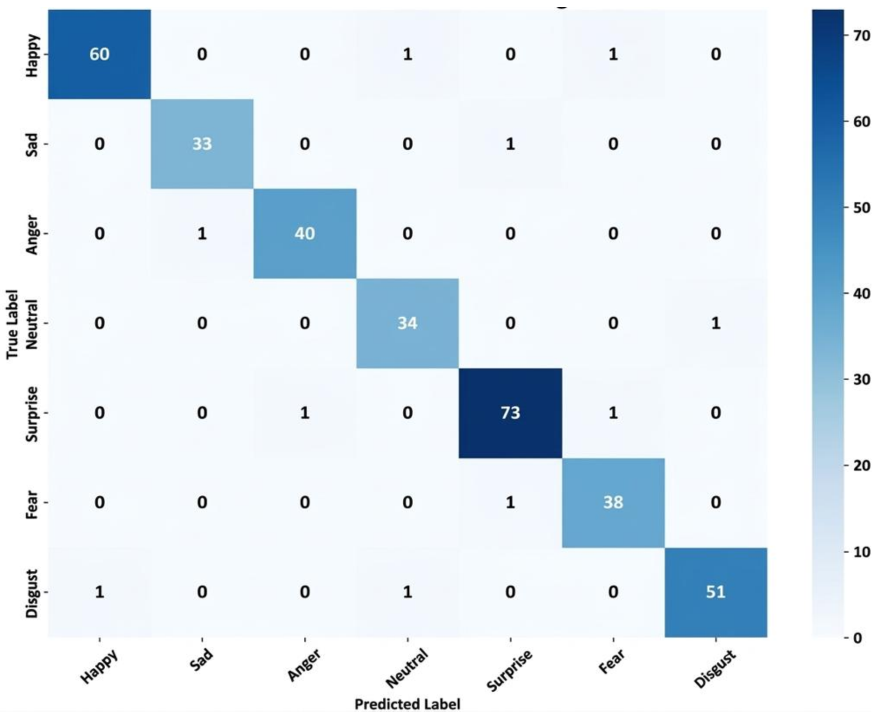

Figure 10. Confusion matrix for CK+ dataset

Table 6. Ablation study of PsySFA-Net

Variant (Remove)	Overall Acc. (%)	Macro Precision	Macro Recall	Macro F1
Full PsySFA-Net (F + S + PA + DFF-MHA + ES+ MBOA)	98.1	0.969	0.980	0.974
– MBOA	97.9	0.967	0.978	0.972
– Dynamic Feature Fusion (use simple concatenation instead of MHA)	96.6	0.953	0.964	0.958
– Psychological Attention (PA)	97.0	0.957	0.968	0.962
– Spectral block (S)	97.5	0.962	0.973	0.967
– Fractal block (F)	97.3	0.960	0.971	0.965
Backbone only	94.7	0.928	0.939	0.933

Notes. F: Fractal module; S: Spectral module; PA: Psychological Attention; DFF-MHA: Dynamic Feature Fusion with multi-head attention; ES: Emotion suppression (attention-guided).

Further, in the spectral block, the effect of frequency band selection is analysed and compared with other selection methods. The results are given in Table 7. The results show that mid-frequency components achieve the highest performance by effectively capturing discriminative structural features and avoiding noise sensitivity. Likewise, the ablation study to evaluate the kernel size impact is given in Table 8. The 7×7 kernel achieves the best trade-off between contextual feature capture and computational efficiency. The smaller kernels fail to capture global dependencies and larger kernels introduce redundancy without significant performance gain. The performance of the model for different feature extraction techniques in the Fractal block is given in Table 9. The results demonstrate that the proposed diagonal pixel difference operation outperforms traditional texture descriptors such as LBP and Gabor filters. The proposed method directly models local intensity variations with lower computational complexity. This enables better preservation of fine-grained facial textures and improves robustness.

Table 7. Effect of frequency band selection in spectral block

Frequency Band Used	Accuracy (%)	Precision	Recall	F1-Score
Low-frequency only	92.8	0.915	0.928	0.921
High-frequency only	91.6	0.903	0.916	0.909
Low + Mid	95.2	0.943	0.952	0.947
Mid-frequency only (Proposed)	96.5	0.958	0.965	0.961
All bands (Low + Mid + High)	95.8	0.949	0.958	0.953

Table 8. Ablation study on kernel size in psychological attention module

Kernel Size	Accuracy (%)	Precision	Recall	F1-Score
3 × 3	94.9	0.931	0.949	0.940
5 × 5	95.7	0.945	0.957	0.951
7 × 7 (Proposed)	96.5	0.958	0.965	0.961
9 × 9	96.3	0.955	0.963	0.959
11 × 11	96.1	0.952	0.961	0.956

The compassion of the proposed model with existing models is given in Table 10. All the models are trained under a standardized setup with 100 epochs. The batch size is set to 18.

The optimiser is the Adam optimizer with identical data augmentation techniques. The learning rate is set to 0.0084.

Table 9. Comparison of texture feature extraction methods in fractal block

Method	Accuracy (%)	Precision	Recall	F1-Score
LBP (Local Binary Pattern)	94.8	0.936	0.948	0.942
Gabor Filter	95.3	0.942	0.953	0.947
LBP + CNN	95.7	0.948	0.957	0.952
Gabor + CNN	96.0	0.952	0.960	0.956
Diagonal Pixel Difference (Proposed)	96.5	0.958	0.965	0.961

Table 10. Comparison with previously proposed models

Model	Accuracy (%)
Vision Transformer (ViT) Model	94.5
Multi-modal Fusion Model	93.5
Weighted Voting	89.5
Two-Stage DL Model	94
GAN-based Model	95
Attention CNN	91.78
ResNet model	95.37
Multimodal Emotion Recognition	94.9
Attention-based Transformer	95.6
Proposed	96.5

Table 11. Statistical significance analysis (accuracy)

Model	Accuracy (%)	Standard Deviation (%)
Vision Transformer (ViT) Model	94.5	0.28
Multi-modal Fusion Model	93.5	0.35
Weighted Voting	89.5	0.45
Two-Stage DL Model	94	0.31
GAN-based Model	95	0.22
Attention CNN	91.78	0.38
ResNet Model	95.37	0.18
Multimodal Emotion Recognition	94.9	0.27
Attention-based Transformer	95.6	0.19
Proposed Model	96.5	0.12

Table 12. Paired t-test Results (Accuracy)

Comparison	p-value	Statistical Significance
Proposed vs. Vision Transformer (ViT)	0.0012	Yes
Proposed vs. Multi-modal Fusion	0.0023	Yes
Proposed vs. Weighted Voting	0.0001	Yes
Proposed vs. Two-Stage DL	0.0035	Yes
Proposed vs. GAN-based Model	0.0048	Yes
Proposed vs. Attention CNN	0.0009	Yes
Proposed vs. ResNet	0.0057	Yes
Proposed vs. Multimodal Emotion Recognition	0.0027	Yes
Proposed vs. Attention-based Transformer	0.0019	Yes

The PsySFA-Net outperforms several previously proposed models and achieves an accuracy of 96.5%. This surpasses other state-of-the-art models like the Attention-based Transformer (95.6%) and GAN-based Model (95%).

To validate the superiority of the PsySFA-Net model, a statistical significance analysis was conducted based on accuracy, as given in Table 11. The models are evaluated over five independent runs to calculate the average accuracy and standard deviation for each model. The baseline models include Vision Transformer (ViT), Multi-modal Fusion, Weighted Voting, Two-Stage DL, GAN-based, Attention CNN, ResNet, Multimodal Emotion Recognition, and Attention-based Transformer.

Paired t-tests are conducted to compare the PsySFA-Net model with other baseline models. The t-tests are performed at a 95% confidence level ($\alpha = 0.05$) to determine if the observed differences were statistically significant. The p-values for the comparison of accuracy are summarized in Table 12. A p-value of less than 0.05 indicates that the proposed model significantly outperforms the corresponding baseline model.

The proposed model demonstrates a clear statistical improvement in accuracy over all baseline models. The p-values below 0.05 across all comparisons indicate that the improvements are statistically significant.

5. CONCLUSION

In this work, a lightweight PsySFA-Net is proposed for emotion detection from facial images. The model significantly enhances classification performance with the integration of different feature learning techniques. The proposed PsySFA-Net model achieves better accuracy of 98.1%, 94.05% and 97.06% on RAF-DB, FER-2013, and CK+ datasets, respectively. The PsySFA-Net stands out with the best overall accuracy of 96.5% when compared with other approaches. The future work will focus on enhancing PsySFA-Net by incorporating more sophisticated attention mechanisms and integrating multimodal data to improve emotion recognition.

REFERENCES

[1] Martinez, A., Du, S. (2012). A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. Journal of Machine Learning Research, 13(5): 1589-1608.

[2] Shan C.F., Gong S.G., McOwan P.W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. Image and vision Computing, 27(6): 803-816. https://doi.org/10.1016/j.imavis.2008.08.005

[3] Carcagnì, P., Del Coco, M., Leo, M., Distante, C. (2015). Facial expression recognition and histograms of oriented gradients: A comprehensive study. SpringerPlus, 4(1): 645. https://doi.org/10.1186/s40064-015-1427-3

[4] Hassouneh, A., Mutawa, A.M., Murugappan, M. (2020). Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods. Informatics in Medicine Unlocked, 20: 100372. https://doi.org/10.1016/j.imu.2020.100372

[5] Jaffar, M.A. (2017). Facial expression recognition using hybrid texture features based ensemble classifier. International Journal of Advanced Computer Science and Applications, 8(6): 449-453. https://doi.org/10.14569/IJACSA.2017.080660

[6] Jiang, D.H., Hu, Y.Z., Dai, L., Peng, J. (2021). Facial expression recognition based on attention mechanism. Scientific Programming, 2021: 6624251. https://doi.org/10.1155/2021/6624251

[7] Chitrapu, P., Morampudi, M.K., Kalluri, H.K. (2025). Robust face recognition using deep learning and ensemble classification. IEEE Access, 13: 99957-99969. https://doi.org/10.1109/ACCESS.2025.3575192

[8] Karamizadeh, S., Chaeikar, S.S., Najafabadi, M.K. (2025). Enhancing facial recognition and expression analysis with unified zero-shot and deep learning techniques. IEEE Access, 3: 43508-43519. https://doi.org/10.1109/ACCESS.2025.3546061

[9] Mukhopadhyay, M., Dey, A., Kahali, S. (2023). A deep-learning-based facial expression recognition method using textural features. Neural Computing and Applications, 35(9): 6499-6514. https://doi.org/10.1007/s00521-022-08005-7

[10] Yao, L. (2024). Facial expression recognition based on multiscale features and attention mechanism. Automatic Control and Computer Sciences, 58(4): 429-440. https://doi.org/10.3103/S0146411624700548

[11] Grover, R., Bansal, S. (2024). Efficient facial expression recognition through lightweight CNN technique on public datasets. SN Computer Science, 6(1): 15. https://doi.org/10.1007/s42979-024-03557-y

[12] Xie, Y., Wang, W., Zhang, Y., Shi, K., Zhang, H., Zhou, N. (2025). Dynamic region features learning for facial expression recognition. Signal, Image and Video Processing, 19(6): 487. https://doi.org/10.1007/s11760-025-04037-3

[13] Li, J., Jin, K., Zhou, D., Kubota, N., Ju, Z. (2020). Attention mechanism-based CNN for facial expression recognition. Neurocomputing, 411: 340-350. https://doi.org/10.1016/j.neucom.2020.06.014

[14] Gautam, C., Seeja, K.R. (2023). Facial emotion recognition using handcrafted features and CNN. Procedia Computer Science, 218: 1295-1303. https://doi.org/10.1016/j.procs.2023.01.108

[15] Chandrakala, M., Devi, P.D. (2021). Two-stage classifier for face recognition using HOG features. Materials Today: Proceedings, 47: 5771-5775. https://doi.org/10.1016/j.matpr.2021.04.114

[16] Xu, R., Huang, A., Hu, Y., Feng, X. (2023). GFFT: Global-local feature fusion transformers for facial expression recognition in the wild. Image and Vision Computing, 139: 104824. https://doi.org/10.1016/j.imavis.2023.104824

[17] Dagher, I., Dahdah, E., Al Shakik, M. (2019). Facial expression recognition using three-stage support vector machines. Visual Computing for Industry, Biomedicine, and Art, 2(1): 24. https://doi.org/10.1186/s42492-019-0034-5

[18] Chaudhari, A., Bhatt, C., Krishna, A., Mazzeo, P.L. (2022). ViTFER: Facial emotion recognition with vision transformers. Applied System Innovation, 5(4): 80. https://doi.org/10.3390/asi5040080

[19] Pei, S., Chen, M., Wang, C. (2024). Multi-scale patch fuzzy decision for face recognition with category information. International Journal of Machine Learning and Cybernetics, 15(10): 4561-4574. https://doi.org/10.1007/s13042-024-02169-5

[20] Guo, D., Xu, F. (2026). STAR-Former: Spatio-temporal adaptive and region-aware transformer for dynamic facial expression recognition. Journal of King Saud University Computer and Information Sciences. https://doi.org/10.1007/s44443-026-00511-1

[21] Yang, W., Yu, J., Chen, T., Liu, Z., Wang, X., Shen, J. (2024). Multi-threshold deep metric learning for facial expression recognition. Pattern Recognition, 156: 110711. https://doi.org/10.1016/j.patcog.2024.110711

[22] Xu, C., Du, Y., Zheng, W., Li, T., Yuan, Z. (2025). Facial expression recognition based on YOLOv8 deep learning in complex scenes. International Journal of Information and Communication Technology, 26(1): 89-101. https://doi.org/10.1504/ijict.2025.144013

[23] Xiang, G., Yao, S., Wu, X., Deng, H., Wang, G., Liu, Y., Peng, Y. (2025). Driver multi-task emotion recognition network based on multi-modal facial video analysis. Pattern Recognition, 161: 111241. https://doi.org/10.1016/j.patcog.2024.111241

[24] Manalu, H.V., Rifai, A.P. (2024). Detection of human

emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm. Intelligent Systems with Applications, 21: 200339. https://doi.org/10.1016/j.iswa.2024.200339

[25] Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., Zareapoor, M. (2018). Hybrid deep neural networks for face emotion recognition. Pattern Recognition Letters, 115: 101-106. https://doi.org/10.1016/j.patrec.2018.04.010

[26] Tang, G., Xie, Y., Li, K., Liang, R., Zhao, L. (2023). Multimodal emotion recognition from facial expression and speech based on feature fusion. Multimedia Tools and Applications, 82(11): 16359-16373. https://doi.org/10.1007/s11042-022-14185-0

[27] Shah, A., Ali, B., Habib, M., Frnda, J., Ullah, I., Anwar, M.S. (2023). An ensemble face recognition mechanism based on three-way decisions. Journal of King Saud University-Computer and Information Sciences, 35(4): 196-208. https://doi.org/10.1016/j.jksuci.2023.03.016

[28] Wang, Y., Li, Y., Song, Y., Rong, X. (2019). Facial expression recognition based on random forest and convolutional neural network. Information, 10(12): 375. https://doi.org/10.3390/info10120375

[29] Bhagat, D., Vakil, A., Gupta, R.K., Kumar, A. (2024). Facial emotion recognition (FER) using convolutional neural network (CNN). Procedia Computer Science, 235: 2079-2089. https://doi.org/10.1016/j.procs.2024.04.197

[30] Nan, Y., Ju, J., Hua, Q., Zhang, H., Wang, B. (2022). A-MobileNet: An approach of facial expression recognition. Alexandria Engineering Journal, 61(6): 4435-4444. https://doi.org/10.1016/j.aej.2021.09.066

[31] Marinova, M., Chona, E., Kotevski, A., Sazdov, B., et al. (2025). Deep learning for facial expression and human activity recognition using smart glasses. IEEE Access, 13: 48257-48270. https://doi.org/10.1109/ACCESS.2025.3551610

[32] Lin, Q., He, R., Jiang, P. (2020). Feature guided CNN for baby's facial expression recognition. Complexity, 2020(1): 8855885. https://doi.org/10.1155/2020/8855885

[33] Tarnowski, P., Kołodziej, M., Majkowski, A., Rak, R.J. (2017). Emotion recognition using facial expressions. Procedia Computer Science, 108: 1175-1184. https://doi.org/10.1016/j.procs.2017.05.025

[34] Ge, H., Zhu, Z., Dai, Y., Wang, B., Wu, X. (2022). Facial expression recognition based on deep learning. Computer Methods and Programs in Biomedicine, 215: 106621. https://doi.org/10.1016/j.cmpb.2022.106621

[35] Gong, Q., Liu, X., Ma, Y. (2025). Real-time facial expression recognition based on image processing in virtual reality. International Journal of Computational Intelligence Systems, 18(1): 8. https://doi.org/10.1007/s44196-024-00729-9

[36] Bakariya, B., Singh, A., Singh, H., Raju, P., Rajpoot, R., Mohbey, K.K. (2024). Facial emotion recognition and music recommendation system using CNN-based deep learning techniques. Evolving Systems, 15(2): 641-658. https://doi.org/10.1007/s12530-023-09506-z

[37] Qadir, I., Iqbal, M.A., Ashraf, S., Akram, S. (2025). A fusion of CNN And SIFT for multicultural facial expression recognition. Multimedia Tools and Applications, 84(28): 33505-33523. https://doi.org/10.1007/s11042-024-20589-x

[38] Yang, D., Alsadoon, A., Prasad, P.C., Singh, A.K., Elchouemi, A. (2018). An emotion recognition model based on facial recognition in virtual learning environment. Procedia Computer Science, 125: 2-10. https://doi.org/10.1016/j.procs.2017.12.003