

## Facial emotion recognition using enhanced multi-verse optimizer method

Ravi Gummula, Vinothkumar Arumugam, Abilasha Aranganathan

Department of Electronics and Communication Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, India

### Article Info

#### Article history:

Received Jul 6, 2023

Revised Sep 29, 2023

Accepted Nov 4, 2023

#### Keywords:

Enhanced multi-verse optimizer  
Machine learning  
Surrey audio-visual expressed emotion  
Text recognition  
YouTube

### ABSTRACT

In recent years, facial emotion recognition has gained significant improvement and attention. This technology utilizes advanced algorithms to analyze facial expressions, enabling computers to detect and interpret human emotions accurately. Its applications span over a wide range of fields, from improving customer service through sentiment analysis, to enhancing mental health support by monitoring emotional states. However, there are several challenges in facial emotion recognition, including variability in individual expressions, cultural differences in emotion display, and privacy concerns related to data collection and usage. Lighting conditions, occlusions, and the need for diverse datasets also impacts accuracy. To solve these issues, an enhanced multi-verse optimizer (EMVO) technique is proposed to improve the efficiency of recognizing emotions. Moreover, EMVO is used to improve the convergence speed, exploration-exploitation balance, solution quality, and the applicability in different types of optimization problems. Two datasets were used to collect the data, namely YouTube and surrey audio-visual expressed emotion (SAVEE) datasets. Then, the classification is done using the convolutional neural networks (CNN) to improve the performance of emotion recognition. When compared to the existing methods shuffled frog leaping algorithm-incremental wrapper-based subset selection (SFLA-IWSS), hierarchical deep neural network (H-DNN) and unique preference learning (UPL), the proposed method achieved better accuracies, measured at 98.65% and 98.76% on the YouTube and SAVEE datasets, respectively.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



### Corresponding Author:

Ravi Gummula

Department of Electronics and Communication Engineering, Dr. M.G.R. Educational and Research Institute  
Chennai, India

Email: ravi.gummula@gmail.com

## 1. INTRODUCTION

Facial expressions serve as the most instinctive form of conveying human emotions and play a vital role in the overall system of expressing emotions. In everyday communication, individuals can quickly communicate their underlying feelings and intentions through facial expressions [1], [2]. Face recognition is always considered an attractive area of research among researchers even today, due to many of its internet of things (IoT) applications [3]. Most popular techniques are now suited for sustaining data utility in terms of pre-defined criteria such as structural similarity or face form, which suffers the censure of insufficient variation and adaptableness [4], [5]. Several techniques use the posture estimation method for single-picture face recognition, but only a minority integrate the pose estimation method in image set-based face identification [6], [7]. The existing studies focused on intended emotion labels, and few also focused on seeming labels for better results to increase emotion recognition accuracy [8]. Multimodal fusion and feature

selection are the two most difficulties in affective computing and multimodal sentiment analysis [9]. The characteristics retrieved from separate modalities are multidimensional, and many of them are redundant and unnecessary [10], [11]. A previously conducted research focused on the reported facial expression classifications of displeasure/anger, sad/unhappy, smiling/happy, afraid, and surprised/astonished [12]. In psychotic syndromes, recognition of positive expressions (happy) is intact, but recognition of negative expressions (fear, anger, disgust, and sadness) is impaired for both positive and negative emotions [13]. Another research focused on overcoming the problems of blurred, low-resolution and noisy images for face detection recognition systems by employing an enhanced multi-verse optimizer (EMVO) technique [14], [15]. Facial emotion recognition is employed for its versatility in various applications. It enhances human-computer interaction, aids mental health monitoring, and improves customer service by gauging emotions [16], [17]. Additionally, it offers valuable insights in fields like market research and advertising, making it a valuable tool for understanding and responding to human emotions [18]. However, while recognizing a face, there are certain challenges such as refined expressions, ambiguous emotions, individual differences, and context dependency. Facial emotion recognition proves to be difficult at times, due to insufficient data and bias issues, making it a continually evolving research. [19], [20]. Hence, this research utilizes EMVO feature selection method to further improve the accuracy in facial emotion recognition.

Kothuri and Rajalakshmi [21] discovered a shuffled frog leaping algorithm-incremental wrapper-based subset selection (SFLA-IWSS) implementation, to enhance emotion detection. The suggested system analyzed the user's emotions using audio, video, text elements and music recommendations to the users/operators. However, the SFLA had limited adaptability towards the dynamic changing nature of the emotion recognition tasks, thus, modifications were required to enhance the algorithm's adaptability under such scenarios. Singh *et al.* [22] developed a hierarchical deep learning-based technique to collect context-dependent emotional elements from the text of both multimodal and unimodal SER systems. Combination of words, spectral information and voice quality (VQ) were used by both global and local level audio descriptors to represent how humans perceive emotion. However, the hierarchical deep learning technique required large amounts of labelled training data to generalize the specific domains. Lei and Cao [23] created novel preference learning algorithms for multimodal emotion recognition that overcame the disparity between alleged labels on different modalities. However, preference learning algorithms often require labelled data in the form of recognition. Acquiring such data is challenging and time-consuming, especially when the number of data is large. Singh *et al.* [24] designed a two-dimensional convolutional neural network (CNN) to identify the most effective features for the task. However, the suggested CNN-based method encountered difficulties in capturing the essential temporal dynamics required for precise emotion recognition. Mao *et al.* [25] implemented an improved CNN called SCAR-NET, to extract features from speech signals, and perform classification. However, SCAR-NET significantly increased the computational complexity of the network. As a result, an Improved multi-verse optimizer is presented in this research to increase the features of emotion identification. As compared to existing methods in emotion recognition, the suggested method achieved superior performance in accuracy. However, the Facial emotion recognition is a challenging task due to the vast amount of various known facial expressions, which are assessed under varying lighting and orientations. So, the existing methods struggled to select optimal features and navigate high-dimensional spaces effectively. Additionally, achieving perfect accuracy often requires a combination of sophisticated feature extraction, model design, and robust data handling. The limitations of the existing methods, coupled with dataset diversity and challenges in emotion recognition, makes it tougher to accomplish better performances on datasets like YouTube and SAVEE. EMVO effectively explores and exploits the solution space in facial emotion recognition, inspired by cosmological principles. Cosmology's multiverse theory uses diverse solution sets in different universes to explore facial emotion feature spaces. Cosmological principles perform efficient exploration, leading to enhanced accuracy, by identifying optimal feature sets for improved emotion recognition models. EMVO offers a more systematic search of features, helping in optimal feature selection and model design. Its balanced approach to exploration and exploitation enhances accuracy by potentially finding more discriminative features. EMVO's ability to efficiently navigate high-dimensional feature space, addresses the limitations in feature selection, leading to better accuracy in facial emotion recognition compared to existing methods.

The contribution of the research are as follows; i) In this research, an enhanced multi-verse optimizer (EMVO) approach is developed for feature selection method to improve the accuracy of facial emotion recognition, ii) EMVO is crucial for facial emotion recognition due to its ability to efficiently explore high-dimensional feature spaces and is therefore used to further improve the accuracy in the facial emotion challenging task, iii) EMVO offers an advantage in emotion recognition by enabling faster and more accurate assessment of facial expressions, voice tonality, and text sentiment. This proves to be useful in applications like customer sentiment analysis, and more precisely helps in such emotion monitoring techniques. The rest of this paper is organized as follows: section 2 illustrates the proposed method. Section 3

denotes the suggested enhanced multi-verse optimization method. The implementation details and its results are demonstrated in section 4. The conclusion of this research is given in section 5.

**2. PROPOSED METHOD**

In this section, the proposed multi-modal facial expression recognition and classification is analyzed. The proposed method is implemented and simulated using MATLAB R2020b, with the system configuration of i7 processor, 16 GB RAM and Windows 10 OS. The data collected from the datasets undergo pre-processing, feature selection, feature extraction and classification, before giving it to the proposed method, as briefly explained below. The flow diagram of the proposed method is shown in Figure 1.

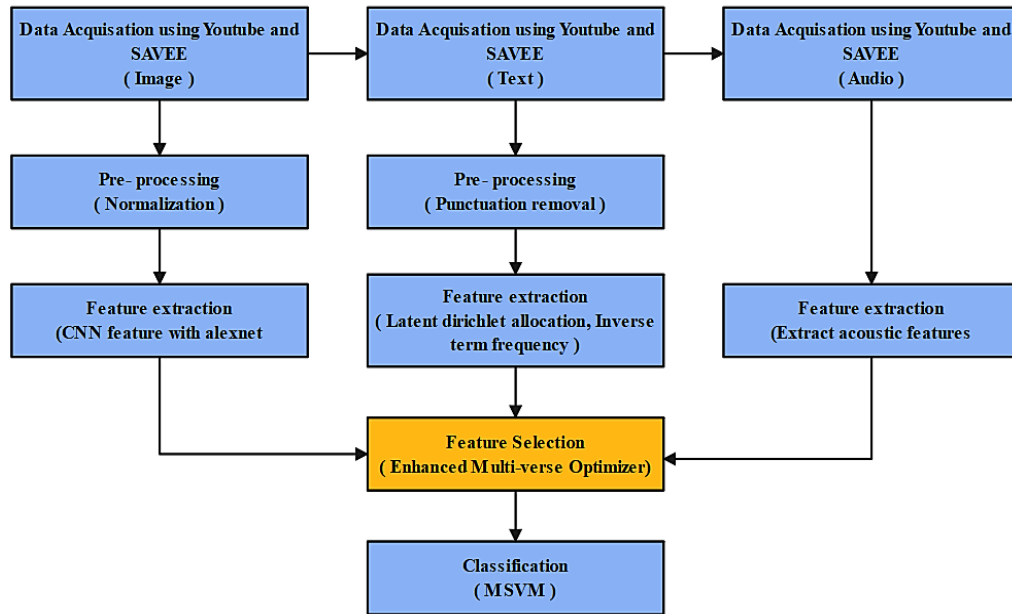


Figure 1. Flow diagram of the suggested techniques

**2.1. Data collection**

In this research, two datasets are used to collect the data, namely YouTube and surrey audio-visual expressed emotion (SAVEE) datasets [21], which are used to evaluate the suggested method’s performance. These datasets are used to create automated emotion identification systems, that support image, video, text and audio. Figure 2 illustrates sample emotions within distinct classes in the SAVEE and YouTube dataset.

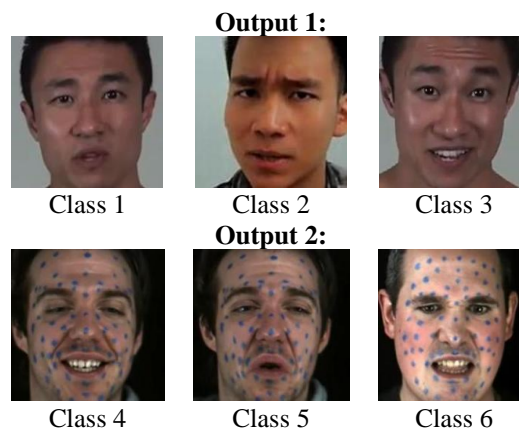


Figure 2. Various emotions detection for the SAVEE dataset

SAVEE dataset contains 480 total words, 3 standards, 2 emotion-specific, and ten general sentences. The YouTube dataset is made up of 7 different emotions: disgust, anger, happiness, fear, surprise, sadness, and neutral, all of which were recorded and assessed by 10 different individuals. After data collection from these datasets, pre-processing is employed to restructure and prepare the data for further analysis or for training the model. This involves various steps that addresses issues such as missing values, outliers, and feature scaling, ensuring that the data is in a suitable format for subsequent tasks.

## 2.2. Pre-processing

After the data collection, normalization and punctuation removal are utilized to pre-process the image data. In this research, data pre-processing is carried out in two ways, one is normalization [26] using image/video and another is Punctuation removal using text. These are briefly explained below.

### 2.2.1. Normalization using image/video

Normalization is a preprocessing approach used to decrease differences in face photographs, such as the degree of lighting, to obtain a better face image. The min-max normalization method is utilized to boost image intensity, resulting in enhanced clarity of information and improved performance of classifiers. The mathematical expression of normalization is illustrated in (1).

$$x = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where  $x_{min}$  describes the image minimum intensity,  $x_{max}$  describes the image maximum intensity, and  $x_n$  describes the normalized image.

### 2.2.2. Punctuation removal using text

Punctuation signals in web social media include sentiment data which are the primary resource of emotional data, and they have significant usefulness in addressing the absence of text content semantics. After performing pre-processing on the data, feature extraction techniques are applied to the image, text, video, and audio data, to extract relevant and meaningful information from each of these modalities. Once the punctual removal is done, the relevant data is given as input to feature extraction stage, where bidirectional encoder representations from transformers (BERT) model is used which is clearly described in the following section.

## 2.3. Feature extraction using BERT model

In this research, the feature extraction is done using the BERT model for text that is fused into video and audio features [27]. BERT is frequently chosen for facial emotion recognition due to its proven performance, extensive adoption, and abundance of pretrained models, while robustly optimized BERT pretraining approach (ROBERTa) requires enhancement in its adoption and validation for specific domain applications like facial emotion recognition. Utilizing BERT in facial emotion recognition increases comprehensibility of textual emotional context that often accompanies images or videos. BERT's pre-trained transformer model captures intricate language nuances, helping in understanding associated emotions within textual descriptions or comments. This textual context, when fused with corresponding facial images or videos, provides a more comprehensive understanding of the conveyed emotions. By considering both textual and visual elements, the accuracy and depth of facial emotion recognition is significantly improved, enabling a more nuanced and precise interpretation of emotional states and expressions depicted in the visual data. The process of BERT using text feature extraction is briefly discussed below.

### 2.3.1. Text using latent Dirichlet allocation, inverse term frequency, and bag of words

Latent Dirichlet allocation (LDA) is a probabilistic generating approach that is given as random mixes over latent themes. The mathematical expression of LDA is shown in (2). Moreover, inverse term frequency, information extraction, text mining, and weighting factors are primarily employed to prevent filtering terms in the text categorization and classification applications. BERT model is balanced by weights, which helps to prevent certain words from being overused, as illustrated in (3).

$$p(\alpha) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1} \quad (2)$$

$$\Phi_{c,v}^{weight} = \eta \times \left[ \frac{SN_{c,v}}{SN_c} \times \log \left( \frac{C}{C_v} \right) \right] + \gamma \quad (3)$$

where  $\Gamma(x)$  is the gamma function. Bag-of-words (BoW) model is an image representation technique for image categorization and annotation tasks. To fuse text features with audio and video data, it typically follows a multi-modal approach. First, it extracts text features using BERT as described above. Then, it integrates them with audio and video features. The fused representation is then used for various multi-modal tasks, such as sentiment analysis of video reviews, speech-to-text in videos, or video summarization, depending on the specific application. The choice of fusion method depends on the nature of the data and the goals of the task, allowing it to leverage the strengths of BERT's text features alongside audio and video information, for a more comprehensive analysis.

### 2.3.2. Image/video using CNN feature with AlexNet

CNN is a branch of deep learning which is widely used alongside AlexNet to improve data. It includes five parts: a convolution layer, an input layer, an output layer, a pooling layer, and a full connection layer. AlexNet contains 5 convolutional layers, 8 layers of neural networks, 3 pooling layers and 3 full connection layers. The mathematical expression was shown in (4).

$$f(x) = \begin{cases} x, & \text{if } y > 0 \\ 0, & \text{if } y \leq 0 \end{cases} \quad (4)$$

### 2.3.3. Using extracted acoustic features in audio

In this research, acoustic features are combined to increase the semantic data of the emotional features in the audio [27]. After performing feature extraction on the data, the feature selection technique called enhanced multi-verse optimizer (EMVO) is applied to identify and select the most significant features. The mathematical expression of the extracted acoustic features is illustrated in (5).

$$x_{rms} = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 \dots + x_n^2}{n}} \quad (5)$$

## 3. ENHANCED MULTI-VERSE OPTIMIZER

In this research, an innovative approach is introduced for facial emotion recognition through the utilization of the enhanced multi-verse optimizer (EMVO). The multi-verse optimization (MVO) faced limitations in terms of exploration and convergence speed in complex optimization tasks. EMVO addresses these issues by introducing multiple universes inspired by cosmological principles [28]. Cosmology's multiverse theory uses diverse solution sets in different universes to explore facial emotion feature spaces. Cosmological principles perform efficient exploration, leading to enhanced accuracy, by identifying optimal feature sets for improved emotion recognition models.

In the facial emotion recognition of EMVO, features are selected by utilizing a cosmology-inspired optimization approach. In EMVO, the cosmology-inspired approach uses principles from the multiverse theory to enhance the optimization process. Similarly, to multiple universes in cosmology, EMVO employs multiple solutions in various universes to search and optimize facial emotion feature spaces. Each universe represents a potential solution set, that explores diverse feature combinations and configurations. The incorporation of cosmological principles helps guide the exploration strategy, allowing for more efficient and effective search across the solution. By utilizing inspiration from the vastness of the cosmos, EMVO significantly improves the accuracy and robustness of facial emotion recognition models by identifying optimal feature sets. EMVO enhances exploration of the solution space by employing multiple universes and optimizing these universes iteratively. This enhanced exploration helps in identifying the most discriminative facial features. Additionally, the optimization process balances the exploration and exploitation, thereby effectively fine-tuning and optimizing the feature selection procedure. By efficiently navigating the high-dimensional feature space and optimization, EMVO improves the accuracy of facial emotion recognition by selecting the most relevant and discriminative features from the emotional facial expressions. Hence, the enhancements of the EMVO are as follows:

- EMVO incorporates adaptive control parameters. These parameters dynamically adjust during the optimization process, ensuring the algorithm can fine-tune its exploration and exploitation strategies based on the evolving problem state. This adaptability enables EMVO to navigate the solution space more efficiently.

- EMVO introduces a crossover operation inspired by genetic algorithms. This operation facilitates the exchange of information among the universes, promoting diversity within the population of solutions, and helping the algorithm escape local optima.
- The velocity update strategy in EMVO is modified to optimize the movement of universes within the search space. Adjustments in velocities based on fitness values, guide the search towards promising regions, which enhances the convergence speed and accuracy.

Feature selection involves the process of carefully choosing a subset of the most pertinent features from a given set of original features, thereby eliminating redundant, irrelevant, or noisy features. To deal with this problem, a wormhole existence probability (WEP) is considered in the algorithm for iterations that are illustrated in (6).

$$WEP = \begin{cases} 0.2 + (0.8) \frac{t}{t_{max}} & t < \frac{t_{max}}{2} \\ w_{min} + (w_{max} - w_{min}) \cdot \frac{t}{t_{max}} & t \geq \frac{t_{max}}{2} \end{cases} \quad (6)$$

### 3.1. Fitness function for EMVO

The EMVO model, classifies into black holes, white holes, and wormholes. White and black holes represent wormholes, while exploration and exploitation phases correspond to different stages. The fitness function's mathematical expressions are defined in (7).

$$x_i^j = \begin{cases} x_k^j & r_1 < normr(U_i) \\ x_i^j & r_1 \geq normr(U_i) \end{cases} \quad (7)$$

where  $normr(U_i)$  is the normalized inflation rate of  $i^{th}$  universe and  $x_i^j$  is the  $j^{th}$  variable.  $r_1$  is a random number between  $[0,1]$ .  $x_k^j$  is the  $j^{th}$  variable of  $k^{th}$  universe. Following the feature selection process, the CNN classification technique is employed to classify the data based on the selected features. The CNN algorithm utilizes multiple support vector machines to handle complex and multi-class classification problems effectively.

### 3.2. Classification using convolutional neural network

Convolutional neural network (CNN) is a specialized neural network for image analysis, and provides excellent results in facial emotion classification by extracting intricate features and recognizing spatial patterns that are crucial for accurate emotion recognition [29]. Its hierarchical learning process enables automated feature extraction, improving the model's ability to interpret and classify emotions from facial images [30]. When contrasted to a deep artificial neural network (ANN), CNNs have a lesser training time. Therefore, CNNs have shown remarkable success in various computer vision tasks, including image segmentation, identification, as well as classification, due to their ability to capture complex patterns and spatial relationships. CNN segmentation as well as the classification system incorporates convolutional layers, pooling layers, fully connected layers, drop-out layers. Figure 3 shows the basic architecture of CNN.

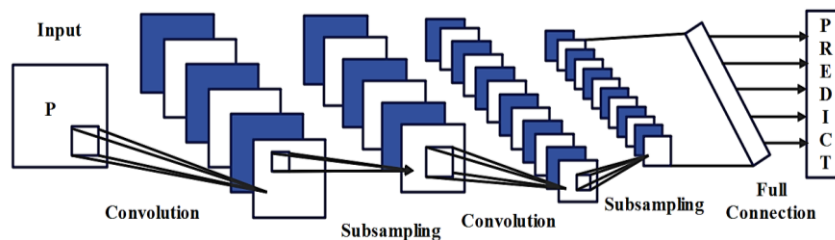


Figure 3. Basic architecture of CNN

In Figure 3, the first layer's feature map is created by combining the input using six convolution kernels. Every convolution kernel is  $5 \times 5$  in size, with a stride of 1. Equations (8) and (9) are used to calculate the feature map size:

$$n_f = \frac{n_i + 2p - f}{s} \quad (8)$$

where,  $n_f$  is feature map size,  $n_i$  is input data size,  $p$  is padding value,  $f$  is kernels size,  $s$  is stride value. The basic formula of a convolution operation is given in (6):

$$a^l = \delta(W^l a^{l-1} + b^l) \quad (9)$$

where,  $a^l$  is  $l^{th}$  convolution layer's output,  $W^l$  is  $l^{th}$  convolution layer's convolution kernel,  $a^{l-1}$  is  $(l-1)^{th}$  convolution layer output,  $b^l$  is  $l^{th}$  convolution layer's bias,  $\delta$  is  $l^{th}$  convolution layer's activation function.

#### 4. EXPERIMENTAL RESULTS

In this research, enhanced multi-verse optimizer is suggested to recognize facial expressions and to increase the hyperparameters of CNN classification. The proposed method is implemented and simulated using MATLAB R2020b with the system configuration of i7 processor, 16 GB RAM and Windows 10 OS. The performance of the feature selection EMVO algorithm is evaluated using common performance measures of sensitivity, specificity, accuracy, Matthews correlation coefficient (MCC) and positive predictive value (PPV). The mathematical equation for the aforementioned performance measures is given in Table 1.

Table 1. Mathematical equation of respective performance measures

Performance Measures	Equations
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$
PPV	$\frac{TP}{TP + FP}$

##### 4.1. Performance analysis of multi-model features for YouTube and SAVEE dataset

Here, Table 2 represent the performance analysis of the enhanced multi-verse optimizer in YouTube dataset with respect to various performance metrics. It shows that proposed method achieves a higher accuracy of 98.65%, a sensitivity of 99.59%, a specificity of 98.76%, MCC of 98.87%, and a PPV of 98.01%. Whereas the text, audio, video, video+text, audio+text and video+Audio, attain their corresponding accuracies at 92.90%, 94.00%, 96.86%, 95.98%, 89.76%, and 97.81%. Table 3 depicts that the hybrid features achieve a higher accuracy of 98.76%, a sensitivity of 98.04%, a specificity of 97.04%, MCC of 98.54% and a PPV of 98.88%. Whereas the text, audio, video, video+text, audio+text and video+audio accomplish their accuracies, respectively at 94.95%, 97.54%, 96.86%, 96.90%, 93.98%, and 97.24%.

Table 2. Performance analysis of the YouTube dataset

Multi-model features	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	MCC (%)
Text	92.90	91.86	92.86	90.11	88.98
Audio	94.00	89.86	91.86	87.07	90.71
Video	96.86	93.43	95.86	94.24	95.37
Video+Text	95.98	96.86	97.86	94.73	93.98
Audio+Text	89.76	94.99	88.86	89.99	90.91
Video+_Audio	97.81	95.69	98.41	96.93	94.36
Hybrid_features(Text,Audio,Video)	98.65	99.59	98.76	98.01	98.87

Table 3. Performance analysis of SAVEE dataset

Multi-model features	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	MCC (%)
Text	94.95	94.90	95.92	96.95	92.89
Audio	97.54	96.76	95.93	95.31	96.98
Video	96.86	97.81	95.54	94.76	96.88
Video+Text	96.90	95.60	93.97	93.94	94.80
Audio+Text	93.98	95.79	96.88	89.85	94.97
Video+Audio	97.24	96.86	95.98	98.33	95.86
Hybrid_features (Text, Audio, Video)	98.76	98.04	97.04	98.54	98.88

#### 4.2. Performance of feature selection methods for YouTube and SAVEE datasets

Here, Table 4 represents the performance analysis of feature selection methods namely particle swarm optimization (PSO), ant colony optimization (ACO), and artificial bee optimization (ABC), in YouTube dataset. Table 4 showed that the suggested method achieves a higher accuracy of 98.65%, the sensitivity of 99.59%, a specificity of 98.76%, a PPV of 98.01% and MCC of 98.87%. Whereas the PSO, ACO, and ABC achieved accuracy of 93.78%, 93.89%, and 90.64% respectively. Table 5 displays that the proposed method produces a higher accuracy of 98.76%, sensitivity of 98.04%, specificity of 97.04%, PPV of 98.54% and MCC of 98.87%. Whereas the PSO, ACO, and ABC achieved accuracy of 95.29%, 94.99%, and 90.64% respectively.

Table 4. Performance analysis of feature selection methods for You tube dataset

Feature Selection methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	MCC (%)
PSO	93.78	92.43	92.74	90.91	92.99
ACO	93.89	92.38	94.98	90.79	92.42
ABC	90.64	93.59	89.54	88.16	92.87
Proposed	98.65	99.59	98.76	98.01	98.87

Table 5. Performance analysis of feature selection methods for SAVEE dataset

Feature Selection methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	MCC (%)
PSO	95.29	94.76	92.86	90.20	93.22
ACO	94.99	96.91	93.98	90.97	93.63
ABC	96.63	95.87	94.92	97.44	94.51
Proposed	98.76	98.04	97.04	98.54	98.88

#### 4.3. Performance analysis of classification methods for YouTube and SAVEE dataset

Here, Table 6 represents the performance analysis of various classification methods such as k-nearest neighbor (KNN), differential evolution (DE) and random forest (RF), in YouTube dataset. After analyzing Table 6, the inference is that the CNN accomplishes a higher accuracy of 98.65%, sensitivity of 99.59%, specificity of 98.76%, PPV of 98.01% and MCC of 98.87%. Whereas the KNN, RF, DE and CNN achieve their respective accuracies at 91.76%, 91.38%, 95.74 and 98.65 %. Likewise, Table 7 portrays the same analysis for the SAVEE dataset. In this, the CNN achieves a higher accuracy of 98.76%, a sensitivity of 98.04%, a specificity of 97.04%, a PPV of 98.54% and MCC of 98.88%. Whereas the KNN, RF, and DE achieved respective accuracies at 92.90%, 90.90%, and 95.79%.

Table 6. Performance analysis of classification methods for YouTube dataset

Classifiers	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	MCC (%)
KNN	91.76	90.87	92.86	89.11	88.88
RF	91.38	89.57	92.97	88.07	90.52
DE	95.74	94.86	96.98	95.98	93.98
CNN	98.65	99.59	98.76	98.01	98.87

Table 7. Performance analysis of classification methods for the SAVEE dataset

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	MCC (%)
KNN	92.90	93.88	94.43	95.89	88.18
RF	90.90	89.36	91.97	89.84	90.62
DE	95.79	93.68	96.98	94.68	97.37
CNN	98.76	98.04	97.04	98.54	98.88

#### 4.4. Comparative analysis

This part, provides a comparative evaluation of the suggested enhanced multi-verse optimizer (EMVO) model with existing models, which are the shuffled frog leaping algorithm incremental wrapped-based subset selection (SFLA-IWSS) [21], hierarchical DNN [22] and unique preference learning [23] to evaluate the performance of the EMVO. The proposed EMVO is compared and analyzed with the existing models in terms of classification accuracy which is shown in Table 8. From the Table 8, the proposed EMVO model achieves a greater accuracy of 98.65% in YouTube and 98.76% in SAVEE dataset, when compared to SFLA-IWSS [21] Hierarchical DNN [22] and unique preference learning [23] models, that respectively attain



97.05%, 81.02% and 85.06% in the SAVEE dataset. This clearly states that, the proposed EMVO implements the classification with greater accuracy and outperforms the existing SFLA-IWSS [21] hierarchical DNN [22] and unique preference learning [23] models.

Table 8. Comparison evaluation of the proposed model with existing models

Models	Classification Accuracy (%)	Dataset
SFLA-IWSS [21]	97.81	YouTube
	97.05	SAVEE
Hierarchical DNN [22]	81.02	SAVEE
Unique preference learning [23]	85.06	SAVEE
EMVO	98.65	YouTube
	98.76	SAVEE

#### 4.5. Discussion

According to the results, the existing methods of shuffled frog leaping algorithm (SFLA) [21], hierarchical deep neural network (DNN) [22] and unique preference learning [23], are compared with the proposed method in terms of accuracy. The developed EMVO in this research study aims to further enhance the accuracy and effectiveness of recognizing emotions in real-time, contributing to advancements in the field of facial emotion recognition technology. It excels in handling complex and non-linear problems, making it suitable for the intricate task of recognizing emotions from facial expressions. EMVO's adaptability allows for efficient learning from limited labeled data, mitigating the need for extensive training datasets. Facial emotion recognition is a challenging task due to varied possibility of expressions, assessed under varying lighting conditions and orientations. Thus, the existing methods struggled in optimal feature selection, and effective navigation of high-dimensional spaces. Additionally, achieving perfect accuracy often requires a combination of sophisticated feature extraction, model design, and robust data handling. The limitations of the existing methods, coupled with dataset diversity and challenges in emotion recognition, prevents achieving better accuracy results, in the datasets of YouTube and SAVEE. In order to mitigate these existing issues, EMVO was utilized as a potential solution to scale up the exploration and exploitation efficiency, during feature selection and model training. EMVO, inspired by a more comprehensive search of the solution space, potentially helps in the discovery of more optimal features for emotion recognition. Its ability to balance exploration and exploitation better, produces a model design with improved feature selection which addresses the limitations of existing methods. According to the results, the existing SFLA [21] achieved accuracy of 97.81% in YouTube and 97.05% in SAVEE dataset. Then, hierarchical DNN [22] achieved an accuracy of 81.02% in SAVEE dataset. Finally, unique preference learning [23] achieved accuracy of 85.06% in SAVEE dataset respectively. When compared to existing methods, the suggested method produced better accuracy performances, that were measured at 98.65% in YouTube and 98.76% in SAVEE dataset.

#### 5. CONCLUSION

The research of video, audio and text that express user emotions, is an intriguing study that needs an effective emotion recognition model. Current approaches use various feature selection strategies to increase emotion identification efficiency. Here, an EMVO technique was suggested to improve the effectiveness of emotion recognition. In this research, two datasets were used to collect the data, namely YouTube and SAVEE datasets. Also, the classification was done using CNN to improve emotion recognition performance. When compared to the existing methods namely SFLA-IWSS, Hierarchical DNN, and Unique preference learning, the suggested method performed better in terms of accuracy, as it accomplished 98.65% in YouTube and 98.76% in SAVEE dataset. These results demonstrate superior performance in emotion identification when compared to existing methods. In the future, the focus will be on training various classifiers and temporal analysis, for much more accurate emotion recognition in audio, text, and video data.





#### AUTHOR CONTRIBUTIONS

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, have been done by 1<sup>st</sup>. The supervision and project administration have been done by 2<sup>nd</sup>, and 3<sup>rd</sup> author.





## REFERENCES

- [1] L. Lu, "Multi-angle face expression recognition based on generative adversarial networks," *Computational Intelligence*, vol. 38, no. 1, pp. 20–37, Feb. 2022, doi: 10.1111/coin.12437.
- [2] P. M. A. Kumar, L. A. Raj, K. M. Sagayam, and N. S. Ram, "Expression invariant face recognition based on multi-level feature fusion and transfer learning technique," *Multimedia Tools and Applications*, vol. 81, no. 26, pp. 37183–37201, Aug. 2022, doi: 10.1007/s11042-022-13538-z.
- [3] F. Nan *et al.*, "Feature super-resolution based facial expression recognition for multi-scale low-resolution images," *Knowledge-Based Systems*, vol. 236, p. 107678, Jan. 2022, doi: 10.1016/j.knosys.2021.107678.
- [4] J. Shen, H. Yang, J. Li, and Z. Cheng, "Assessing learning engagement based on facial expression recognition in MOOC's scenario," *Multimedia Systems*, vol. 28, no. 2, pp. 469–478, Oct. 2022, doi: 10.1007/s00530-021-00854-x.
- [5] Y. Qiu, Z. Niu, B. Song, T. Ma, A. Al-Dhelaan, and M. Al-Dhelaan, "A novel generative model for face privacy protection in video surveillance with utility maintenance," *Applied Sciences (Switzerland)*, vol. 12, no. 14, p. 6962, Jul. 2022, doi: 10.3390/app12146962.
- [6] J. Lin, L. Xiao, T. Wu, and W. Bian, "Image set-based face recognition using pose estimation with facial landmarks," *Multimedia Tools and Applications*, vol. 79, no. 27–28, pp. 19493–19507, Mar. 2020, doi: 10.1007/s11042-019-08408-0.
- [7] H. Cevikalp and G. G. Dordinejad, "Video based face recognition by using discriminatively learned convex models," *International Journal of Computer Vision*, vol. 128, no. 12, pp. 3000–3014, Jul. 2020, doi: 10.1007/s11263-020-01356-5.
- [8] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification," *Computational Intelligence*, vol. 36, no. 2, pp. 861–881, Jan. 2020, doi: 10.1111/coin.12274.
- [9] C. Li, Y. Huang, W. Huang, and F. Qin, "Learning features from covariance matrix of Gabor wavelet for face recognition under adverse conditions," *Pattern Recognition*, vol. 119, p. 108085, Nov. 2021, doi: 10.1016/j.patcog.2021.108085.
- [10] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," *SN Applied Sciences*, vol. 2, no. 3, Feb. 2020, doi: 10.1007/s42452-020-2234-1.
- [11] G. Pinto, J. M. Carvalho, F. Barros, S. C. Soares, A. J. Pinho, and S. Brás, "Multimodal emotion evaluation: A physiological model for cost-effective emotion classification," *Sensors (Switzerland)*, vol. 20, no. 12, pp. 1–13, Jun. 2020, doi: 10.3390/s20123510.
- [12] L. A. Steenhuis *et al.*, "The longitudinal association between preadolescent facial emotion identification and family factors, and psychotic experiences in adolescence (The TRAILS Study)," *Child Psychiatry and Human Development*, vol. 51, no. 2, pp. 187–199, Sep. 2020, doi: 10.1007/s10578-019-00922-4.
- [13] M. F. H. Siddiqui and A. Y. Javaid, "A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images," *Multimodal Technologies and Interaction*, vol. 4, no. 3, pp. 1–21, Aug. 2020, doi: 10.3390/mti4030046.
- [14] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human-computer interaction applications," *Neural Computing and Applications*, vol. 35, no. 32, pp. 23311–23328, Apr. 2021, doi: 10.1007/s00521-021-06012-8.
- [15] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep CNN," *Electronics (Switzerland)*, vol. 10, no. 9, p. 1036, Apr. 2021, doi: 10.3390/electronics10091036.
- [16] A. A. Abdulmunem, N. D. Al-Shakarchy, and M. S. Safoq, "Deep learning based masked face recognition in the era of the COVID-19 pandemic," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 2, pp. 1550–1559, Apr. 2023, doi: 10.11591/ijece.v13i2.pp1550-1559.
- [17] J. S. Hussain, A. Al-Khazzar, and M. N. Raheema, "Recognition of new gestures using Myo armband for myoelectric prosthetic applications," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 6, pp. 5694–5702, Dec. 2020, doi: 10.11591/ijece.v10i6.pp5694-5702.
- [18] V. Sekar and A. Jawaharlalnehru, "Semantic-based visual emotion recognition in videos-a transfer learning approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 4, pp. 3674–3683, Aug. 2022, doi: 10.11591/ijece.v12i4.pp3674-3683.
- [19] M. Moussa, M. Hamila, and A. Douik, "Face recognition using fractional coefficients and discrete cosine transform tool," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 1, pp. 892–899, Feb. 2021, doi: 10.11591/ijece.v11i1.pp892-899.
- [20] M. Wafi, F. A. Bachtar, and F. Utaminigrum, "Feature extraction comparison for facial expression recognition using adaptive extreme learning machine," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 1, pp. 1113–1122, Feb. 2023, doi: 10.11591/ijece.v13i1.pp1113-1122.
- [21] S. R. Kothuri and N. R. Rajalakshmi, "A hybrid feature selection model for emotion recognition using shuffled frog leaping algorithm (SFLA)-incremental wrapper-based subset feature selection (IWSS)," *Indian Journal of Computer Science and Engineering*, vol. 13, no. 2, pp. 354–364, Apr. 2022, doi: 10.21817/indjcs/2022/v13i2/221302040.
- [22] P. Singh, R. Srivastava, K. P. S. Rana, and V. Kumar, "A multimodal hierarchical approach to speech emotion recognition from audio and text [Formula presented]," *Knowledge-Based Systems*, vol. 229, Oct. 2021, doi: 10.1016/j.knosys.2021.107316.
- [23] Y. Lei and H. Cao, "Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels," *IEEE Transactions on Affective Computing*, pp. 1–16, 2023, doi: 10.1109/TAFFC.2023.3234777.
- [24] J. Singh, L. B. Saheer, and O. Faust, "Speech emotion recognition using attention model," *International Journal of Environmental Research and Public Health*, vol. 20, no. 6, Mar. 2023, doi: 10.3390/ijerph20065140.
- [25] K. Mao, Y. Wang, L. Ren, J. Zhang, J. Qiu, and G. Dai, "Multi-branch feature learning based speech emotion recognition using SCAR-NET," *Connection Science*, vol. 35, no. 1, Apr. 2023, doi: 10.1080/09540091.2023.2189217.
- [26] J. R. Kaka and K. Satya Prasad, "Differential evolution and multiclass support vector machine for Alzheimer's classification," *Security and Communication Networks*, vol. 2022, pp. 1–13, Jan. 2022, doi: 10.1155/2022/7275433.
- [27] M. B. Er, "A novel approach for classification of speech emotions based on deep and acoustic features," *IEEE Access*, vol. 8, pp. 221640–221653, 2020, doi: 10.1109/ACCESS.2020.3043201.
- [28] L. Abualigah, "Multi-verse optimizer algorithm: a comprehensive survey of its results, variants, and applications," *Neural Computing and Applications*, vol. 32, no. 16, pp. 12381–12401, Mar. 2020, doi: 10.1007/s00521-020-04839-1.
- [29] M. Jiang and S. Yin, "Facial expression recognition based on convolutional block attention module and multi-feature fusion," *International Journal of Computational Vision and Robotics*, vol. 13, no. 1, pp. 21–37, 2022, doi: 10.1504/ijcvr.2023.127298.
- [30] A. M. Ashir, A. Eleyan, and B. Akdemir, "Facial expression recognition with dynamic cascaded classifier," *Neural Computing and Applications*, vol. 32, no. 10, pp. 6295–6309, Mar. 2020, doi: 10.1007/s00521-019-04138-4.





**BIOGRAPHIES OF AUTHORS**

**Ravi Gummula**     is a PhD scholar of electronics and communication engineering at the Dr. M.G.R. Educational and Research Institute in Chennai, Tamil Nadu, India. He obtained his M. Tech in VLSI design in 2012 from Shadan College of Engineering and Technology, and his BE in Electronics and Communication Engineering from Deccan College of Engineering and Technology in 2007. He can be contacted at email: ravi.gummula@gmail.com.



**Vinothkumar Arumugam**     is a professor of electronics and communication engineering at the Dr. M.G.R. Educational and Research Institute in Chennai, Tamil Nadu, India. He obtained his Ph.D. in machine learning in 2017 and his M.Tech. in applied electronics in 2010 from Dr. M.G.R. Educational and Research Institute, and his BE in electronics and communication engineering from Anna University in 2008. He received an M.Sc. in real estate valuation from Annamalai University in 2016. He can be contacted at email: dravinoth@gmail.com.



**Abilasha Aranganathan**     is an assistant professor of electronics and communication engineering at the Dr. M.G.R. Educational and Research Institute in Chennai, Tamil Nadu, India. She obtained her M.Tech. in nanotechnology in 2014 and her BE in electronics and communication engineering in 2012 from Anna University. She can be contacted at email: vabilasha90@gmail.com.