Journal of Information Systems Engineering and Management
2025, 10(23s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Multi-modal Emotion Recognition using Speech and Facial Expressions using novel PSO-AOA-MAUT and Deep Convolution Neural Network

Shwetkranti Taware[1], Anuradha D. Thakare[2], Manisha D. Kitukale[3]

[1]Research Scholar, Department of Computer Engineering, Pimpri Chichwad College of Engineering, Savitribai Phule Pune University, Pune, India. shweta.taware@gmail.com

[2]Department of Computer Science & Engineering(AI&ML), Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University Pune, India. anuradha.thakare@pccoepune.org

[3]Professor, P. Wadhwani College of Pharmacy, Yavatmal, India, kitukalemanisha5@gmail.com

ARTICLE INFO

Received: 14 Dec 2024

Revised: 30 Jan 2025

Accepted: 18 Feb 2025

ABSTRACT

A vast growth in automation and robotics leads to effective human-machine interaction considering human emotions. Deep learning (DL) based Multimodal emotion recognition (MER) has shown higher reliability, accuracy, and security compared with unimodal emotion recognition systems (UER). However, the effectiveness of the MES is limited due to the more significant feature vector that increases the intricacy and total trainable parameters of the DL framework. This work provides an MER system using speech and facial expressions using a Deep Convolutional Neural Network (DCNN). It uses a novel hybrid combination of Particle Swarm Optimization, Archimedes Optimization Algorithm, and Multi-Attribute Utility Theory algorithm (PSO-AOA-MAUT) for the prominent feature selection. The AOA algorithm helps to attain better convergence and balanced optimization using the exploration and exploitation of particles. The multi-criteria decision-based MAUT algorithm is utilized to compute the weights of the fitness function of the PSO-AOA-based feature selection scheme. The results of the suggested MER are evaluated on the BAUM dataset. The suggested MER provides improved accuracy of 98.33%, recall rate of 0.98, precision of 0.97 and F1-score of 0.97 compared with traditional methods.

Keywords: Emotion recognition, deep convolution neural network, particle swarm optimization, archimedes optimization algorithm, multi-attribute utility theory.

INTRODUCTION

Emotion recognition is vital in many human-machine interaction systems where commands or control signals are provided using human emotion. It also helps to analyze the mental health of the human being (Abdullah et al., 2021) & (Koromilas & Giannakopoulos, 2021). Emotion recognition systems are widely used in many applications such as online learning platforms, affective agents, call centers, narcotics analysis, computer games, robotics and automation, audio and video conferencing, etc. The human-machine interaction needs interpretation, recognition, and simulation of the specific human emotion for effective communication between the human and computer agent (He et al., 2020) & (Bhangale & Kothandaraman, 2023).

The human emotions can be captured using various physiological and behavioral signals. The physiological modalities consist of the physical characteristics of the human being for the affect sensing such as face, body signal, electroencephalographs (EEG), Electrocardiograms (ECG), eye movement, gait, action, etc. The behavioral modalities for emotion recognition include speech, text, video, etc (Liu et al., 2021) & (Cimtay et al., 2020).

UER considers a single modality for the affect sensing of the human. However, MER systems consider multiple modalities for the affect sensing of the human. The UER are often subjected to poor recognition rates, less security, less robustness, less reliability, and less trust in the emotion recognition systems. At the same time, multi-modal emotion recognition systems are more accurate, reliable, secure, robust, and trustworthy than unimodal systems. The multi-modal system provides recognition results by fusing the modalities at the sensor, feature, decision, and score levels. The system's complexity depends upon the modality fusion level (Tan et al., 2021),(Bhangale & Kothandaraman, 2022),(Marechal et al., 2019),(Ahmed et al., 2023).

This article provides multi-modal emotion recognition systems using speech and face modalities of humans. The chief contributions of the article are summarized as follows:

•Emotion recognition using facial images using different texture and shape features such as local binary pattern (LBP) and Histogram of oriented gradients (HOG) features and DCNN

•Emotion recognition using speech signal using multiple audio features (MAF) that comprises spectral, temporal and voice-quality features and DCNN

•Feature selection using Hybrid Collaborative Particle Swarm Optimization, Archimedes Optimization Algorithm, and Multi-Attribute Utility Theory (PSO-AOA-MAUT)

•MER using face and facial images using speech and audio features using proposed PSO-AOA-MAUT and DCNN.

RELATED WORK

Various modalities have been considered for MER, such as face, gait, motion, speech, body signals, ECG, and EEG. Most of the time, the flexibility of combining features from two different modalities could be better. While selecting two or more modalities, it is crucial to consider correlation and connectivity in two modalities is essential. Face (Kumar A et al.,2023) and speech modalities are highly correlated with each other as it depicts the changes in facial expression for “what” and “how” a person is speaking. (Middya et al., 2022) presented model-level fusion for MER using video and audio data. It provided 86% and 99% accuracy for RAVDESS and SAVEE datasets. (Chen et al., 2022) proposed MER using speech and facial expression. It uses time domain and frequency domain features of speech modal-ity and gray pixels of facial images. Further, it utilized kernel canonical correlation analysis for multimodal feature fusion. It used k-mean clustering to select the crucial features. It is observed that k-mean clustering-based feature selection helps to achieve 4.7% and 2.77% improvement over speech recognition without feature selection. (Sharafi et al., 2022) explored spatio-temporal CNN for MER. It used a hybrid model combining two CNN and BiLSTM for learning fea-tures. The CNNs are used for extracting spatial and temporal attributes of the vid-eo frames. The MFCC and energy features are obtained for audio signal using the BiLSTM network. It provided 99.23%, 94.99%, and 99.75% accuracy for RML, RAVDESS, and SAVEE datasets, respectively. (Chen et al., 2022) suggested MER in conversation. It uses HU-Dialogue to characterize the conversation's emotion.

(Jia et al., 2022) offered MER using video, audio, and motion capture. It utilizes a dual spectrogram to combine the global and local representation of speech us-ing CNN, gated recurrent unit (GRU), and attention unit. Further, 3D-CNN with an attention mechanism is used to acquire emotional features from videos. After-ward, 3-layered BiLSTM is employed for capturing the movement of the head and hand. It is noted that decision-level fusion leads to better generalization abil-ity and higher precision. (Kumar et al., 2023) explored a novel interpretability technique for identifying important features of images and speech. It resulted in 83.29% accuracy for the IIT-R SIER dataset.

It is noted that MER provides a noteworthy boost in accuracy compared with UER. It has shown more robustness, generalization ability, security, and reliability. However, a combination of multiple modalities increases the complexity of sys-tems because of more prominent features and complexity in deep learning frameworks. Therefore, it is essential to select the critical features to improve the feature distinctiveness to boost the accuracy of the MER system.

MATERIAL AND METHODOLOGY

This section provides the details about the material and methodology utilized for the implementation foe the proposed MER system.

Dataset

The usefulness of the suggested system are evaluated on the BAUM-2 dataset, which consists of audio-visual emotional expressions (Erdem et al., 2018) & (Eroglu Erdem et al., 2015). We have selected the facial images from the peak frames of the BAUM dataset videos. It consists of six emotions: happiness, anger, disgust, sadness, fear, and surprise. It encompasses 1047 video samples of 247 subjects. The facial images from the BAUM-2 dataset are shown in **Figure 1**. The audio samples from the dataset has variable length, therefore to maintain the uniformity in the feature we have cropped or appended the audio to the 4 sec duration signal. The dataset is split in the ratio of 70:15:15 for the training, testing and validation.

Figure 1. samples facial images from BAUM-2 dataset [18]

Methodology

The flow diagram of the proposed MER scheme is illustrated in **Figure 2.** which considers speech and image as the input signal. The MAF consists of various spectral, temporal and voice quality features to represent the emotional attributes of the speech signal. However, HOG and LBP features describes the shape changes and texture changes over the facial image due to emotions. The features are concatenated to form the combined feature vector. The PSO-AOA-MAUT algorithm is employed for selecting the imperative features from the speech and image features. The AOA algorithm offers improvement in convergence and balance in exploitation and exploration of PSO algorithm. The MAUT algorithms selects the weight factors of the fitness function of the suggested PSO-AOA based feature selection technique.

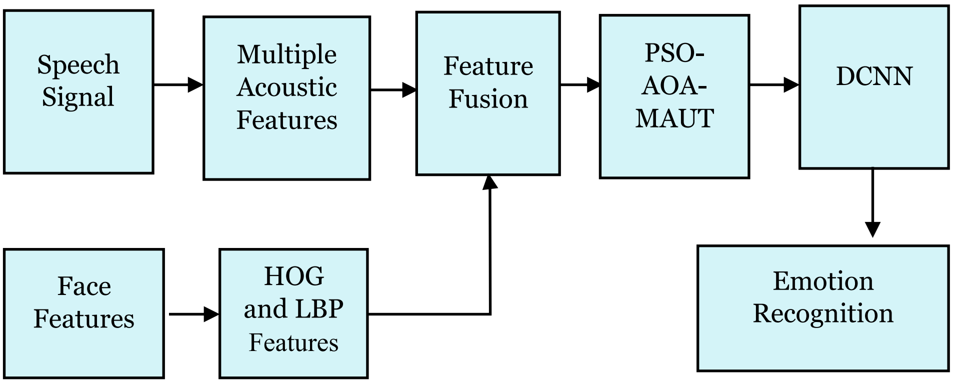


Figure 2. Overall flow diagram of the system

Feature Representation Of Speech And Face

Speech Feature Extraction. Multiple audio features comprising spectral, time-domain, and voice quality features are extracted to represent the speech signal. The distribution of the features is given in **Table 1**. Traditional Mel Frequency Cepstrum Coefficient (MFCC) algorithm of speech representation suffers from low-frequency resolution problems and spectral leakage problems; to avoid this, the proposed work considers multi-taper MFCC (MTMFCC) features. The MTMFCC provides better resolution at low frequencies and minimizes the spectral leakage problem (Bhangale & Kothandaraman, 2023). The time-domain features consist of pitch, zero crossing rate (ZCR), and

Hjorth's parameters, such as mobility, activity and complexity. The voice quality feature includes jitter and shimmer that depict the changes in time and amplitude, respectively.

Table 1: Multiple audio features description

Type of Features	Features	Number of Features
Spectral Features	MTMFCC	13
	MTMFCC-Δ	13
	MTMFCC-ΔΔ	13
	LPCC	12
	Formant	1
	Mean of Formant	1
	Spectral kurtosis	257
	Skewness	1
	WPT features	56
Time domain features	Pitch frequency	1
	ZCR	1
	Hjorth Parameters	3
Voice Quality features	Jitter	1
	Shimmer	1
Total Speech Features		374

Facial Image Feature Extraction. The face features are represented using LBP based texture descriptor and HOG based shape descriptor. The LBP and HOG captures the textural and shape changes occurs in the distinct facial points such as nose, eye, mouth, forehead and chicks due to emotions. It represents the changes in the lips movement due to utterance of the specific emotions. The fea-tures are computed for the peak frame of the BAUM dataset. The LBP texture descriptor provides the feature vector length of 256. The HOG features are calcu-lated for the cell size of 8×8 pixel, block size of 2×2 cells, 50% overlapping of the block and 9 bin histogram of the orientation. For facial image of 128×128 resolution, the HOG descriptor provides feature vector of 8100. The HOG and LBP based facial expression description is illustrated in **Figure 3.** The combined feature vector forms final facial image feature descriptor of 8356. The LBP and HOG captures both local and global features of the images.

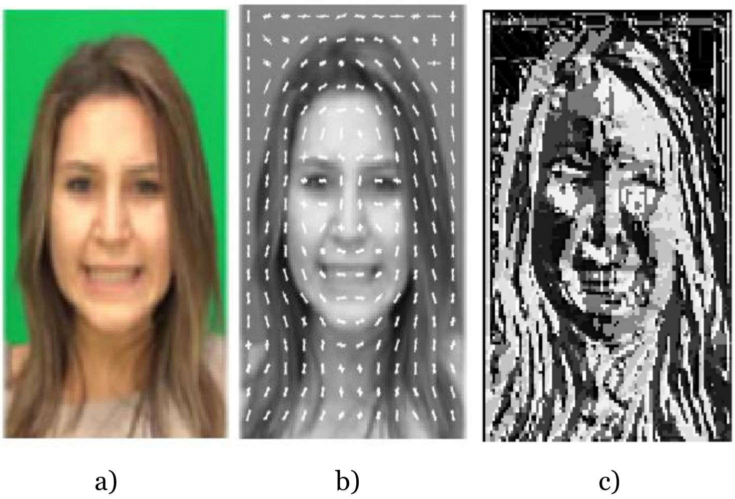


Figure 3. Facial expression description a) Original Image b) HOG features c) LBP descriptor

Feature Selection Using Pso-Aoa-Maut

The higher feature vector length in multi-modal emotion recognition increases the computational burden on the DL framework. Additionally, it leads to the larger training and recognition time of the system. This work presents a novel feature selection scheme based on the PSO-AOA-MAUT algorithm. The traditional PSO suffers from the lack of global convergence, poor balance in exploration and exploitation, limited multi-objective optimization, and pre-mature convergence (Marini & Walczak, 2015) & (Poli et al., 2007). This leads to improper selection of the potential solution. In this work, PSO algorithm is used for the feature selection. The AOA algorithm is utilized to surge the exploration and exploitation of the particles of PSO. The weight of the fitness function are selected using MAUT algorithm, where the weights are decided automatically based on the utility ranking of three attributes, representing the covariance, entropy, and ratio of inter-class to intra-class variance of the features. The flow diagram of the PSO-AOA-MAUT based feature selection scheme is described in **Figure 4.**

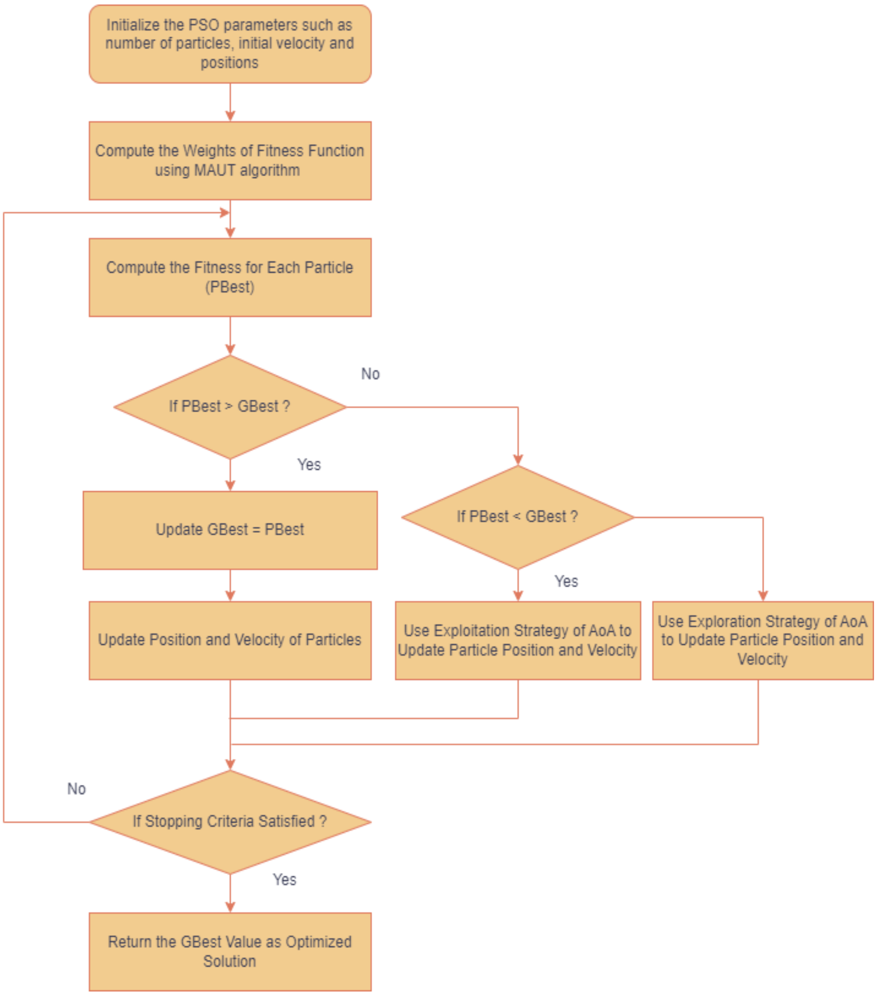

Initialize the PSO parameters such as number of particles, initial velocity and positions
Compute the Weights of Fitness Function using MAUT algorithm
Compute the Fitness for Each Particle (PBest)
If PBest > GBest ?
No
Yes
Update GBest = PBest
If PBest < GBest ?
Yes
Update Position and Velocity of Particles
Use Exploitation Strategy of AoA to Update Particle Position and Velocity
Use Exploration Strategy of AoA to Update Particle Position and Velocity
No
If Stopping Criteria Satisfied ?
Yes
Return the GBest Value as Optimized Solution

Figure 4. PSO-AOA-MAUT-based feature selection

PSO-AOA

PSO is a bio-inspired population-based stochastic algorithm that is easy, quicker, and can be utilized to non-differentiable problems. Here, PSO selects the crucial features from the available feature set. Here, the PSO population represents the possible feature set. The initial population is chosen arbitrarily by considering the lower and upper bounds as total features and minimum features(Marini & Walczak, 2015) & (Poli et al., 2007). The initial population (P) and velocity (V) of particles in PSO are given using equation 1 and 2.

$$P = [p_1, p_2, p_3, p_4, \dots \dots . . p_{n]} \qquad (1)$$

$$V = [v_1, v_2, v_3, v_4, \dots \dots . . v_{n\ ]} \qquad (2)$$

The fitness function is computed for every particle in the population, which is based on covariance (Cov), entropy (EN), and the ratio of inter-class to intra-class variance (RI) of the features as given in equation 3.

$$Fitness(p) = w_1 * Cov + w_2 * EN + w_3 * RI \qquad (3)$$

Here, the weights of the attributes are decided using the MAUT algorithm such that $w_1 + w_2 + w_3 = 1$.

If the PBest > GBest then the velocity (v) and position (p) of the particles (P) in PSO is modified using conventional PSO using equations 4 and 5.

$$v = w * v + c_1 * rand1 * (PBest - p) + c_2 * rand2 * (GBest - p) \quad (4)$$

$$p = p + v \quad (5)$$

Where w is inertia, c_1 and c_2 are intellectual and societal acceleration, $rand1$and $rand2$ are random weights between 0 and 2.

If the PBest < GBest, then the v and p of the particles in PSO is altered using the exploitation strategy of the AOA algorithm (Hashim et al., 2021) & (Dhal et al., 2023)

If the PBest = GBest, then the v and p of the particles in PSO is altered using the exploration strategy of the AOA algorithm.

MAUT

The Multi-attribute utility theory (MAUT) algorithm is based on every attribute's aggregate marginal utility score (MUS). The flow diagram of the MAUT-based attribute weight computation is described in **Figure 5** (Von Winterfeldt & Fischer, 1975) & (Jansen, 2011). The procedure of the MAUT-based attribute ranking is described as follows:

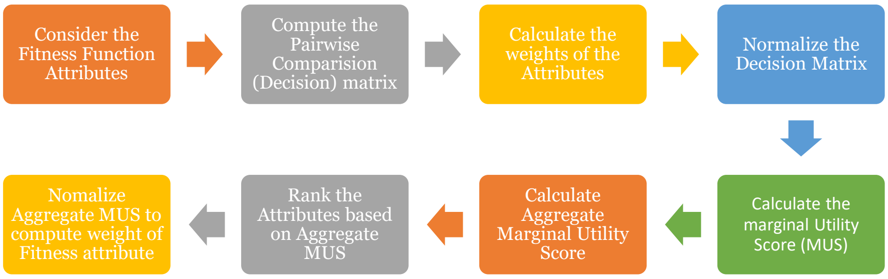

Consider the Fitness Function Attributes
Compute the Pairwise Comparison (Decision) matrix
Calculate the weights of the Attributes
Normalize the Decision Matrix
Calculate the marginal Utility Score (MUS)
Calculate Aggregate Marginal Utility Score
Rank the Attributes based on Aggregate MUS
Nomalize Aggregate MUS to compute weight of Fitness attribute

Figure 5. Process of the fitness function weight computation of PSO-AOA-based feature selection

Step 1: Form the decision matrix using a pairwise comparison of each criterion (A_{ij}) and compute the weight of every criterion (W_j) for three attributes such as covariance, entropy and the ratio of inter to the intra-class variance of the features.

Step 2: Normalize the decision matrix using the following equation.

$$A_{ij} = \frac{A_{ij} - \min_i A_{ij}}{\min_i A_{ij} - \max_i A_{ij}} \quad (6)$$

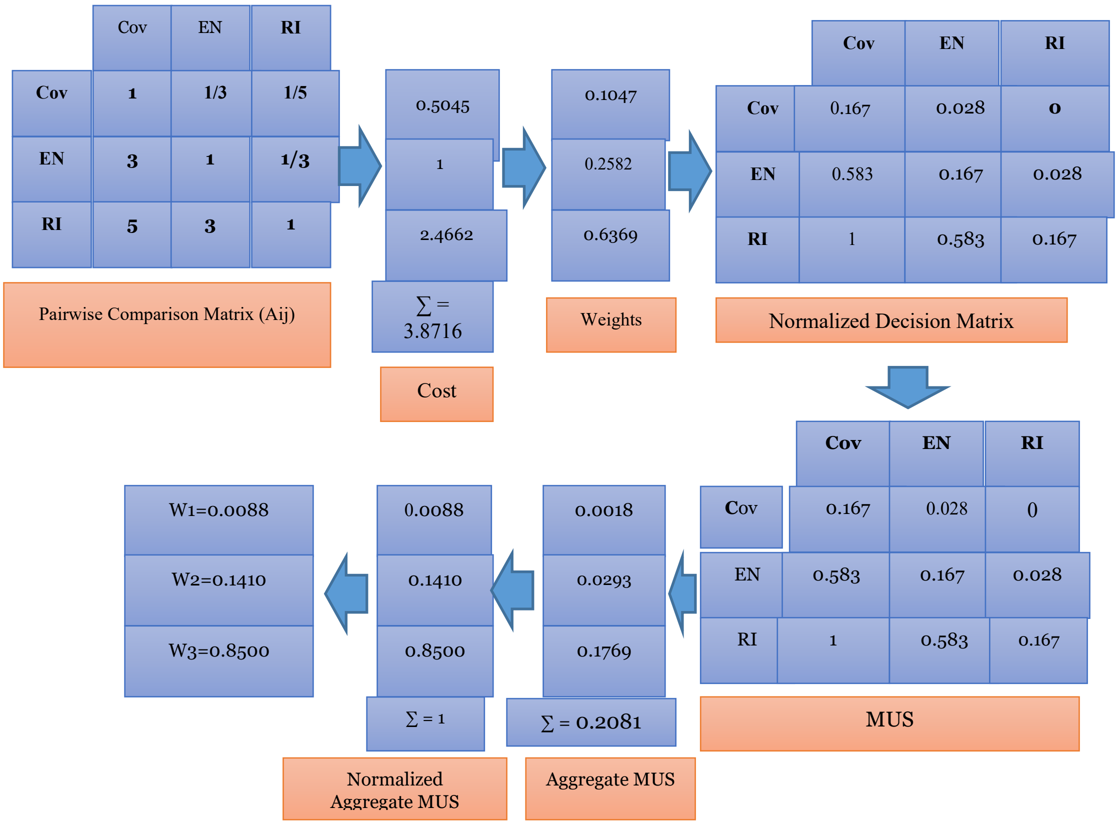
Step 3: Compute the MUS using the following equation.

$$MUS_{ij} = \frac{e^{(A_{ij})^2} - 1}{1.71} \quad (7)$$

Step 4: Calculate the aggregate MUS_i using the weight of j^{th} criterion

$$MUS_i = \sum_j MUS_{ij} \times W_j \quad (8)$$

Step 5: Rank the attributes as per the descending order of aggregate MUS. The attribute with the highest MUS is ranked higher. Fig. 6 illustrates the computation of the weights w1, w2 and w3 using MAUT algorithm for optimizing the fitness function of the PSO-AoA based feature selection.



	Cov	EN	RI
Cov	1	1/3	1/5
EN	3	1	1/3
RI	5	3	1

Pairwise Comparison Matrix (Aij)

Cost
0.5045
1
2.4662
∑ = 3.8716

Weights
0.1047
0.2582
0.6369

	Cov	EN	RI
Cov	0.167	0.028	0
EN	0.583	0.167	0.028
RI	1	0.583	0.167

Normalized Decision Matrix

	Cov	EN	RI
Cov	0.167	0.028	0
EN	0.583	0.167	0.028
RI	1	0.583	0.167

MUS

Aggregate MUS
0.0018
0.0293
0.1769
∑ = 0.2081

Normalized Aggregate MUS
0.0088
0.1410
0.8500
∑ = 1

W1=0.0088
W2=0.1410
W3=0.8500

Figure 6. Weight calculation using MAUT algorithm

Architecture Of 1-D Dcnn

The recommended 1-D DCNN (Jang Bahadur Saini et al., 2024) comprised of a total eleven layers. These levels are as follows: three convolution layers (Conv), three Rectified Linear Unit layers (ReLU), three batch normalization layer (BN), one fully connected layers (FC), and a Softmax classifier layer. The compact 1-D deep convolutional neural network that has been proposed takes the *Feat* features selected using MAUT-PSO-AOA algorithm as input. At each successive layer of the convolutional filtering process, the one-dimensional input undergoes a convolving operation with the Conv filter. It offers high-level information on the features of the spoken emotion signal. In equation 9, you may get the result of the convolution, which is denoted by the value z(n), of the features Feat(n) and the filter w(n) with a size of l. Equation 9 is used to represent the convolution feature map. In this equation, z_i^l is used to describe the i^{th} feature of the lth layer, z_j^{l-1} is used to represent the jth feature of the $(l-1)^{th}$layer, w_{ij}^l is used to explain the filter kernel of the l^{th} layer connected to the jth feature, b_i^l is used to display the ReLU activation function, and b_i^l stands for bias. The ReLU layer is an activation function that is uncomplicated and speedy, and it provides a solution to the problem of a vanishing gradient, which is outlined in equation (10).

$$z(n) = Feat(n) \times w(n) = \sum_{m=0}^{i-1} Feat(m). w(n-m) \qquad (9)$$

$$z_i^l = \sigma\left(b_i^l + \sum_j z_j^{l-1} \times w_{ij}^l\right) \qquad (10)$$

In the FC layer, the input feature vector is subjected to a linear transformation that is performed utilizing the weight matrix. The non-linear activation function (NLAF) is used throughout the whole of the non-linear transformation process, as seen by equation (11).

$$y_{jk}(x) = f(\sum_{i=1}^{n_H} W_{jk}x_i + wj_0) \tag{11}$$

Where x_i is a value from the flattened vector, w_0 is the bias term, w denotes the weight matrix, f represents the NLAF, y is the output of the non-linear transformation, and n_H does not provide any hidden layers.

Lastly, the Softmax classifier delivers the likelihood of the output, where the label with the highest probability reveals the output class label, as demonstrated by the formulas (12-14). This is the case because the biggest probability of class label discloses the output class label [4].

$$z_i = \sum_j h_j w_{ji} \tag{12}$$

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{n} \exp(z_j)} \tag{13}$$

$$\hat{y} = \arg \max_i p_i \tag{14}$$

z_i denotes the input to the softmax layer, the probability of the class label is denoted by p_i, and the predicted class label is denoted by $\hat{y}$. In this case, h_j stands for the weight of the penultimate layer, while w_{ji} and h_j signifies the weights of the Softmax and penultimate layers, respectively. The feature vector's length has an effect on the entire trainable parameters and computational time. To train the network that is being recommended, the procedure known as stochastic gradient descent with momentum (SGDM) is used. The system is trained for batch sizes of up to 64 to get around the memory restriction that is being faced. In the training approach, 200 epochs, an initial learning rate of 0.01, a cross-entropy loss function, and a momentum value of 0.9 are all taken into consideration. The training accuracy and loss of the DCNN are shown in **Figure 7 and 8**, correspondingly.

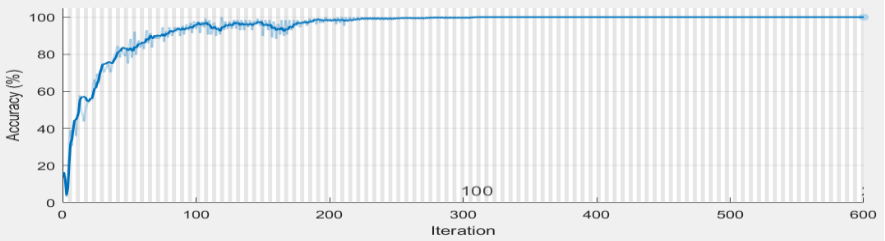


Figure 7. DCNN Training accuracy

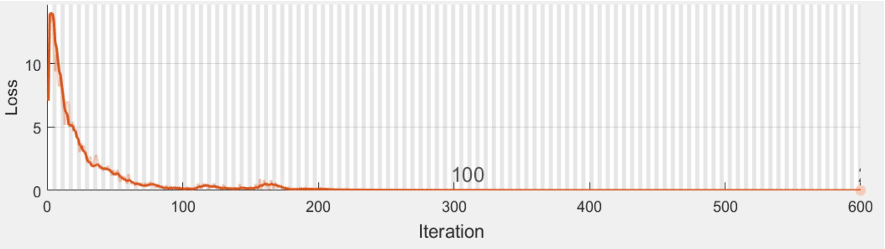


Figure 8. DCNN Training loss

EXPERIMENTAL RESULTS AND DISCUSSIONS

The suggested MER system uses MATLAB R2020b software on a Personal Computer with 20GB RAM and a Windows operation environment.

Performance Metrics

The outcomes are estimated using accuracy, recall, precision, F1-score, trainable parameters and training time as described in equation 15 to 18 respectively.

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TN}{TN + FN} \tag{16}$$

$$Accuracy(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{17}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{18}$$

Results And Discussions

The outcomes of the suggested MER are estimated for the face, speech and face-speech modality for emotion recognition of the six emotion. The face, speech and face-speech provides total features of 8356, 374 and 8730 respectively. The optimal feature set is selected using PSO-AoA-MAUT algorithm. **Figure 9-12** shows precision, recall, f1-score and accuracy for the PSO-AoA-MAUT-DCNN and PSO-AoA-AHP-DCNN based multiomodal emotion recognition using combination of speech and face features. The suggested scheme provides better results for the 600 features selected using PSO-AoA-MAUT algorithm and DCNN based emotion recognition system. The PSO-AoA-MAUT based MER system outperforms PSO-AoA-AHP based feature selection scheme for MER system. The PSO-AoA-AHP-DCNN provides precision of 0.9736, recall of 0.98, F1-score of 0.9767 and accuracy of 97.67% for speech+face based MER. Whereas, PSO-AoA-MAUT-DCNN provides precision of 0.9801, recall of 0.9837, F1-score of 0.9834 and accuracy of 98.33% for speech+face based MER.

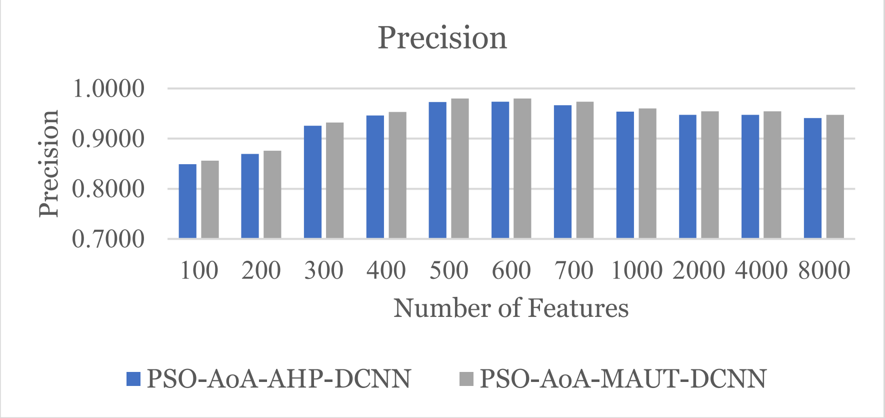


Figure 9. Precision of proposed MER system

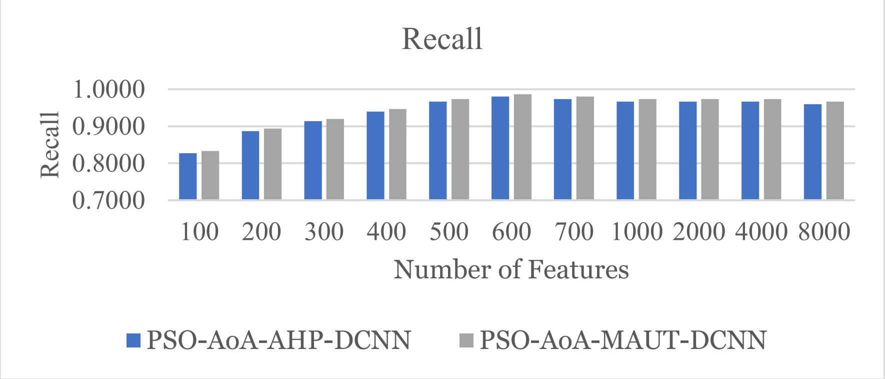


Figure 10. Recall of proposed MER system

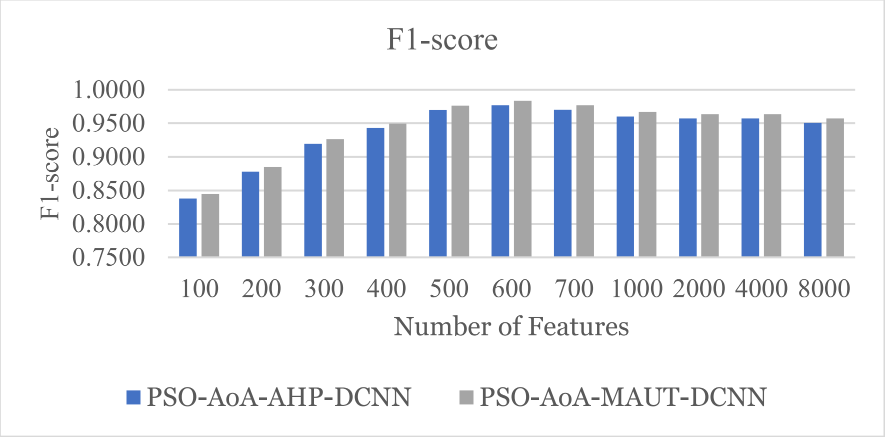


Figure 11. F1-score of proposed MER system

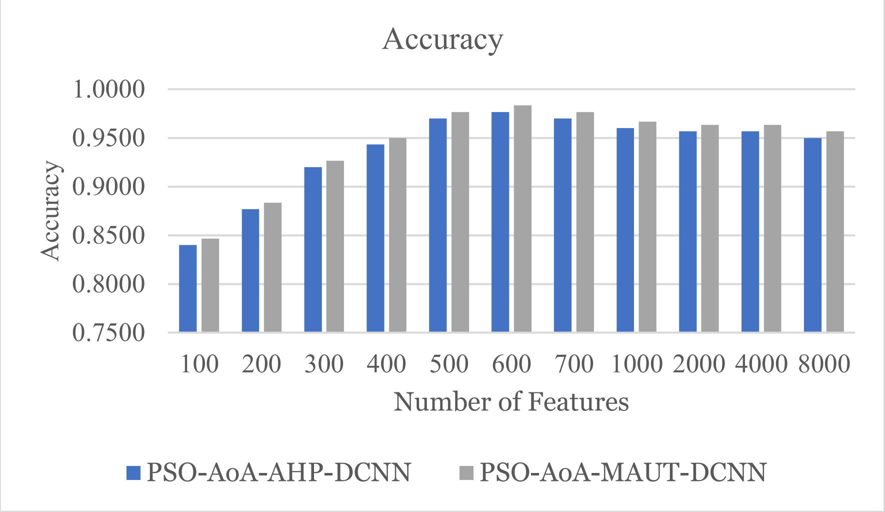
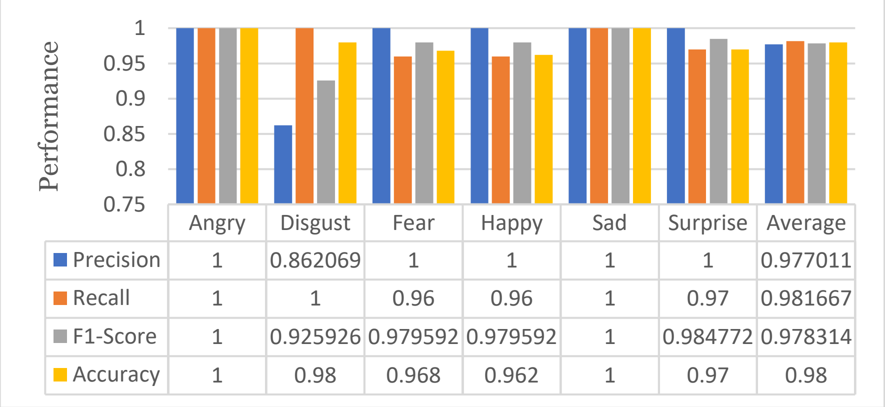


Figure 12. Accuracy of proposed MER system

Figure 13 provides the performance of PSO-AoA-MAUT-DCNN based MER scheme for six emotions from BAUM dataset. It is observed that the anger and sad emotion has higher arousal and hence results in 100% accuracy. The disgust, happy and surprise has lower arousal and hence results in an accuracy of 98%, 96.2% and 97% respectively.



	Angry	Disgust	Fear	Happy	Sad	Surprise	Average
🟦 Precision	1	0.862069	1	1	1	1	0.977011
🟧 Recall	1	1	0.96	0.96	1	0.97	0.981667
⬜ F1-Score	1	0.925926	0.979592	0.979592	1	0.984772	0.978314
🟨 Accuracy	1	0.98	0.968	0.962	1	0.97	0.98

Figure 13. Performance of PSO-AoA-MAUT-DCNN for distinct emotions

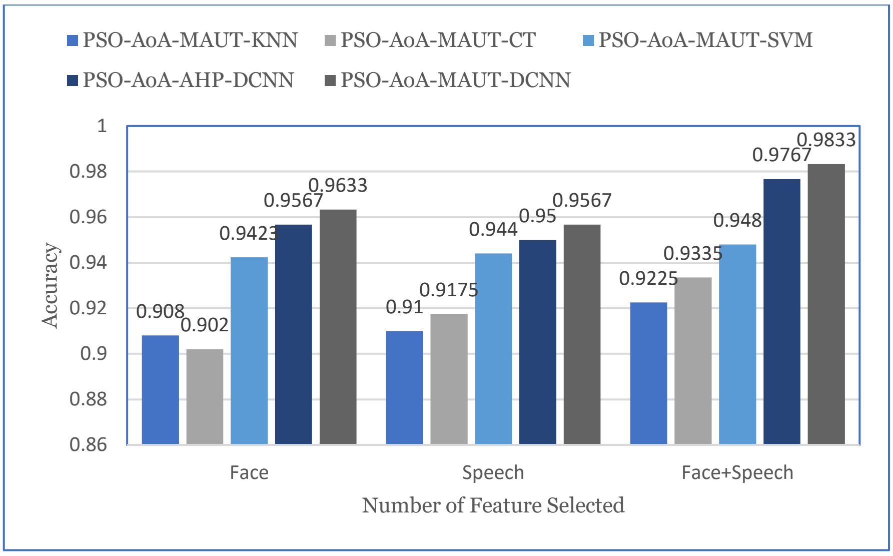


Figure 14. Performance comparison for different modality

The results of the suggested system is estimated for the face, speech, and combination of face and speech as shown in **Figure 14**. MER improves emotion recognition accuracy over traditional ML-based classifiers such as KNN, SVM, and CT. The MAUT-based weight decision provides superior results than the based weight selection for the feature selection. The proposed PSO-AoA- MAUT-DCNN provides the highest 98.33% accuracy for the speech+face modality, which is better than the face (96.33%) and speech (95.67%) modality. The PSO-AoA- MAUT-KNN (k=3) provides an accuracy of 90.8%, 91%, and 92.25% for the face, speech, and face+speech, respectively. The PSO-AoA-MAUT-CT resulted in emotion recognition accuracy of 90.20%, 91.75%, and 93.35% for face, speech, and face+speech modality. The PSO-AoA- MAUT-SVM gives an accuracy of 94.23%, 94.40%, and 94.80% for face, speech, and face+speech modalities, respectively. The feature selection helps to select the prominent features from the multimodal features, and the based multi-criteria decision algorithm assists in assigning proper weights to the feature selection fitness function. The DCNN helps to enhance the feature representation capability of the raw feature set and thus boosts the recognition results. PSO-AoA—AHP-DCNN provides an accuracy of 95.67% for face modality, 95% for speech modality, and 97.67% for face+speech modality. The utility theory-based MAUT algorithm provides superior results than the AHP and provides an accuracy of 96.33% for face, 95.67% for speech and 98.33% for speech+face modality.

CONCLUSIONS AND FUTURE SCOPES

Thus, this article offers a MER system using speech and face modalities based on an effective feature selection scheme and DCNN. It uses the PSO algorithm for feature selection where convergence of PSO is optimized using the AoA algorithm, and weights of fitness function of the PSO-based feature selection are decided using the MAUT algorithm. The DCNN-based lightweight architecture helps to boost the feature correlation, representation, and connectivity in different modalities. The proposed PSO-AoA- MAUT-DCNN-based algorithm provides an accuracy of 96.33% for face, 95.67% for speech, and 98.33% for speech+face modality. It is observed that the effective feature extraction scheme helps to achieve the salient features that assist in minimizing the computational intricacy of the DCNN architecture. In the future, the performance of the anticipated MER scheme can be enhanced by optimizing the DCNN algorithm. Further, focus can be given to developing a deep learning scheme that learns the spatial, spectral, and temporal attributes of the diverse emotion modalities.

REFERENCES

[1] Abdullah, S. M. S., Ameen, S. Y., Ameen, M. A. M. S., & Zeebaree, S. (2021). Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends, 2*(2), 52-58.

[2] Ahmed, N., Al Aghbari, Z., & Girija, S**.** (2023). A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications, 17*, 200171.

[3] Bhangale, K., & Kothandaraman, M. (2023). Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. *Electronics, 12*(4), 839.

[4] Bhangale, K. B., & Kothandaraman, M. (2022). Survey of deep learning paradigms for speech processing. *Wireless Personal Communications, 125*(2), 1913-1949.

[5] Bhangale, K. B., & Kothandaraman, M. (2023). Speech emotion recognition using the novel PEmoNet (Parallel Emotion Network). *Applied Acoustics, 212*, 109613.

[6] Chen, F., Shao, J., Zhu, A., Ouyang, D., Liu, X., & Shen, H. T. (2022). Modeling hierarchical uncertainty for multimodal emotion recognition in conversation. *IEEE Transactions on Cybernetics*.

[7] Chen, L., Wang, K., Li, M., Wu, M., Pedrycz, W., & Hirota, K. (2022). K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human–robot interaction. *IEEE Transactions on Industrial Electronics, 70*(1), 1016-1024.

[8] Cimtay, Y., Ekmekcioglu, E., & Caglar-Ozhan, S. (2020). Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access, 8*, 168865-168878.

[9] Dhal, K. G., Ray, S., Rai, R., & Das, A. (2023). Archimedes optimizer: Theory, analysis, improvements, and applications. *Archives of Computational Methods in Engineering, 30*(4), 2543-2578.

[10] Erdem, C., Turan, C., & Aydın, Z. (2018). BAUM-2. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5HC8C

[11] Eroglu Erdem, C., Turan, C., & Aydın, Z. (2015). BAUM-2: A multilingual audio-visual affective face database. *Multimedia Tools and Applications, 74*, 7429–7459. https://doi.org/10.1007/s11042-014-1986-2

[12] Hashim, F. A., Hussain, K., Houssein, E. H., Mabrouk, M. S., & Al-Atabany, W. (2021). Archimedes optimization algorithm: A new metaheuristic algorithm for solving optimization problems. *Applied Intelligence, 51*, 1531-1551.

[13] He, Z., Li, Z., Yang, F., Wang, L., Li, J., Zhou, C., & Pan, J. (2020). Advances in multimodal emotion recognition based on brain–computer interfaces. *Brain Sciences, 10*(10), 687.

[14] Jang Bahadur Saini, D., Choubey, S., Choubey, A., Kidwai, M., Mehrotra, M., Kolekar, S., & Raut, Y. (2024). Early detection of glaucoma integrated with deep learning models over medical devices. BioSystems, 238, 105156. https://doi.org/10.1016/j.biosystems.2024.105156

[15] Jansen, S. J. T. (2011). The multi-attribute utility method. In *The Measurement and Analysis of Housing Preference and Choice* (pp. 101-125). Dordrecht: Springer Netherlands.

[16] Jia, N., Zheng, C., & Sun, W. (2022). A multimodal emotion recognition model integrating speech, video, and MoCAP. *Multimedia Tools and Applications, 81*(22), 32265-32286.

[17] Kumar, A., Joshi, P., Bala, A., Sudhakar Patil, P., Jang Bahadur Saini, D. K., & Joshi, K. (2023). Smart transaction through an ATM machine using face recognition. Indian Journal of Information Sources and Services, 13(2), 7-13.

[18] Kumar, A., Yadav, R. K., & Jang Bahadur Saini, D. K. (2023). Create and implement a new method for robust video face recognition using convolutional neural network algorithm. e-Prime - Advances in Electrical Engineering, Electronics and Energy, 5, 100241. https://doi.org/10.1016/j.prime.2023.100241

[19] Koromilas, P., & Giannakopoulos, T. (2021). Deep multimodal emotion recognition on human speech: A review. *Applied Sciences, 11*(17), 7962.

[20] Kumar, P., Malik, S., & Raman, B. (2023). Interpretable multimodal emotion recognition using hybrid fusion of speech and image data. *Multimedia Tools and Applications*.

[21] Liu, W., Qiu, J.-L., Zheng, W.-L., & Lu, B.-L. (2021). Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems, 14*(2), 715-729.

[22] Marini, F., & Walczak, B. (2015). Particle swarm optimization (PSO): A tutorial. *Chemometrics and Intelligent Laboratory Systems, 149*, 153-165.

[23] Marechal, C., Mikolajewski, D., Tyburek, K., Prokopowicz, P., Bougueroua, L., Ancourt, C., & Wegrzyn-Wolska, K. (2019). Survey on AI-based multimodal methods for emotion detection. *High-Performance Modelling and Simulation for Big Data Applications, 11400*, 307-324.

[24] Middya, A. I., Nag, B., & Roy, S**.** (2022). Deep learning-based multimodal emotion recognition using model-level fusion of audiovisual modalities. *Knowledge-Based Systems, 244*, 108580.

[25] Poli, R., Kennedy, J., & Blackwell, T. (2007). Particle swarm optimization: An overview. *Swarm Intelligence, 1*, 33-57.

[26] Sharafi, M., Yazdchi, M., Rasti, R., & Nasimi, F. (2022). A novel spatio-temporal convolutional neural framework for multimodal emotion recognition. *Biomedical Signal Processing and Control, 78*, 103970.

[27] Tan, Y., Sun, Z., Duan, F., Solé-Casals, J., & Caiafa, C. F. (2021). A multimodal emotion recognition method based on facial expressions and electroencephalography. *Biomedical Signal Processing and Control, 70*, 103029.

[28] Von Winterfeldt, D., & Fischer, G. W. (1975). Multi-attribute utility theory: Models and assessment procedures. In *Utility, Probability, and Human Decision Making: Selected Proceedings of an Interdisciplinary Research Conference, Rome, 3–6 September, 1973* (pp. 47-85). Dordrecht: Springer Netherlands.