Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Research Article

Hybrid Model of Attention-Guided Deep Learning and Metaheuristic Optimization for Speech Emotion Recognition

Kenan DONUK[1*]

[1] Şırnak University, Department of Computer Engineering, kenandonuk@sirnak.edu.tr, Orcid No: 0000-0002-7421-5587

ARTICLE INFO

Article history:

Received 20 February 2026
Received in revised form 7 March 2026
Accepted 26 March 2026
Available online 27 July 2026

Keywords:

Speech emotion recognition, Attention mechanism, Metaheuristic optimization

Doi: 10.24012/dumf.1893980

* Corresponding author

ABSTRACT

Speech Emotion Recognition (SER) technologies, which have become a significant part of the human-computer interaction field, are artificial intelligence-based systems that automatically detect emotional states in human speech signals. The working principle of these systems is the process of matching acoustic features extracted from voice signals with emotional labels through machine learning. Despite significant advancements, current SER technologies still face some challenges. Chief among these is the limitation of model performance due to the inadequate representation of local and global patterns and the presence of non-distinctive features. This study presents an innovative hybrid model to overcome these challenges in emotion representation. The model consists of two phases. The first phase represents the baseline model, which involves an end-to-end learning process with the extraction of feature representations of emotions and their matching with emotion labels. The second phase is the process where feature optimization takes place. The first phase consists of steps including local feature extraction, channel-level feature activation, temporal dependency representation, and finally, focusing on key time periods. The second phase is the process of optimizing features using eighteen different optimization configurations derived from the integration of six metaheuristic algorithms with three machine learning paradigms. Experiments were conducted on the EMO-DB dataset. In a 5-fold cross-validation process, the proposed hybrid model increased the unweighted accuracy (UA) by 7.8% to 83.73% compared to the baseline model, and reduced feature dimensionality by 54.8%. These results confirm that the proposed hybrid model provides high accuracy and efficient feature representations.

Introduction

Speech Emotion Recognition (SER) involves the automatic detection and classification of emotional states from voice signals. It plays a crucial role in improving human-computer interaction and enhancing decision support processes in many different application areas. As artificial intelligence research has advanced, SER technologies have become an integral part of various fields. In healthcare, it helps in the early diagnosis of mental disorders such as depression [1,2] and anxiety [3], while in educational applications it enables the assessment of students' emotional involvement [4]. SER also improves customer service through adaptive call systems [5,6], supports safety by identifying stress indicators [7], and enhances driver safety in automotive technologies [8]. Its importance continues to grow in emerging fields such as social robotics [9,10], gaming [11], and multimodal emotion analysis [12].

Acoustic sound characteristics used in SER research are divided into different subgroups, primarily prosodic and spectral. Prosodic characteristics include descriptors such as fundamental frequency (F0) and energy variation. Spectral characteristics, on the other hand, encompass descriptors such as Mel-Frequency Cepstral Coefficients (MFCC) and spectrogram-based representations.

With advancements in machine learning, the first SER approaches relied on traditional machine learning algorithms. Algorithms such as Support Vector Machines (SVM), Random Forests, k-Nearest Neighbors (k-NN), Gaussian Mixture Models (GMM), and Hidden Markov Models (HMM) were widely adopted and continue to be used to this day [13-17]. These classical machine learning algorithms rely on handcrafted feature extraction methods where the feature extraction process is manual. Feature selection and dimensionality reduction play a crucial role in the use of such algorithms. Therefore, methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Independent Component Analysis (ICA) are frequently used [18-20]. These methods provide low computational complexity and high interpretability.

Since the early 2010s, algorithms in the field of deep learning such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM), and Bidirectional Gated Recurrent Unit (Bi-GRU) have significantly improved SER technologies, outperforming classical machine learning approaches [21-24]. While CNNs efficiently extract features from 2D representations such as spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs), RNN-based models excel at modeling the temporal structure of

speech signals. LSTM and GRU networks further enhance long-term dependency learning and address issues such as vanishing gradients. Key advantages of these deep learning methods include automated feature extraction, strong generalization, and efficient learning from large datasets.

Transformer architectures [25], widely used in natural language processing (NLP), stand out in SER research due to their effectiveness in modeling long-range dependencies. Multi-head attention mechanisms have become important for SER tasks by enabling simultaneous analysis of time and frequency features. Transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers) [26], wav2vec 2.0 [27] and HuBERT (Hidden-Unit BERT) [28] achieve strong SER performances.

In modern SER systems, hybrid architectures combining multiple neural network paradigms are widely adopted. These hybrid architectures make significant contributions to SER technology by leveraging the strengths of different architectural approaches. Configurations such as CNN-LSTM [29,30], CNN-BiLSTM [31,32] and CNN-GRU [33,34] create powerful synergies by combining the spatial feature extraction of convolutional networks with the temporal processing power of recurrent units. Similarly, Transformer-CNN [35,37,38] hybrid architectures produce more robust feature representations by combining global context with local pattern detection. This research makes three important contributions to the field of emotion recognition through speech:

The first contribution presents a baseline model that integrates Conv1D layers, Efficient Channel Attention (ECA) mechanisms, and BiLSTM-Attention frameworks, offering high representational power despite its lightweight structure. This model is designed to provide strong initial performance in speech-based emotion recognition tasks.

The second contribution proposes a hybrid model that integrates the developed baseline model with advanced feature optimization strategies.

The third contribution advances computational efficiency through systematic dimensionality reduction, demonstrating that metaheuristic feature selection can reduce the 1024-dimensional deep feature space by up to 69.7% (ACO+SVM) while simultaneously improving classification accuracy thereby addressing the practical deployment constraints of SER systems in resource-limited environments.

Related Works

Akinpelu et al. [36] propose the implementation of a lightweight Vision Transformer (ViT) architecture to extract spatial features from mel-spectrogram representations. Testing the proposed approach on the TESS dataset yielded 98% accuracy and on the EMO-DB dataset 91% accuracy, demonstrating the effective generalization ability of the proposed approach across different emotional speech datasets. Xia and Zhao [39] introduced a hybrid CNN-BiLSTM architecture for enhanced temporal modeling of speech emotions using three-dimensional representations derived from MFCC features and their derivatives. The proposed approach uses attention-based weighting to highlight critical emotional features and improves emotional pattern recognition through bidirectional processing. Tests on the IEMOCAP and EMO-DB datasets yielded Unweighted Average Recall rates of 67.44% and 91.89%, respectively. Kundu et al. [40], developed a dual-channel attention-based framework combining Attention-Based Local Feature Blocks (ALFB) and Global Feature Blocks (GFB) with ECA-Net, 1D CNN, and BiLSTM components to extract localized patterns and temporal dependencies from MFCC, mel-spectrograms, RMS energy, and zero-crossing rate features. 5-fold cross-validation on TESS, RAVDESS, BanglaSER, SUBESCO, and Emo-DB datasets yielded superior classification accuracies of 99.65%, 94.88%, 98.12%, 97.94%, and 97.19%, respectively, outperforming existing methodologies. Xu et al. [41] aimed to capture both local and global contextual features by addressing the problem of fragmented emotional information in speech signals. In their proposed network model, they combined Bi-directional GRU (Bi-GRU) networks with multi-head attention mechanisms. Experimental validations on the IEMOCAP and Emo-DB datasets demonstrated high performance with unweighted accuracy of 75.04% and 88.93%, respectively. Ullah et al. [37] proposed CTENet, a hybrid architecture integrating parallel convolutional layers with Transformer encoders to simultaneously capture spatial and temporal emotion patterns from MFCC spectrograms. The framework demonstrated high performance with accuracy of 82.31% in RAVDESS (eight emotions) and 79.42% in IEMOCAP (five emotions), while maintaining computational efficiency with a compact 4.54 MB model size. Lee et al. [42] developed Fusion-ConvBERT, a parallel fusion model combining convolutional neural networks and bidirectional encoder representations from Transformers. The model demonstrated strong performance with speaker-dependent/independent unweighted accuracy rates of 92.1%/86.04% in EMO-DB and 71.36%/67.12% in IEMOCAP datasets. Maji et al. [43] developed a dual-channel speech emotion recognition framework that processes Mel-spectrograms via convolutional capsule networks (Conv-Caps), while processing the remaining spectral features (MFCCs, chromatograms, spectral contrast, zero-crossing rates, and RMS values) via a Bi-Directional Gated Recurrent Unit (Bi-GRU) and enhanced with self-attention mechanisms for selective emotional cue weighting. The proposed approach achieves weighted (WA)-unweighted (UA) accuracy rates of 90.31%-87.61% (EMO-DB), 76.84%-70.34% (IEMOCAP), and 87.52%-86.19% (SITB-OSED), respectively. Tang et al. [38] developed a SER framework that combines a CNN-Transformer architecture with a multidimensional attention mechanism to process speech information at different levels of detail. The approach uses CNN blocks for local time-frequency feature extraction, while a three-dimensional attention mechanism (time-channel-space) enhances interdimensional representations. Global and local dependencies are modeled through large depth-separable

convolutions paired with efficient Transformer components. In experiments on IEMOCAP and Emo-DB datasets, weighted/unweighted accuracy rates of 71.64%-72.72% and 90.65%-89.51% were achieved, respectively. Barhoumi and BenAyed [44] developed a Speech Emotion Recognition system that combines MFCC, ZCR, Mel spectrograms, RMS, and chroma features with data enhancement techniques. In the experiments conducted, Hybrid CNN+BiLSTM models showed better classification performance, outperforming MLP and CNN architectures. Accuracy rates of 100% were achieved in the TESS dataset, 99.50% in the EmoDB dataset, and 90.12% in the RAVDESS dataset.

Material and Methods

This section provides comprehensive information on the dataset used in the proposed approach, the feature extraction procedures, the implemented metaheuristic and classification algorithms, the attention mechanisms, and the LSTM and BiLSTM architectures. The proposed approach is explained by referring to these fundamental elements.

Proposed Model

This study proposes a hybrid SER approach that combines an attention-guided and deep learning-based network model with metaheuristic optimization algorithms to achieve high emotion recognition performance from speech signals. The approach is structured in two sequential phases. In the first phase, the Conv1D-ECA-BiLSTM-Attention baseline model network is used for robust extraction of distinctive emotion representations. In the second phase, metaheuristic optimization techniques are applied in conjunction with machine learning classifiers to improve feature selection and increase prediction accuracy.

The proposed model's process begins with pre-processing the audio segments in the dataset, which will be given as input to the baseline model. This involves resampling them to 16 kHz and normalizing them to a fixed duration of 4 seconds. The pre-processed signals are then divided into 124 frames. For each frame, 29 features are extracted: 20 MFCC coefficients, 6 Spectral Contrast (SC) coefficients, and one feature each for Spectral Rolloff (SR), Spectral Bandwidth (SB), and Zero Crossing Rate (ZCR). Consequently, each recording is represented as a 124×29 matrix. This representation matrix is then given as input to the baseline model.

The baseline model neural network structure, the first stage of our approach, consists of three parts. The first part comprises three parallel Conv1D-ECA-BatchNormalization branches. Each branch starts with a 1-dimensional convolution layer containing 64 filters. The first and second branches use 3 kernel sizes, while the third branch uses 5 kernel sizes. These convolution layers are designed to capture different time-dependent local spectral patterns and temporal dynamics. After convolution, the Efficient Channel Attention (ECA) mechanism adaptively recalibrates the feature maps on a channel basis. Batch normalization is applied to stabilize training and improve convergence. The outputs of the three parallel branches are concatenated to form a matrix (124 x 192). This concatenation is fed into the second part of the baseline model, a 256-unit BiLSTM, to capture temporal dependencies. Hidden states (124 × 512) containing forward-backward context for each time step and the last hidden state (512) for the last time step are obtained. The third part, the temporal attention mechanism, first generates scalar scores (124 × 1) by passing the states at all time steps through dense layers. These scores are scaled and normalized with Softmax to calculate importance weights (alpha) for each time step. The weights generated by the attention mechanism indicate how important the hidden states at each time step are. Using these weights, the hidden states at all time steps are weighted and combined to obtain a single context vector (512 dimensions) summarizing the most emotionally distinctive parts of the audio signal. This context vector is then concatenated with the final hidden state (512) of the BiLSTM, which represents the overall structure of the sequence [45]. As a result, a representation with 1024 features is obtained, encoding both local patterns and sequential temporal relationships. This representation is fed into the final dense layer to generate probabilities for each class, and classification is performed.

In the second stage of our approach, the 1024-dimensional feature vector obtained by concatenation of the context vector (512) and the final hidden state of BiLSTM (512) is subjected to feature selection using six metaheuristic algorithms: Particle Swarm Optimization, Dragonfly Algorithm, Artificial Bee Colony, Ant Colony Optimization, Genetic Algorithm, and Simulated Annealing. The applied algorithms aim to minimize the number of selected features while maximizing classification accuracy. The optimized features are evaluated using a number of conventional machine learning classifiers. A schematic overview of the proposed hybrid framework is presented in Fig. 1.

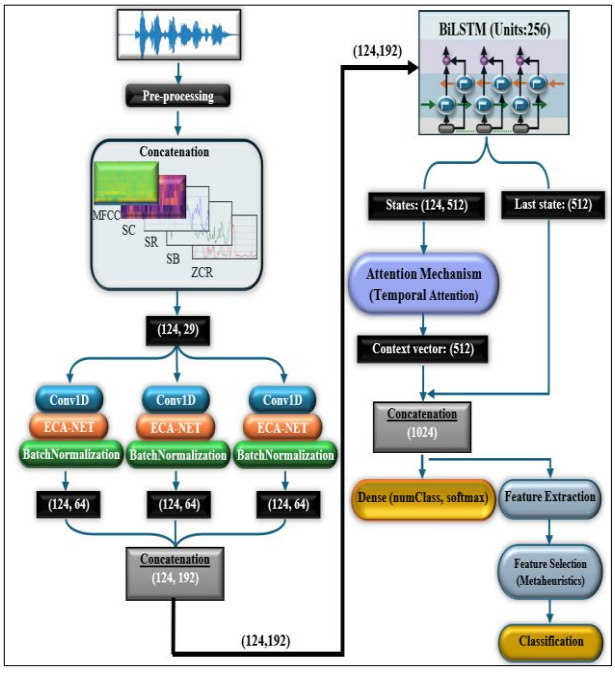

Figure 1. Proposed hybrid model.

Datasets

Created at the Technical University of Berlin, EMO-DB is a publicly available database of German emotional speech containing 535 emotion-labeled high-quality audio recordings of seven emotions (Anger, Boredom, Disgust, Fear, Happiness, Sadness, Neutral) voiced by 10 professional actors (equal numbers of men and women). Recorded in an anechoic chamber, the sampling frequency of the data was reduced from 48 kHz to 16 kHz. Each actor spoke daily sentences of ten different lengths, with a total of approximately 800 recordings including alternate takes [46].

Feature Extraction

In this study, handcrafted acoustic features primarily Mel-Frequency Cepstral Coefficients (MFCC), along with Spectral Contrast (SC), Spectral Rolloff (SR), Spectral Bandwidth (SB), and Zero Crossing Rate (ZCR) were extracted for the detection of emotions encoded in audio signals. These features will provide a more robust feature set for an emotion recognition system by considering the spectral and temporal characteristics of the signal.

i) Mel-Frequency Cepstral Coefficients

Davis and Mermelstein developed Mel-Frequency Cepstral Coefficients (MFCCs) in 1980, creating a fundamental feature extraction framework for audio applications [47]. These coefficients improve the accuracy of audio analysis by converting audio signals into spectral representations that are closest to human auditory perception. MFCC extraction approach is shown in Fig. 2.

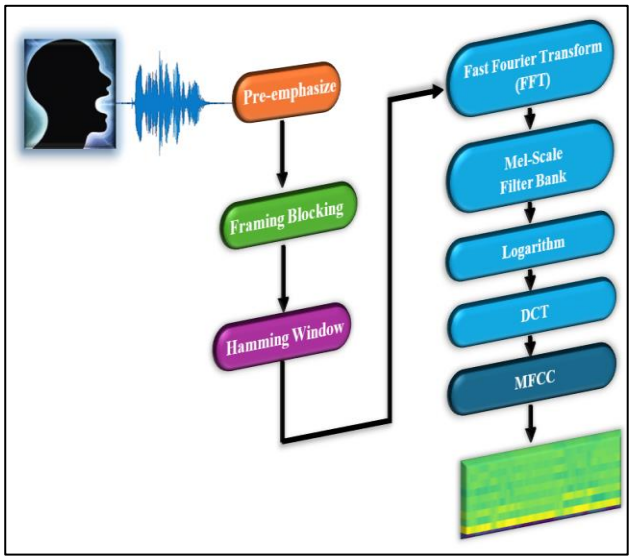

Figure 2. MFCC extraction process.

In speech signals, low frequencies are naturally more dominant. Pre-emphasis filtering is necessary to boost high-frequency components and reduce noise. The pre-emphasis filter is used to achieve spectral balance and improve signal quality for subsequent processing stages. Eq. 1 shows the pre-emphasis filter. In the equation, x(n) represents the original audio signal, x′(n) represents the filtered audio signal, and α represents the pre-emphasis coefficient, which varies between 0.9 and 1.0.

$$x'(n) = x(n) - \alpha x(n-1) \quad (1)$$

Temporal segmentation is necessary for effective signal analysis. The signal is divided into overlapping frames (usually 50% overlap) lasting 20-40 milliseconds to maintain analytical continuity. The total number of frames is given in Eq. 2. In the equation, N represents the total number of samples, L the frame length, R the shift amount, and M the total number of frames.

$$M = \left\lfloor \frac{N-L}{R} \right\rfloor + 1 \quad (2)$$

To reduce discontinuities between generated signal frames and minimize frequency domain distortions, a Hamming window is applied to the frames. When applied correctly, it significantly reduces spectral leakage. In this context, w(n) defines the Hamming window, x(n) the original frame signal, and x'(n) the windowed output, as shown in Eq. 3 and 4.

$$w(n) = 0.54 - 0.46\, cos\left(\frac{2\pi n}{M-1}\right) \quad (3)$$

$$x'(n) = x(n) * w(n) \quad (4)$$

The Fast Fourier Transform (FFT) converts frequency-domain features of time-domain audio signals into spectral forms, yielding representations vital for speech emotion recognition. These frequency-based features provide computationally efficient inputs for machine learning models in classifying emotional states. Eq. 5 details the FFT formulation.

$$X_{(k)} = \sum_{n=0}^{N-1} x_{(n)} \cdot e^{-j\left[\frac{2\pi kn}{N}\right]} \quad (5)$$

In the equation, $X_{(k)}$ represents the frequency-domain components, $x_{(n)}$ represents the windowed frame signal in the time domain, k is the frequency index, n is the time index, and N is the total number of samples. Fig. 3 illustrates the transformation from the time domain to the frequency domain.

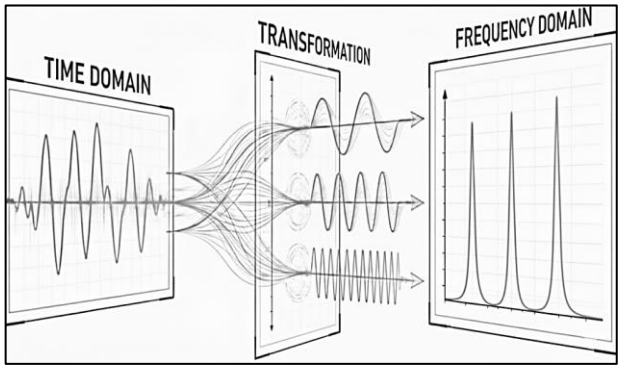

Figure 3. Transformation from the time domain to the frequency domain [48].

After the FFT process, frequencies are scaled to the Mel scale to better reflect human auditory perception and to enable the application of Mel-scale filter banks. This is described in Eq. 6. The human ear is more sensitive to

changes at low frequencies and less sensitive at high frequencies; the Mel filter bank mimics this by using narrow and closely spaced filters at low frequencies and wide and widely spaced filters at high frequencies. It usually consists of 20-40 triangular filters and is applied to the power spectrum output of the FFT, and the energy value for each filter band is calculated as given in Eq. 7. In the equation, E_m represents the total energy obtained from the m-th filter band, $H_m[k]$ represents the weighting value of the m-th Mel filter at the k-th frequency, and $|X(k)|^2$ represents the power from the FFT at the k-th frequency point.

$$Mel(f) = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \quad (6)$$

$$E_m = \sum_{k=0}^{N-1} |X(k)|^2 \cdot H_m[k] \quad (7)$$

The energy outputs from the Mel-scale filter bank are transformed into the logarithmic domain to mimic the auditory system's sensitivity to intensity changes. This compression provides more robust features by emphasizing perceptually significant changes. In the transformation used in Eq. 8, E_m represents the energy parameter.

$$L(m) = ln(E_m) \quad (8)$$

Nasir Ahmed, T. Natarajan, and K. R. Rao introduced the Discrete Cosine Transform (DCT) in 1974, making a seminal contribution to signal processing [49]. This orthogonal transformation method removes the correlation between values from the Mel filter bank using DCT. It compresses many related outputs from the Mel filter bank into 12-13 independent coefficients, preserving critical spectral information for speech recognition systems. By reducing the dimensionality to only 12-13 coefficients, it lowers the computational cost and allows machine learning models to operate more efficiently. In speech representation, DCT is the essential final step in generating Mel-Frequency Cepstral Coefficients (MFCCs). The DCT calculation is given in Eq. 9. The representation of the MFCC coefficients is given in Fig. 4.

$$C[n] = \sum_{m=0}^{M-1} L[m] \cdot cos\left(\frac{\pi n(m+0.5)}{M}\right) \quad (9)$$

In the equation, the parameters $C[n]$ represent the DCT coefficients, $L[m]$ the log energy values, M the number of Mel filters, m the filter index, and n the cepstral coefficient index.

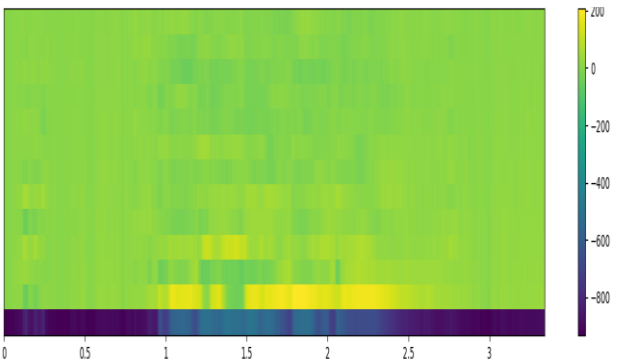

Figure 4. MFCC coefficients

ii) Spectral and Temporal Features

Spectral measurements, which significantly characterize the frequency structure of the audio signal, play an important role in speech emotion recognition (SER) studies. Spectral Contrast (SC) indicates the signal's clarity by measuring the difference between the highest and lowest energy regions in the frequency spectrum on a logarithmic scale. Spectral Rolloff (SR) reflects the signal's brightness by defining the frequency at which a certain percentage (usually 85% or 95%) of the total spectral energy remains below a certain level. Spectral Bandwidth (SB) indicates the width of the frequency distribution by measuring how far the energy is spread around the spectral center. In our study, Zero Crossing Rate (ZCR), although a time-domain measurement, is often used in conjunction with spectral characteristics due to its strong correlation with high-frequency content. This characteristic measures how many times the audio signal crosses the zero axis in a given time interval. Higher ZCR values indicate higher-frequency or noisier signals, while lower ZCR values indicate lower-frequency and more regular structures. Visual representations of the features other than MFCC used in our study are given in Fig. 5.

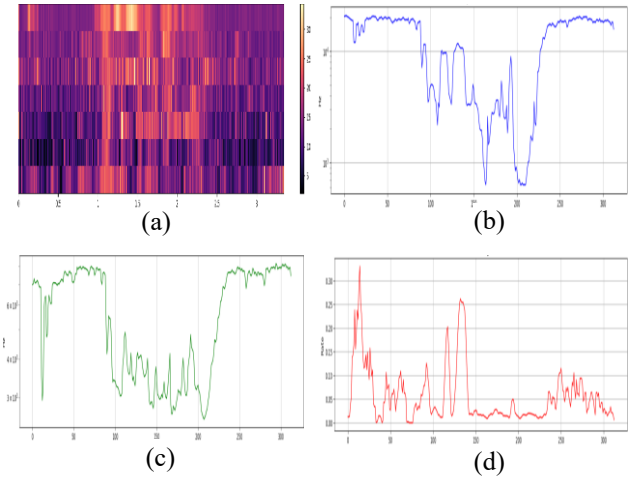

Figure 5. (a) Spectral Contrast (SC), (b) Spectral Rolloff (SR), (c) Spectral Bandwidth (SB), (d) Zero Crossing Rate (ZCR)

Metaheuristic Algorithms

Metaheuristic algorithms are advanced optimization methods inspired by nature, biology, or physical processes. This study aims to improve model performance and reduce computational cost by selecting the most informative features representing emotion in the audio signal, which is the second stage of the proposed hybrid model. Six different metaheuristic optimization methods were used for this purpose: Particle Swarm Optimization, Simulated Annealing, Genetic Algorithm, Dragonfly Algorithm, Artificial Bee Colony Algorithm, and Ant Colony Optimization.

Developed by Kennedy and Eberhart in 1995, Particle Swarm Optimization (PSO) is a metaheuristic algorithm that mimics swarm behavior in birds [50]. Particles in the

swarm represent a population of candidate solutions in the search space by updating their positions through personal experience (PBest) and global knowledge (GBest). By continuously improving their positions until satisfactory results are obtained, the particles effectively solve nonlinear optimization problems [50-52].

The Simulated Annealing (SA) method, introduced by Kirkpatrick et al. (1983) [53], iteratively explores neighboring solutions using a temperature-dependent acceptance probability that balances exploration with convergence. The algorithm starts from a single solution point and avoids local minima by accepting poor solutions with a certain probability at high temperatures, becoming more selective as the temperature decreases and converging to the optimum solution. The temperature is gradually reduced until a termination criterion is met.

The Genetic Algorithm (GA) originated from John Holland's groundbreaking 1975 work "Adaptation in Natural and Artificial Systems" and, particularly through the research of David E. Goldberg [54,55], has created a biologically inspired computational framework. The algorithm starts with a group of potential solutions (population) and selects the best solutions (selection), applies crossover to generate new offspring (crossover), introduces random changes (mutation), and improves the population over generations by repeating this cycle. This cycle continues until a predefined number of iterations is completed or the desired fitness level is reached.

The Dragonfly Algorithm (DA), developed by Seyedali Mirjalili in 2016, is a swarm intelligence optimization method inspired by the collective behavior of dragonflies [56]. This approach encompasses five behavioral mechanisms: avoiding collisions with neighbors (Separation), aligning with neighbor speeds (Alignment), center-seeking tendency (Cohesion), resource-oriented orientation (Attraction), and avoiding enemies (worst-case solutions) (Distraction).

Developed in 2005 by Derviş Karaboğa [57], inspired by the foraging behavior of honeybees, the Artificial Bee Colony (ABC) algorithm is a population-based optimization method. It models three types of artificial bees (employed bees, onlooker bees, and scout bees) that collaboratively explore the search space. Employed bees generate new solutions by performing local searches, onlooker bees are attracted to probabilistically promising solutions, and scout bees randomly discover new solutions when stagnation occurs. The algorithm evaluates solution quality through a fitness function and guides the swarm towards optimal solutions through an adaptive, heuristically guided search process.

Ant Colony Optimization (ACO) is a metaheuristic algorithm developed by Marco Dorigo in 1992, inspired by the pheromone trails left by ants foraging for food [58]. The algorithm simulates how ants use pheromone trails and heuristic cues to create paths representing potential solutions. Ants iteratively explore the solution space by guiding their search toward optimal routes through pheromone updates via evaporation and reinforcement.

Classification Algorithms

Classification algorithms are machine learning techniques used to separate data into predefined classes. In this study, the suitability of the solutions found by metaheuristic algorithms in the search space during the feature selection phase will be evaluated using three different classification methods. These methods are Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Multilayer Perceptrons (MLP).

Support Vector Machines (SVM) were first introduced in 1995 by Vladimir Vapnik and Corinna Cortes [59]. SVM is a supervised learning method that aims to determine the hyperplane that best distinguishes between classes. The decision boundary is based on critical data points called support vectors. While effective for linear problems, it can also achieve high performance in nonlinear datasets thanks to its kernel functions.

The K-Nearest Neighbor (KNN) algorithm was developed in 1967 by Thomas Cover and Peter Hart [60]. KNN is an algorithm used for pattern recognition and classification. During classification, the K nearest neighbors of the test sample in the training dataset are determined, and the class of the majority of these neighbors is assigned as the class of the test sample. The correct selection of the K parameter directly affects the success of the model.

The Multilayer Perceptron (MLP) was popularized in 1986 by David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams along with the backpropagation algorithm [61]. MLP is a feedforward neural network model with at least one hidden layer. Neurons in each layer weight the inputs and transmit them to the next layer by transforming them through activation functions. It has the ability to model nonlinear relationships.

Attention Mechanism

Attention mechanisms are a method that enhances salient contextual representations of a model by weighting significant portions of input data. In this study, we use Efficient Channel Attention (ECA) at the channel level to highlight distinctive channel features and the Bahdanau attention mechanism to model global temporal dependencies.

i) ECA-Net

Introduced in the article “ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks” [62], designed as a simplified attention module for convolutional neural networks, ECA-Net offers a lightweight alternative to complex designs such as Squeeze-and-Excitation Networks (SENet) [63] in tasks such as image classification and object detection without incurring high computational costs. It uses global average pooling followed by simple 1D convolution and sigmoid activation to generate channel-based attention weights to effectively highlight informative features and improve original representations. The Efficient Channel Attention (ECA) module is shown in Fig. 6.

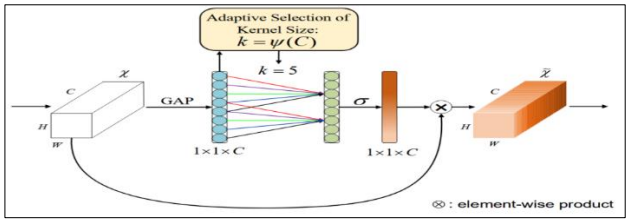

Figure 6. Efficient Channel Attention (ECA)

ii) Temporal Attention

In 2014, Bahdanau, Cho, and Bengio introduced attention mechanisms to address information bottlenecks in neural machine translation [64]. Unlike conventional encoder-decoder systems that compress inputs into fixed vectors, attention enables dynamic focus on relevant sequence elements during generation. This innovation resolves long-dependency challenges and substantially improves model accuracy. The approach later became foundational to Transformer architectures, now central to contemporary NLP and vision systems [25]. Equations 10 through 13 present the mathematical formulations for the sequential stages of the attention mechanism: h_t' transformation of hidden representations, e_t generation of attention scores, α_t normalization through softmax operation, and c computation of the final context vector, respectively. In the proposed model, the temporal attention mechanism is designed as an adaptive weighting system applied to the output sequence of the BiLSTM layer. This mechanism assigns varying importance to different temporal positions in the time series, allowing the model to selectively focus on performance-critical time periods. Fig. 7 illustrates the structure of the temporal attention mechanism referenced from Bahdanau et al. and employed in this study.

$$h_t' = tanh(W_1 \cdot h_t) \tag{10}$$

$$e_t = W_2 \cdot h_t' \tag{11}$$

$$\alpha_t = \frac{ex\,p(e_t)}{\sum_{k=1}^{T} ex\,p(e_k)} \tag{12}$$

$$c = \sum_{t=1}^{T} \alpha_t \cdot h_t \tag{13}$$

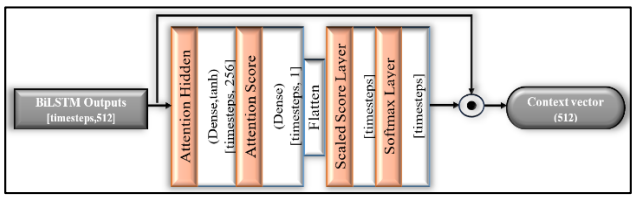

Figure 7. Temporal Attention

Fig. 7 illustrates attention mechanism used in our model that generates weighted contextual representations using hidden states produced by BiLSTM. The bidirectional LSTM produces forward and backward hidden states for each time step, collected in the states variable. These states pass through a dense layer with tanh activation to form attention-based representations, which are subsequently transformed into scores via linear projection. Softmax normalization converts these scores into attention weights (alpha) that quantify temporal importance. The final context vector (context vector) emerges from element-wise multiplication between attention weights and original states using the Dot operation. This context vector concatenates with the BiLSTM's final hidden state, providing enriched representations. This structure enhances overall model effectiveness by allowing focus on specific time periods.

LSTM and BiLSTM

LSTM networks regulate memory through gated mechanisms, with the forget gate (f_t) playing a central role in filtering past information. It combines the current input (x_t) and previous hidden state (h_{t-1}) via a weight matrix (W_f), applies a sigmoid function, and generates values between 0 and 1. These values modulate the prior cell state (C_{t-1}) through element-wise multiplication, allowing the model to retain only information relevant for long-term memory. The input gate (i_t) regulates the incorporation of new information into memory by applying a sigmoid activation to the processed inputs (h_{t-1}, x_t) via the weight matrix W_i. Its output modulates the update vector ($\widetilde{C_t}$) through element-wise multiplication, determining the new content added to long-term memory. The update vector ($\widetilde{C_t}$) processes inputs (h_{t-1}, x_t) through the weight matrix (W_C) followed by a tanh activation to generate candidate values for memory update. The new cell state (C_t) is then computed by combining the forgotten content filtered by the forget gate with newly selected information regulated by the input gate. The output gate (o_t) processes the inputs (h_{t-1}, x_t) through weight matrix (W_o) followed by a sigmoid activation to regulate the cell's new hidden state (h_t). This hidden state is computed by element-wise multiplying the output gate with the tanh-activated updated cell state (C_t) [65]. Equations 14-19 detail the full formulation. Fig. 8 illustrates the working principle of the LSTM.

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{14}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{15}$$

$$\widetilde{C_t} = \tanh(W_C[h_{t-1}, x_t] + b_C) \tag{16}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \tag{17}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{18}$$

$$h_t = o_t * \tanh(C_t) \tag{19}$$

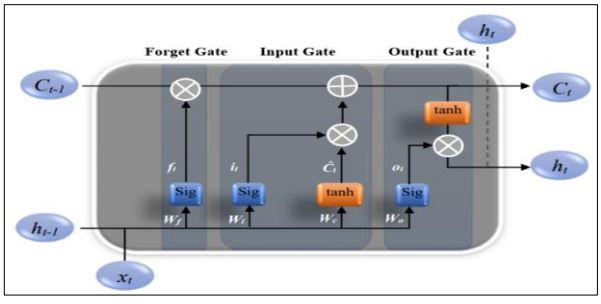

Figure 8. LSTM [65]

Bidirectional Long Short-Term Memory (BiLSTM) networks enhance traditional recurrent neural networks by processing input sequences in both forward and backward temporal directions. Building upon the foundational LSTM architecture introduced by Hochreiter and Schmidhuber, BiLSTMs utilize two LSTM layers - one handling the sequence from past to future and the other from future to past - whose outputs are typically combined to create a richer temporal representation [66]. By leveraging memory cells and gating mechanisms, this design effectively addresses challenges related to capturing long-range dependencies in sequential data. Consequently, BiLSTM architectures are particularly well-suited for applications requiring extensive contextual understanding, such as speech recognition and emotion detection. Fig. 9 illustrates the working principle of the BiLSTM.

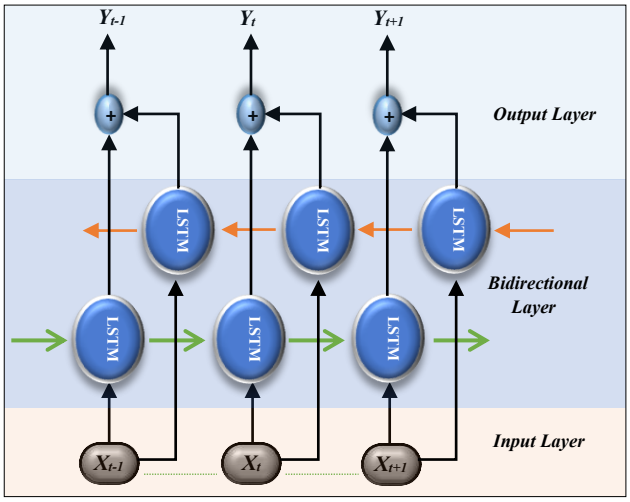

Figure 9. BiLSTM

Experimental Evaluation

This section discusses the dataset and experimental setup used in the study, performance metrics, and results. First, the experimental environment, methods, and configurations are described, followed by a comparative presentation of the metrics used in performance evaluation and the obtained results.

Experimental Setup

In this study, the Berlin Emotional Speech Database (EMO-DB) was used to conduct the experiments. This dataset consists of a total of 535 voice samples recorded in German and includes seven different emotion classes: anger, boredom, disgust, fear, happiness, sadness, and neutral. Raw audio recordings were divided into 1024-sample-long frames, with a 512-sample shift applied between each frame. As a result of this process, each audio recording was divided into 124 frames, resulting in a feature vector with 29 components for each frame. The extracted features include MFCC (obtained from Pre-emphasis, Framing, Windowing, Fast Fourier Transform, Mel-Scale Filter Bank, Logarithm, and Discrete Cosine Transform stages), Spectral Contrast, Spectral Bandwidth, Spectral Rolloff, and Zero Crossing Rate. The entire dataset was divided into 80% training + validation (428 samples) and 20% testing (107 samples) per fold. To prevent data leakage, scaling and other preprocessing steps were applied only to the training data, and these transformations were directly transferred to the test data. Furthermore, the 107-sample test set was held out entirely from both the baseline model training and the metaheuristic optimization-classifier process; the test set was accessed only for final performance evaluation after all training and optimization procedures were complete. After completing the training of the proposed deep learning-based baseline model, 1024-dimensional deep feature vector was extracted from the layer before the classification layer. These vectors were incorporated into the optimization-based feature selection integrated with classifiers process within the baseline architecture. Finally, the model's performance was measured using 5-fold evaluation protocols. This study is based on a speaker-dependent protocol. Since the primary objective is to demonstrate the contribution of metaheuristic-based dimensionality reduction to computational efficiency rather than cross-speaker generalization, the Leave-One-Speaker-Out (LOSO) protocol was not applied within the scope of this work. Figure 10 shows the 5-fold assessment partitioning of the EMO-DB dataset with per-fold (training+val) - test partitioning. Fig. 11 also shows the emotion distribution numbers for the EMO-DB dataset.

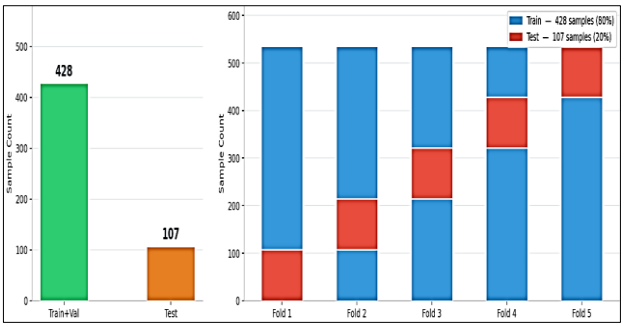

Figure 10. Visualizations of EMO-DB data allocations, depicting the train-validation versus test split, the 5-fold cross-validation scheme

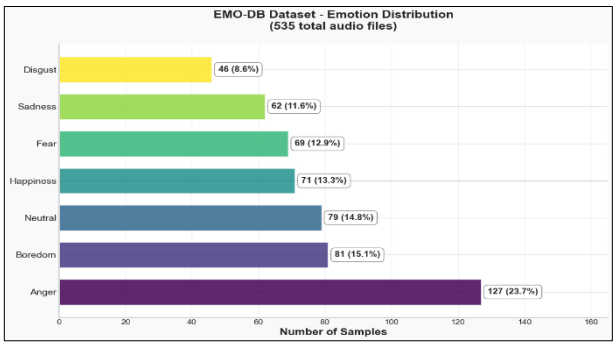

Figure 11. EMO-DB emotion distribution

Evaluation Metrics

In this study, the evaluation framework employs multiple performance metrics. Weighted Accuracy (WA) measures the overall classification success, while Unweighted Accuracy (UA) eliminates class imbalance, providing equal averages across all emotion categories. The F1-score

provides a balanced combination of precision and recall. Confusion matrices (CM) visualize the model's category-wise performance profile. The relevant formulations are presented in Equations 20-24. A 5-fold cross-validation methodology was adopted for robust validation, dividing the dataset into equal parts and using a different part as a test in each iteration. The UA metric is a performance metric that represents the arithmetic mean of the accuracy values calculated for each class to ensure all classes are evaluated with equal weight.

$$\mathrm{WA} = \frac{\sum_{j=1}^{m} C_j}{\sum_{j=1}^{m} T_j} \tag{20}$$

$$\mathrm{UA} = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{C_j}{T_j}\right) \tag{21}$$

$$Precision_j = \frac{TP_j}{TP_j + FP_j} \tag{22}$$

$$Recall_j = \frac{TP_j}{TP_j + FN_j} \tag{23}$$

$$F1_j = \frac{2 \cdot Precision_j \cdot Recall_j}{Precision_j + Recall_j} \tag{24}$$

In the equations above, the notation is interpreted as follows: m denotes the total number of classes, and j indexes an individual class. T_j represents the number of instances belonging to class j, while C_j is the count of class-j instances that are correctly classified. TP_j (true positives) are cases that genuinely belong to class j and are predicted as j; FP_j (false positives) are cases that do not belong to class j but are nevertheless predicted as j; and FN_j (false negatives) are cases that belong to class j but are predicted as some other class.

Experimental Results Analysis

This section details the performance findings of the developed hybrid methodology on the EMO-DB dataset. The aim of the proposed approach in this research is to create a powerful hybrid system by combining metaheuristic optimization methods and machine learning classifiers based on deep feature representations extracted from a deep learning framework. The training parameters of the model are presented in Table 1, the parameters for metaheuristic algorithms in Table 2, and the parameters for traditional classifiers in Table 3. Experimental studies were conducted on a computer with an Intel Core i7-13700KF processor (3.40 GHz), 32 GB RAM, NVIDIA RTX 3060 graphics card, and 64-bit Windows 11 Pro operating system. The training and testing processes for the Baseline and Hybrid (baseline + optimization-classification) models in the 5-fold cross-validation took a total of 908.5 minutes.

Table 1. Training-Testing parameters and experimental configuration

Parameter	Value	Description
Batch Size	16	Mini-batch size
Maximum Epochs	100	With early stopping
Optimizer	Adam	Learning rate = 0.001
Loss Function	Categorical Crossentropy	Multi-class classification
Early Stopping	Patience = 10	Based on validation accuracy
Learning Rate Reduction	Factor = 0.5, Patience = 7	Adaptive learning rate
Random Seed	42	For reproducibility
CV Strategy	5-Fold Stratified	Stratified cross-validation

Table 2. Metaheuristic parameters and experimental configuration

Algorithm	Initial Population	Iterations	Main Parameters
PSO	50	200	w=0.6, c1=2.0, c2=2.0
GA	50	200	Crossover=0.9, Mutation=0.05, Tournament=3
SA	1 (no population)	200	T_initial=500, T_final=0.01
ABC	50	200	Employed=25, Onlooker=25, Limit=100
ACO	50	200	α=1.0, β=2.5, ρ=0.1
DA	50	200	w=0.9, s=2.0, a=2.0, c=2.0, f=2.0, e=1.0

Table 3. Classifiers parameters and experimental configuration

Classifier	Main Parameters
SVM	Kernel=RBF, C=1.0, Gamma=Scale
KNN	k=5, Weights=Uniform, Metric=Minkowski
MLP	Hidden=(100,), Max_iter=300

Evaluation Baseline and Hybrid Model

This section presents the experimental results of a 5-fold cross-validation performed to evaluate the performance of the proposed speech emotion recognition system's baseline model and hybrid model, respectively. Weighted Accuracy (WA), Unweighted Accuracy (UA), and F1-Score metrics were used for performance evaluation, and the results of the baseline and hybrid models were analyzed comparatively.

In the experiments conducted, a UA test performance of 77.69% was measured after baseline model training. The confusion matrix showing this is given in Fig. 12.

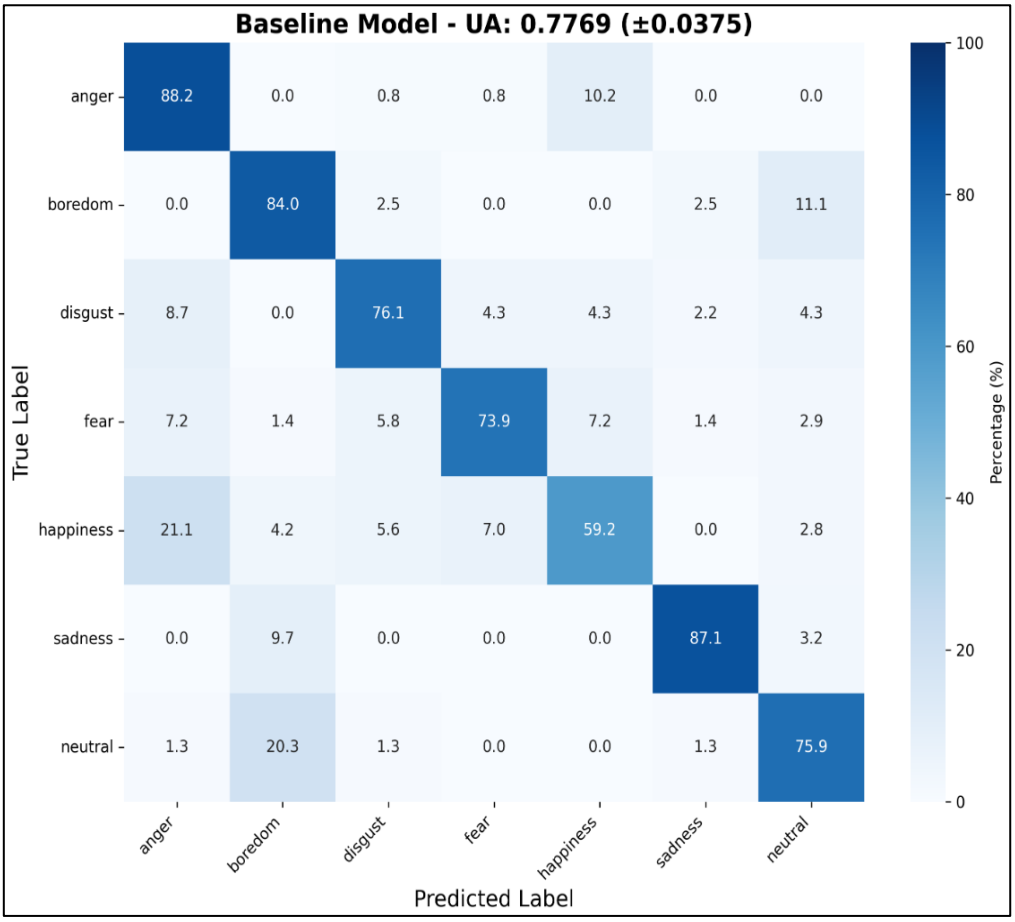

Figure 12. Baseline model confusion matrix (5-FOLD)

Examining the confusion matrix of the baseline model in Fig. 12, it is seen that the baseline model performed well in recognizing the emotions of anger (88%), sadness (87%), and boredom (84%), while the happiness class showed the lowest performance (59%) due to significant confusion with the emotion of anger (21%). The low recognition performance of the happiness and neutral (75.9%) classes indicates that their acoustic properties overlap with the emotions of anger and boredom, respectively. Remarkably, the emotion of disgust achieved strong accuracy (76%) despite limited representation (n=46, 8.6%). This suggests that disgust has distinctive acoustic properties and a characteristic spectral structure. The relatively low performance of the fear class (73.9%) and the error profile distributed across all classes reflect the acoustic diversity of the expression of fear. In addition to all this, our baseline model measured 77.69% unweighted accuracy and 78.88% weighted accuracy in overall classification performance.

Following the baseline model, an enhanced hybrid model incorporating metaheuristic optimization techniques integrated with three different classification techniques was subjected to a 5-fold cross-validation evaluation. Table 4 presents the test performance metrics of the hybrid model enhanced with optimization and classification algorithms. Fig. 13 also shows the heat maps of the UA test results of the optimization and classifier combination.

Table 4. Performance metrics of the proposed hybrid model under 5-fold cross-validation

Algorithm	Classifier	WA Mean±Std	UA Mean±Std	F1 Score ±Std	Features	Reduction
PSO	SVM	0.8411±0.0345	0.8305±0.0334	0.8360±0.0329	490±13.3	52.1±1.3%
PSO	KNN	0.8037±0.0350	0.7790±0.0305	0.7867±0.0326	496±5.4	51.5±0.5%
PSO	MLP	0.8262±0.0456	0.8135±0.0452	0.8169±0.0470	495±13.5	51.7±1.3%
GA	SVM	0.8486±0.0286	0.8373±0.0289	0.8440±0.0274	463±8.2	54.8±0.8%
GA	KNN	0.8131±0.0374	0.7914±0.0342	0.8017±0.0335	474±22.1	53.7±2.2%
GA	MLP	0.8393±0.0478	0.8284±0.0457	0.8312±0.0510	485±25.6	52.6±2.5%
SA	SVM	0.8355±0.0268	0.8269±0.0225	0.8310±0.0239	511±7.4	50.1±0.7%
SA	KNN	0.7888±0.0367	0.7628±0.0345	0.7706±0.0338	517±9.4	49.6±0.9%
SA	MLP	0.8336±0.0448	0.8215±0.0430	0.8231±0.0460	522±4.9	49.0±0.5%
ABC	SVM	0.8393±0.0254	0.8293±0.0229	0.8353±0.0243	414±23.6	59.6±2.3%
ABC	KNN	0.7925±0.0398	0.7702±0.0323	0.7781±0.0360	448±23.3	56.3±2.3%
ABC	MLP	0.8430±0.0320	0.8324±0.0285	0.8369±0.0328	474±11.1	53.7±1.1%
ACO	SVM	0.8336±0.0361	0.8219±0.0316	0.8276±0.0331	311±6.2	69.7±0.6%
ACO	KNN	0.8093±0.0305	0.7879±0.0209	0.7938±0.0260	340±37.0	66.8±3.6%
ACO	MLP	0.8093±0.0386	0.7991±0.0366	0.7998±0.0428	365±50.8	64.3±5.0%
DA	SVM	0.8374±0.0337	0.8266±0.0330	0.8326±0.0332	496±12.1	51.6±1.2%
DA	KNN	0.8019±0.0280	0.7795±0.0204	0.7882±0.0226	499±13.6	51.3±1.3%
DA	MLP	0.8393±0.0342	0.8274±0.0333	0.8310±0.0348	492±13.5	51.9±1.3%

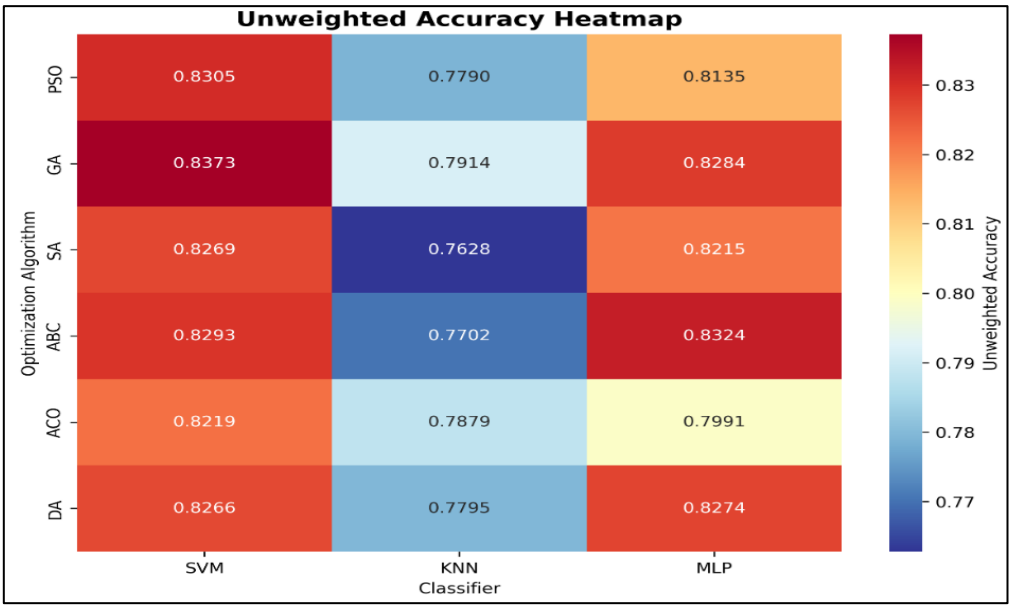

Figure 13. Proposed Hybrid Model Unweighted Accuracy Heatmap (5-Fold)

When examining Table 4 and Fig. 13 of the experimental results of the hybrid model, the consistent difference between the WA and UA metrics is noteworthy. The fact that WA values are systematically higher than UA values in all configurations indicates that class imbalance in the dataset and superior performance achieved in larger classes positively influence the weighted metrics. This suggests that the UA metric should be preferred for a balanced performance evaluation. Table 4 shows that the GA and SVM combination provides the highest performance metrics, with WA at 84.86%, UA at 83.73%, and F1-Score at 84.40%. The second best combination in the performance ranking is ABC+MLP, and the third is PSO+SVM. When the GA+SVM results are compared with the baseline model, a 7.8% increase in accuracy is achieved in the UA metric, and a 54.8% dimensionality reduction in feature representation is achieved, resulting in computational efficiency. The lowest performance values were observed in the SA+KNN combination. The highest compression ratio in feature reduction was achieved with the ACO+SVM combination at 69.7%, and overall, the highest feature reduction rate was obtained with the ACO algorithm. The Support Vector Machine classifier outperformed the KNN and MLP classifiers across all six optimization algorithms, with an average UA value of 0.8288. This finding clearly demonstrates that classifier selection is a critical factor in emotion recognition systems.

The convergence graphs and confusion matrices of the optimization algorithms of the three best hybrid models are given in Figures 14 and 15, respectively. These convergence curves and confusion matrices provide detailed information about the iterative learning dynamics and classification details of the most successful combinations.

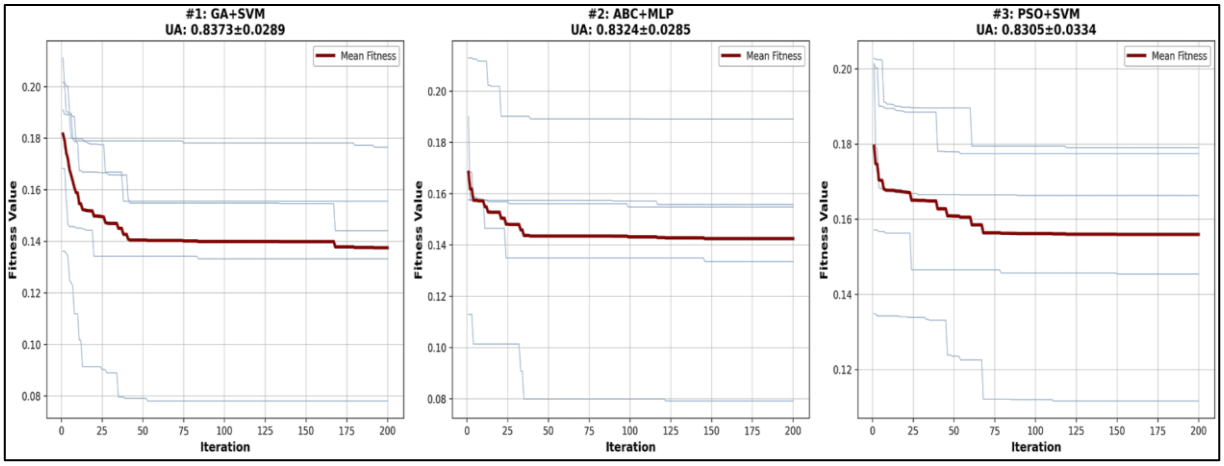

Figure 14. Convergence curves of the optimization process for the three best performing

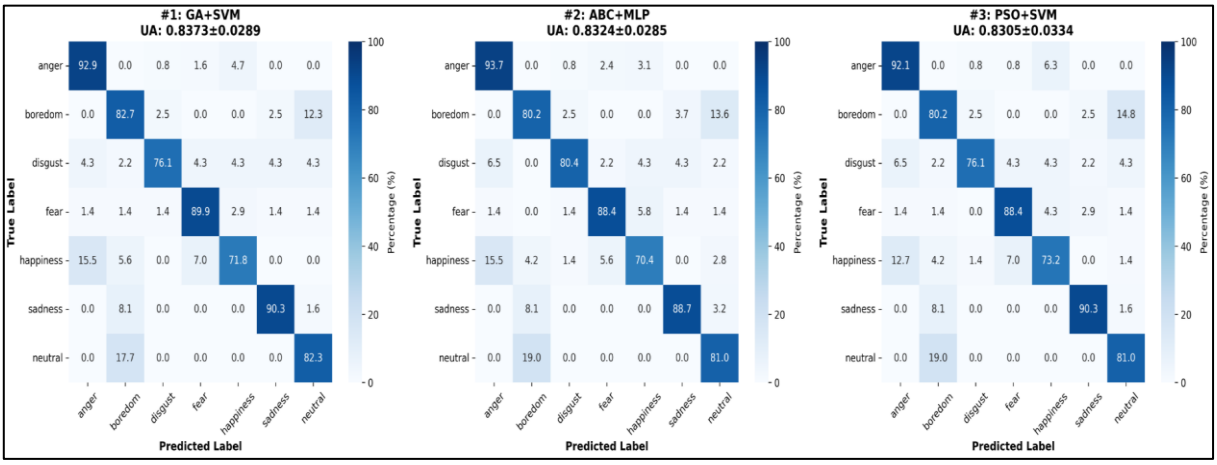

Figure 15. Confusion matrices of the top 3 combinations

As shown in Fig. 14, the GA+SVM combination exhibited both the fastest and best convergence, reaching the lowest fitness value at approximately 50 iterations in the 5-fold cross-validation average. The gray lines in the graph show the individual convergence curves of each fold, while the dark main line shows the average convergence. On the other hand, compared to the baseline model shown in Fig. 12 with a performance of UA: 0.7769 (±0.0375), the hybrid model (Baseline-GA+SVM) shown in Fig. 15 achieved a UA: 0.8373 (±0.0289) with a 7.8% absolute improvement. When examined at the emotion level, a remarkable jump was observed in the anger class (88.2% to 92.9%), a dramatic

improvement was recorded in fear perception (73.9% to 89.9%), sadness recognition success increased (87.1% to 90.3%), and most importantly, the happiness class, which had the lowest classification score in the baseline, rose from 59.2% to 71.8%. In the baseline model, the emotion of happiness was misclassified as anger 21.1% of the time, while this rate decreased to 15.5% in the hybrid model. Conversely, the misclassification of the emotion of neutral as boredom occurred 20.3% in the baseline model and 17.7% in the hybrid model. This finding indicates that pairs of neutral and boredom emotions, which have low arousal levels, pose a systematic classification challenge for both models. This may be due to the spectral similarity of these emotions and may point to the need to add prosodic features to the model. In the overall evaluation, it is seen that genetic algorithm-based feature selection offers a balanced performance increase in all classes by optimizing the decision limits of SVM. The computational cost of the GA+SVM combination arises exclusively during the offline training phase; at inference, classification is performed over 463 selected features (54.8% dimensionality reduction from the original 1024-dimensional space). Figure 16 presents a comprehensive comparison of the baseline and hybrid models. It also details the changes in accuracy levels, optimization times, and feature representation among different optimizer-classifier combinations. The results show that the hybrid model delivers better results than the baseline model in terms of both accuracy and feature representation.

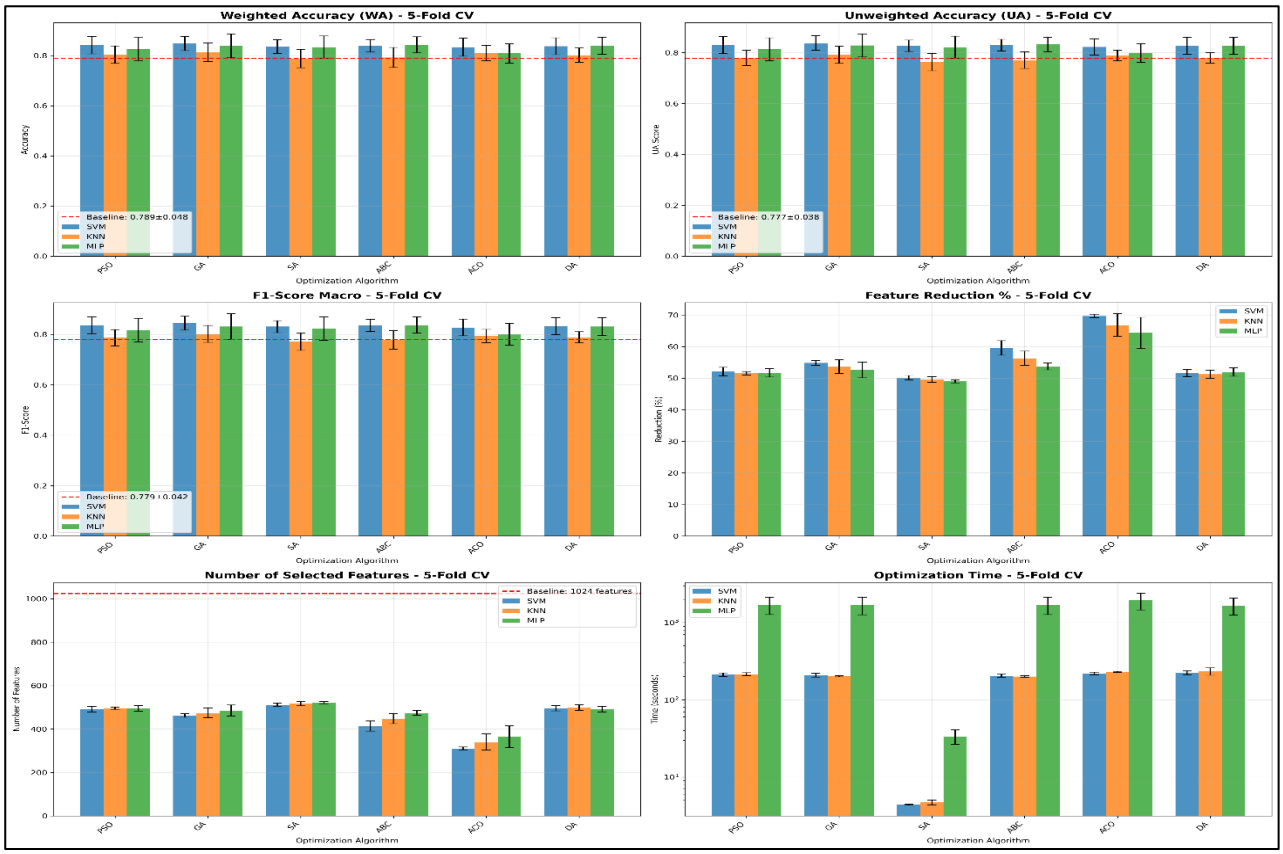

Figure 16. Comparison of hybrid and baseline model in terms of accuracy and feature representation

Speech emotion recognition experiments were conducted using the EMO-DB dataset with both a baseline and a hybrid model. Table 5 shows the unweighted accuracy (UA) and dimensionality reduction achieved through feature selection between the baseline and hybrid models after the experiments.

Table 5. Comparison of the baseline model and the proposed hybrid model

MODEL	**5-FOLD (UA%)**
Baseline	77.69
Hybrid	83.73
Increase Accuracy (%)	7.8
Feature Reduction (%)	54.8 (1024 →463)

A performance comparison of our proposed approach with those in the literature is presented in Table 6. Some results in the literature are not included in the table because they did not use 5-fold cross-validation for reliability, 7 emotion categories, and UA. Examining the table, we see that both our baseline model and hybrid model approaches achieved successful results in 5-fold cross-validation. The main objective of this study is to demonstrate the effectiveness of the proposed hybrid model in improving classification accuracy while reducing feature dimensionality, using features obtained with an attention-based deep learning model.

Table 6. The proposed hybrid approach is compared with literature methods

Authors/Reference	Method	CV Strategy	Accuracy (UA%)
Singh et al.[67]	CQT-MSF + DNN-SVM	5-FOLD	76.17
This Paper	**Baseline Model**	**5-FOLD**	**77.69**
Roh & Lee [68]	CNN + BiLSTM	5-FOLD	78.16
Liu et al.[69]	SVM+RBF	5-FOLD	78.66
Li et al.[70]	Bi-A2CEmo	5-FOLD	80.16
Ancilin & Milton[71]	SVM	5-FOLD	81.50
This Paper	Hybrid Model (PSO+SVM)	5-FOLD	83.05
This Paper	Hybrid Model (ABC+MLP)	5-FOLD	83.24
Hou et al. [72]	DSNMF	5-FOLD	83.30
This Paper	**Hybrid Model (GA+SVM)**	**5-FOLD**	**83.73**
Mishra et al.[73]	DNN	5-FOLD	84.72
Issa et al.[74]	1D CNN	5-FOLD	86.10
Mishra et al.[75]	DNN	5-FOLD	91.59

Conclusion

This study proposes a hybrid classifier model for speech emotion recognition. A hybrid model based on feature selection and accuracy improvement was developed by integrating six optimization algorithms (PSO, GA, SA, ABC, ACO, DA) and three classifiers (SVM, KNN, MLP) into the attention-based Conv1D-ECA-BiLSTM-Attention baseline deep network model. In experiments conducted with the EMO-DB dataset, the hybrid model performed better than the baseline model in a 5-fold cross-validation evaluation. In tests on the hybrid model, the GA+SVM configuration showed the best performance in the UA metric, reducing the feature dimensionality by 54.8% compared to the baseline model and achieving an overall accuracy of 83.73% with a 7.8% improvement in UA. In the overall evaluation, the SVM classifier demonstrated the most consistent performance. Experimental findings reveal that the proposed hybrid approach offers both high classification accuracy and computational efficiency. These results demonstrate that the model provides a balanced and effective solution in terms of performance and computational efficiency. Future studies are recommended to combine spectral and prosodic features and test the developed hybrid approaches on different datasets.

Ethics committee approval and conflict of interest statement

There is no need to obtain permission from the ethics committee for the article prepared.

There is no conflict of interest with any person / institution in the article prepared.

Authors' Contributions

Donuk K: Study concept and design; data analysis and interpretation; preparation of the manuscript draft; critical review and revision.

References

[1] J. Ye et al., "Multi-modal depression detection based on emotional audio and evaluation text", J. Affect. Disord., vol. 295, pp. 904-913, 2021.

[2] F. Yin, J. Du, X. Xu, and L. Zhao, "Depression Detection in Speech Using Transformer and Parallel Convolutional Neural Networks", Electronics 2023, vol. 12, no. 2, p. 328, 2023.

[3] E. W. McGinnis et al., “Giving Voice to Vulnerable Children: Machine Learning Analysis of Speech Detects Anxiety and Depression in Early Childhood”, IEEE J. Biomed. Health Inform., vol. 23, no. 6, pp. 2294-2301, 2019.

[4] W. Xian, “Speech Emotion Recognition Application for Education”, BCP Education & Psychology, vol. 7, pp. 378-383, 2022.

[5] M. Płaza et al., “Emotion Recognition Method for Call/Contact Centre Systems”, Applied Sciences 2022, vol. 12, no. 21, p. 10951, 2022.

[6] Y. Feng and L. Devillers, “End-to-End Continuous Speech Emotion Recognition in Real-life Customer Service Call Center Conversations”, 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2023, 2023.

[7] I. Gurowiec and N. Nissim, “Speech emotion recognition systems and their security aspects”, Artif. Intell. Rev., vol. 57, no. 6, pp. 1-45, 2024.

[8] Z. Jia, J. Yu, H. Long, and D. Tang, “Coverage-Guaranteed Speech Emotion Recognition via Calibrated Uncertainty-Adaptive Prediction Sets”, Accessed: Jul. 29, 2025. [Online]. Available: https://arxiv.org/abs/2503.22712v4

[9] R. Mishra, A. Frye, M. M. Rayguru, and D. O. Popa, “Personalized Speech Emotion Recognition in Human-Robot Interaction using Vision Transformers”, Accessed: Jul. 29, 2025. [Online]. Available: http://arxiv.org/abs/2409.10687

[10] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, “On the Robustness of Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks”, IEEE International Conference on Intelligent Robots and Systems, pp. 854-860, 2018.

[11] L. Matsouliadis, E. Siamtanidou, N. Vryzas, and C. Dimoulas, “Speech Emotion Recognition and Serious Games: An Entertaining Approach for Crowdsourcing Annotated Samples”, Information, vol. 16, no. 3, p. 238, 2025.

[12] S. Padi, S. O. Sadjadi, D. Manocha, and R. D. Sriram, “Multimodal Emotion Recognition using Transfer Learning from Speaker Recognition and BERT-based models”, pp. 407-414, 2022.

[13] M. Jain et al., “Speech Emotion Recognition using Support Vector Machine”, Lecture Notes in Networks and Systems, vol. 1003 LNNS, pp. 519-532, 2020.

[14] S. Yan, L. Ye, S. Han, T. Han, Y. Li, and E. Alasaarela, “Speech Interactive Emotion Recognition System Based on Random Forest”, 2020 International Wireless Communications and Mobile Computing, IWCMC 2020, pp. 1458-1462, 2020.

[15] M. Venkata Subbarao, S. K. Terlapu, N. Geethika, and K. D. Harika, “Speech Emotion Recognition Using K-Nearest Neighbor Classifiers”, pp. 123-131, 2022.

[16] I. Shahin, A. B. Nassif, and S. Hamsa, “Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network”, IEEE Access, vol. 7, pp. 26777-26787, 2019.

[17] S. Mao, D. Tao, G. Zhang, P. C. Ching, and T. Lee, “Revisiting Hidden Markov Models for Speech Emotion Recognition”, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6715-6719, 2019.

[18] R. Kingeski, E. Henning, and A. S. Paterno, “Fusion of PCA and ICA in Statistical Subset Analysis for Speech Emotion Recognition”, Sensors 2024, vol. 24, no. 17, p. 5704, 2024.

[19] J. T. Chien and B. C. Chen, “A new independent component analysis for speech recognition and separation”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, pp. 1245-1254, Jul. 2006.

[20] P. Tiwari, H. Rathod, S. Thakkar, and A. D. Darji, “Multimodal emotion recognition using SDA-LDA algorithm in video clips”, Journal of Ambient Intelligence and Humanized Computing, vol. 14, pp. 6585-6602, 2023.

[21] J. Zhao, X. Mao, and L. Chen, “Speech emotion recognition using deep 1D & 2D CNN LSTM networks”, Biomed. Signal Process. Control, vol. 47, pp. 312-323, 2019.

[22] C. Sun, H. Li, and L. Ma, “Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network”, Front. Psychol., vol. 13, p. 1075624, 2023.

[23] K. Deshpande, M. S. Sodhi, N. Raniyer, and M. Rao, “A Time-Distributed CNN-LSTM with Attention Model for Speech Based Emotion Recognition”, DMIP 2024 - Proceedings of 2024 7th International Conference on Digital Medicine and Image Processing, pp. 67-71, 2025.

[24] H. Kang, Y. Xu, G. Jin, J. Wang, and B. Miao, “FCAN : Speech emotion recognition network based on focused contrastive learning”, Biomed. Signal Process. Control, vol. 96, p. 106545, 2024.

[25] A. Vaswani et al., “Attention Is All You Need”, NIPS'17: Proceedings of the 31st International

Conference on Neural Information Processing Systems, p. 6000-6010, 2017.

[26] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 4171-4186, 2019.

[27] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations", NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, p. 12449-12460, 2020.

[28] W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units", IEEE/ACM Trans. Audio Speech Lang. Process., vol. 29, pp. 3451-3460, 2021.

[29] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation", Proc. Workshop on Speech, Music and Mind (SMM 2018), pp. 21–25, 2018.

[30] F. Makhmudov, A. Kutlimuratov, and Y. I. Cho, "Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition", Applied Sciences 2024, vol. 14, no. 23, p. 11342, 2024.

[31] S. S. Poorna, V. Menon, and S. Gopalan, "Hybrid CNN-BiLSTM architecture with multiple attention mechanisms to enhance speech emotion recognition", Biomed. Signal Process. Control, vol. 100, p. 106967, 2025.

[32] S. Mishra, N. Bhatnagar, P. Prakasam, and S. T. R, "Speech emotion recognition and classification using hybrid deep CNN and BiLSTM model", Multimed. Tools Appl., vol. 83, no. 13, pp. 37603-37620, 2024.

[33] L. T. Van, T. T. Le Dao, T. Le Xuan, and E. Castelli, "Emotional Speech Recognition Using Deep Neural Networks", Sensors (Basel)., vol. 22, no. 4, p. 1414, 2022.

[34] A. P. Gopi, P. Deepika, G. A. Gayathri, C. Keerthana, and B. V. Mounika, "Building Emotion Identification System from Speech Using CNN-GRU Model", Advances in Science, Technology and Innovation, pp. 53-59, 2025.

[35] R. Hasan, M. Nigar, N. Mamun, and S. Paul, "EmoFormer: A Text-Independent Speech Emotion Recognition using a Hybrid Transformer-CNN model", 2024 27th International Conference on Computer and Information Technology (ICCIT), pp. 2881-2886, 2024.

[36] S. Akinpelu, S. Viriri, and A. Adegun, "An enhanced speech emotion recognition using vision transformer", Scientific Reports 2024, vol. 14, 13126, 2024.

[37] R. Ullah et al., "Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer", Sensors (Basel)., vol. 23, no. 13, p. 6212, 2023.

[38] X. Tang, J. Huang, Y. Lin, T. Dang, and J. Cheng, "Speech Emotion Recognition via CNN-Transformer and multidimensional attention mechanism", Speech Commun., vol. 171, p. 103242, 2025.

[39] Y. Xia and L. Zhao, "CNN-BLSTM with Attention Model for Speech Emotion Recognition", Oct. 2023, doi: 10.21203/RS.3.RS-3392008/V1.

[40] N. K. Kundu, S. Kobir, Md. R. Ahmed, T. Aktar, and N. Roy, "Enhanced Speech Emotion Recognition with Efficient Channel Attention Guided Deep CNN-BiLSTM Framework", 2024, Accessed: Jul. 21, 2025. [Online]. Available: http://arxiv.org/abs/2412.10011

[41] C. Xu, Y. Liu, W. Song, Z. Liang, and X. Chen, "A New Network Structure for Speech Emotion Recognition Research", Sensors 2024, vol. 24, no. 5, p. 1429, 2024.

[42] S. Lee, D. K. Han, and H. Ko, "Fusion-ConvBERT: Parallel Convolution and BERT Fusion for Speech Emotion Recognition", Sensors (Basel)., vol. 20, no. 22, p. 6688, 2020.

[43] B. Maji, M. Swain, and Mustaqeem, "Advanced Fusion-Based Speech Emotion Recognition System Using a Dual-Attention Mechanism with Conv-Caps and Bi-GRU Features", Electronics 2022, vol. 11, no. 9, p. 1328, 2022.

[44] C. Barhoumi and Y. BenAyed, "Real-time speech emotion recognition using deep learning and data augmentation", Artif. Intell. Rev., vol. 58, 49, 2025.

[45] GitHub- rcantini/speech_emotion_recognition: How to detect emotions from speech using Bi-directional LSTM networks and attention mechanism in Keras." Accessed: Aug. 11, 2025. [Online]. Available: https://github.com/rcantini/speech_emotion_recognition.

[46] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional

speech", 9th European Conference on Speech Communication and Technology, pp. 1517-1520, 2005.

[47] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. Acoust., vol. 28, no. 4, pp. 357-366, 1980.

[48] K. DONUK, "KONUŞMA DUYGU TANIMA İÇİN ÖZELLİK ANALİZİ", Yapay Zekâ Tabanlı Sistemler: Teori, Uygulama ve Gelecek Perspektifleri-2, BİDGE Yayınları , Dec. 2025, 213-235.

[49] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete Cosine Transform", IEEE Transactions on Computers, vol. C–23, no. 1, pp. 90-93, 1974.

[50] J. Kennedy and R. Eberhart, "Particle swarm optimization", Proceedings of ICNN'95 - International Conference on Neural Networks, vol. 4, pp. 1942-1948, 1995.

[51] K. Donuk, N. Özbey, M. Inan, C. Yeroğlu, and D. Hanbay, "Investigation of PIDA Controller Parameters via PSO Algorithm", 2018 International Conference on Artificial Intelligence and Data Processing, IDAP 2018, 2018.

[52] K. DONUK, A. ARI, M. F. ÖZDEMİR, and D. HANBAY, "Deep Feature Selection for Facial Emotion Recognition Based on BPSO and SVM", Politeknik Dergisi, vol. 26, no. 1, pp. 131-142, 2023.

[53] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing", Science (1979)., vol. 220, no. 4598, pp. 671-680, 1983.

[54] D. E. Goldberg and J. H. Holland, "Genetic Algorithms and Machine Learning", Mach. Learn., vol. 3, no. 2, pp. 95-99, 1988.

[55] J. H. Holland, "Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence", Adaptation in Natural and Artificial Systems, Apr. 1992.

[56] S. Mirjalili, "Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems", Neural Comput. Appl., vol. 27, no. 4, pp. 1053-1073, 2016.

[57] D. Karaboga and B. Basturk, "Artificial Bee Colony (ABC) optimization algorithm for solving constrained optimization problems", Lecture Notes in Computer Science, vol. 4529, pp. 789-798, 2007.

[58] M. Dorigo, V. Maniezzo, and A. Colorni, "Ant system: Optimization by a colony of cooperating agents", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 26, no. 1, pp. 29-41, 1996.

[59] C. Cortes and V. Vapnik, "Support-vector networks", Machine Learning 1995, vol. 20, no. 3, pp. 273-297, 1995.

[60] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification", IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21-27, 1967.

[61] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors", Nature 1986, vol. 323, no. 6088, pp. 533-536, 1986.

[62] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 11531-11539, 2019.

[63] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks", IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 8, pp. 2011-2023, 2017.

[64] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2014.

[65] K. Donuk et al., "Konuşma Duygu Tanıma için Akustik Özelliklere Dayalı LSTM Tabanlı Bir Yaklaşım", Bilgisayar Bilimleri, vol. 7, no. 2, pp. 54-67, Dec. 2022.

[66] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks" IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, 1997.

[67] P. Singh, M. Sahidullah, and G. Saha, "Modulation spectral features for speech emotion recognition using deep neural networks", Speech Commun., vol. 146, pp. 53-69, 2023.

[68] K. M. Roh, S. P. Lee, "Enhanced Speech Emotion Recognition Using Conditional-DCGAN-Based Data Augmentation", Applied Sciences 2024, vol. 14, no. 21, p. 9890, 2024.

[69] Z. T. Liu, A. Rehman, M. Wu, W. H. Cao, and M. Hao, "Speech emotion recognition based on formant characteristics feature extraction and phoneme type

convergence", Inf. Sci. (N Y)., vol. 563, pp. 309-325, 2021.

[70] H. Li, X. Zhang, S. Duan, and H. Liang, "Speech emotion recognition based on bi-directional acoustic–articulatory conversion", Knowl. Based. Syst., vol. 299, 112123, 2024.

[71] J. Ancilin and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient", Applied Acoustics, vol. 179, 108046, 2021.

[72] M. Hou, J. Li, and G. Lu, "A supervised non-negative matrix factorization model for speech emotion recognition", Speech Commun., vol. 124, pp. 13-20, 2020.

[73] S. P. Mishra, P. Warule, and S. Deb, "Deep Learning Based Emotion Classification Using Mel Frequency Magnitude Coefficient", 1st IEEE International Conference on Innovations in High Speed Communication and Signal Processing, IHCSP 2023, pp. 93-98, 2023.

[74] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks", Biomed. Signal Process. Control, vol. 59, 101894, 2020.

[75] S. P. Mishra, P. Warule, and S. Deb, "Variational mode decomposition based acoustic and entropy features for speech emotion recognition", Applied Acoustics, vol. 212, 109578, 2023.