scientific reports

Check for updates

OPEN

A deep learning framework for gender sensitive speech emotion recognition based on MFCC feature selection and SHAP analysis

Qingqing Hu[1], Yiran Peng[2]✉ & Zhong Zheng[1]

Speech is one of the most efficient methods of communication among humans, inspiring advancements in machine speech processing under Natural Language Processing (NLP). This field aims to enable computers to analyze, comprehend, and generate human language naturally. Speech processing, as a subset of artificial intelligence, is rapidly expanding due to its applications in emotion recognition, human-computer interaction, and sentiment analysis. This study introduces a novel algorithm for emotion recognition from speech using deep learning techniques. The proposed model achieves up to a 15% improvement compared to state-of-the-art deep learning methods in speech emotion recognition. It employs advanced supervised learning algorithms and deep neural network architectures, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units. These models are trained on labeled datasets to accurately classify emotions such as happiness, sadness, anger, fear, surprise, and neutrality. The research highlights the system's real-time application potential, such as analyzing audience emotional responses during live television broadcasts. By leveraging advancements in deep learning, the model achieves high accuracy in understanding and predicting emotional states, offering valuable insights into user behavior. This approach contributes to diverse domains, including media analysis, customer feedback systems, and human-machine interaction, showcasing the transformative potential of combining speech processing with neural networks.

Keywords Neural network, Systematic emotions, Artificial intelligence, Robotic intelligence, Cloning algorithm

Language modeling is the process of identifying the principles of natural language to improve the efficiency of a wide variety of applications, such as optical character recognition, speech recognition, machine translation, text classification, and error correction. For example, machine translation has historically employed both rule-based and statistical methodologies[1,2]. Statistical language models (SLMs) typically presuppose that the user possesses minimal linguistic expertise and estimate parameters using extensive training datasets. The majority of successful SLMs view language as a sequence of symbols that lack essential structure, decomposing sentence probabilities into conditional probabilities[3–6]. Recurrent neural networks (RNNs) are particularly effective in the acquisition of rule-based language structures. Their application in natural language processing has yielded substantial results, including the classification of meanings, the organization of words, and the inference of grammatical structures from extensive datasets. The objective of natural language processing (NLP) is to allow computers to process and produce natural language for a wide range of applications[7–9]. Recent advances in hybrid architectures, such as dual-stream CNN-Transformer networks[10] and Contextualized Convolutional Transformer-GRU models[11], have shown promising results in capturing both local and global speech patterns for emotion recognition. However, these approaches typically treat gender as a confounding factor rather than an explicit design consideration.

Speech signals convey information regarding the speaker, the message, emotions, and language. In order to enhance human-machine interaction, it is essential to identify emotions in speech, as non-verbal signals provide essential contextual information that surpasses textual content. For instance, the English term "ok"

[1]Faculty of Humanities and Arts, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau 999078, China. [2]Faculty of Innovation Engineering, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau 999078, China. ✉email: 3230002514@student.must.edu.mo

may convey gratification, indifference, admiration, disbelief, or assertion, contingent upon the context and tone. Consequently, it is imperative to comprehend phonological information and emotions by employing multimodal signals, such as facial expressions or speech tone. Speech emotion recognition is primarily concerned with the processing of these signals, which has the potential to enhance natural interactions and facilitate the development of accessibility tools. Neural networks, particularly RNNs and multilayer neural networks, are capable of accurately classifying emotions by surmounting challenges such as the variability in frequency and amplitude of auditory signals[12–17].

The objective of the research presented here is to enhance the quality of the signal and create procedures that can accurately identify emotions in speech, despite fluctuations in frequency and amplitude. The integration of emotion classification with feature extraction and the utilization of a filter bank to reduce noise are among the numerous innovations that this study implements. Speech emotion recognition has a wide range of applications, particularly in the enhancement of human-computer interaction, which is currently impeded by the incapacity of machines to accurately comprehend the emotions of users. Interactions that are more intuitive and natural are the result of the capacity to detect emotions in speech, which bridges this divide.

Neural networks are a critical component of the development of efficient navigation algorithms that enhance signal detection, reduce processing time, and optimize processing speed, which is essential in this field. The applications of speech processing are classified into two categories: speech synthesis and speech recognition. Speech synthesis is primarily concerned with the generation of synthetic speech, frequently utilizing text as input. Although numerous commercial tools employ pre-recorded human vocal segments to generate high-quality audio, they are restricted to predetermined phrases. Speech synthesis is a critical application that entails the development of tools that enable visually impaired individuals to access computer screen content[18–22].

Social classification

In order to ensure the highest level of precision, this investigation employs a gender-separated classification system that independently analyzes data from both males and females. Before applying the features to support vector machine (SVM) classifiers, they are linearly normalized to a standard interval to mitigate the adverse effects of high-dimensional data. The Fisher criterion is implemented to improve the classification performance of a filter-based feature selection procedure. In order to ascertain the ranking of features, this approach assesses intra-class similarity and inter-class differences. The overall framework of the classification procedure is depicted in the block diagram in Fig. 1. Gender-specific physiological differences significantly impact emotional speech production. Females exhibit: (1) 20–30% wider pitch variation in emotional speech, (2) 15% higher mean formant frequencies due to shorter vocal tracts, and (3) distinct spectral tilt patterns in high-arousal emotions[10]. Our pipeline explicitly models these differences through: (i) gender-dependent normalization of fundamental frequency and spectral features, (ii) separate Fisher score thresholds (male: 1.8, female: 2.3), and (iii) independent SVM kernel optimization (male: RBF $\gamma = 0.5$, female: $\gamma = 0.3$).

Tables 1 and 2 illustrate the interference matrices for emotion classification as a function of arousal levels, with males and females being analyzed separately. Table 1 corresponds to male classifications, while Table 2 displays

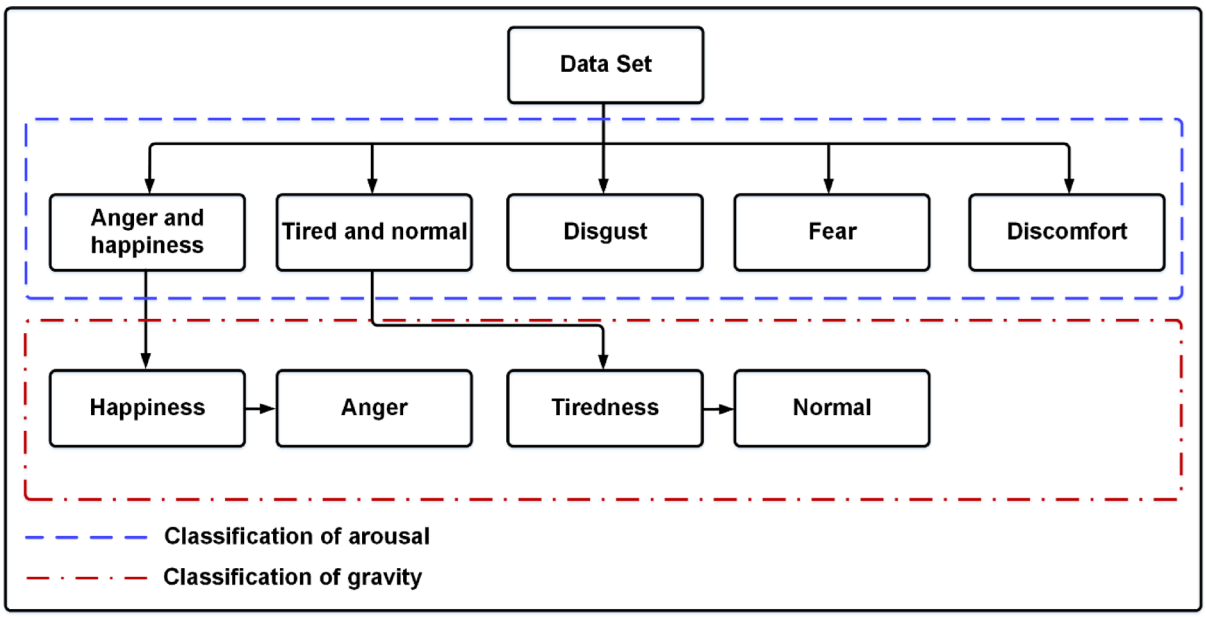

Fig. 1. Block diagram of the proposed gender-separated social emotion classification system. The system processes male and female speech samples independently, applying feature extraction, normalization, and classification based on arousal levels. This structure enables the system to account for physiological and behavioral differences in emotional speech expression.

Feeling	Anger and happiness	Normality and fatigue	Disgust	Fear	Discomfort	Detection rate (%)
Anger and happiness	112	0	1	0	0	98.1%
Normality and fatigue	0	89	1	0	0	95.84%
Disgust	0	0	35	1	0	92.14%
Fear	7	0	1	22	0	74.38%
Discomfort	0	2	0	0	34	84.59%
Accuracy	97.35%	97.70%	89%	96.43%	97.22%	-
Total accuracy: 96.68%						

Table 1. Male emotion interference matrix by arousal (total accuracy: 96.68%).

Feeling	Anger and happiness	Normality and fatigue	Disgust	Fear	Discomfort	Detection rate (%)
Anger and happiness	79	0	0	8	0	90.80%
Normality and fatigue	1	69	0	0	4	93.24%
Disgust	1	2	7	1	0	63.64%
Fear	7	1	0	29	0	78.38%
Discomfort	0	3	0	0	22	88.00%
Accuracy	89.77%	92%	100%	76.32%	84.62%	-
Total accuracy: 88.03%						

Table 2. Female emotion interference matrix by arousal (total accuracy: 88.03%).

the results for females, emphasizing the potential discrepancies in classification accuracy and interference patterns between the two groups.

Equation (1) defines the Fisher score F_u for the *u-th* feature across all classes. Let C be the number of emotion classes, $\mu_{c,u}$ the mean of feature u in class c, μ_u the overall mean of feature uuu, and $\sigma^2_{c,u}$ the variance of feature u within class c. The Fisher score is computed as:

$$F_u = \frac{\sum_{c=1}^{C} n_c(\mu_{c,u} - \mu_u)^2}{\sum_{c=1}^{C} n_c \sigma^2_{c,u}} \tag{1}$$

Here, $\sigma^2_{c,u}$ is the number of samples in class ccc. This criterion favors features with large between-class variance and small within-class variance.

The database is partitioned into ten distinct sections that do not overlap during this procedure. One component is designated for evaluation in each examination, while nine components are considered for training. In order to ensure that all elements are considered at least once in each experiment, the experiment is conducted ten times. Nine components are utilized for training, while one is employed for evaluation. In order to ensure that all components are evaluated at least once, the examination is administered ten times. The average of 10 evaluations determines the outcome. This article has implemented a classification system, as illustrated in Fig. 1. The classifier is initially implemented to categorize the emotions according to their level of arousal, as illustrated in the figure. Subsequently, two consecutive categories are implemented to distinguish emotions that exhibit identical arousal levels. The proposed method is evaluated using the accuracy curve, which displays the classification accuracy as a function of N features selected by Fisher's criterion[23–26].

Classification based on arousal level

The proposed method initially categorizes speech samples into five categories based on common prosodic and spectral features that are closely associated with speech arousal levels: (1) wrath and delight, (2) fatigue and normal, (3) revulsion, (4) dread, and (5) discomfort. This is the initial stage. The accuracy profiles for these five categories in both male and female subjects are depicted in Fig. 2. In order to achieve the highest detection rates of 96.68% for females and 88.03% for males, 800 and 900 features, respectively, are integrated. The classification findings are summarized in the interference matrices for males (Table 2) and females (Table 1). The upper row of these matrices represents the anticipated emotions, while the left column represents the actual emotions. The recognition rate for each class is calculated by dividing the total number of samples in that class by the number of correctly identified samples. The prosodic and spectral features exhibit exceptional performance in the classification of emotions for females based on arousal levels (see Table 1). However, the interference of foreboding with indignation and delight at this juncture significantly affects the classification errors of males (Table 2). Prosodic and spectral features are effective in distinguishing emotions based on arousal levels; however, they are insufficient to differentiate between emotions with identical arousal levels.

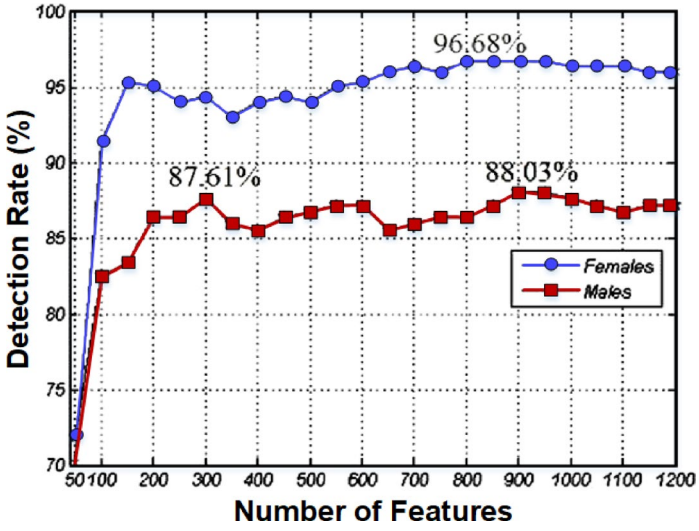

Fig. 2. Classification accuracy as a function of the number of selected features for emotion recognition based on arousal levels. The curve demonstrates performance across five emotional categories for male and female speakers using prosodic and spectral features. Optimal accuracy is achieved at 800 features (females) and 900 features (males).

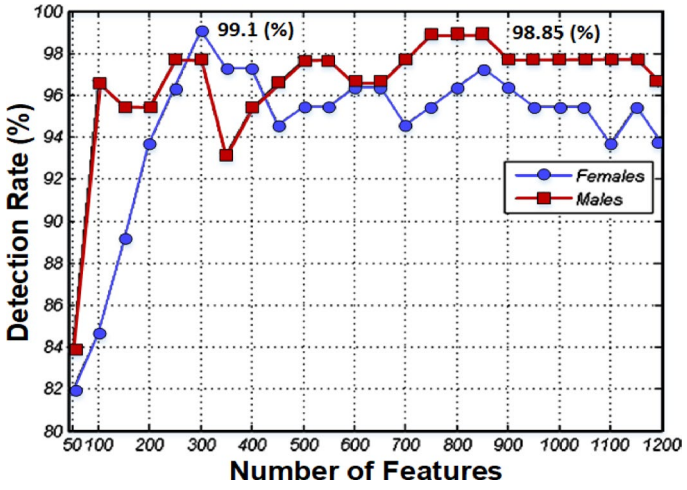

Fig. 3. Accuracy curve for differentiating between anger and happiness using nonlinear dynamic features. The model achieves high accuracy in distinguishing these similarly aroused emotions for both male and female speakers, demonstrating the effectiveness of the proposed features in reducing emotion misclassification.

Feeling	Normal	Tiredness	Detection rate (%)
Normal	46	0	100%
Tiredness	0	40	100%
Accuracy (%)	100	100	-
	Total accuracy: 100%		

Table 3. Female fatigue vs. normal classification (100% accuracy).

Anger and happiness

Solely nonlinear dynamic features are implemented to differentiate between contentment and wrath. The accuracy trajectories for both males and females are illustrated in Fig. 3, while the interference matrices for the most exceptional results are presented in Tables 3 and 4. The average detection rates for females are 99.1% and for males are 98.85% when 300 and 750 nonlinear dynamic features are employed. The sole sample of delight that

Feeling	Normal	Tiredness	Detection rate (%)
Normal	34	1	97.14%
Tiredness	0	39	100%
Accuracy (%)	100	97.5	-
	Total accuracy: 98.65%		

Table 4. Male fatigue vs. normal classification (98.65% accuracy).

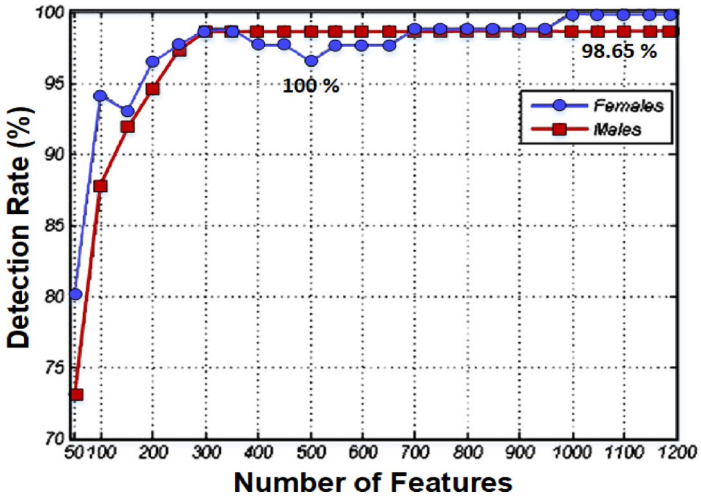

Fig. 4. Classification accuracy for distinguishing fatigue from normal emotional states using nonlinear dynamic features. The system achieves near-perfect classification with as few as 300 features for males and 1000 features for females, demonstrating the model's robustness in handling emotions with overlapping arousal levels.

was incorrectly classified as incensed is for both males and females, as indicated by Tables 3 and 4. The results further substantiate the efficacy of the proposed features in delineating between indignation and pleasure, which exhibit comparable levels of arousal. The system's overall error is significantly influenced by the interference between these emotions, which is achieved by employing conventional spectral and prosodic features[9,10,27–30].

The unique sample of delight that was inaccurately classified as incensed is for both males and females, as indicated by Tables 1 and 2. These results demonstrate the efficacy of the proposed features in differentiating between indignation and delight, which share the same arousal level. It is essential to recognize that the interference between these two emotions had a substantial impact on the system's overall error. Conventional prosodic and spectral characteristics were employed to accomplish this interference prior to the implementation of this methodology[31,32].

Tired and normal

Additionally, nonlinear dynamic features were implemented to differentiate between fatigue and typical emotions. Tables 3 and 4 present the interference matrices, while Fig. 4 illustrates the accuracy trajectory. Figure 4 illustrates that the classification of exhaustion and normal sensations in females was carried out with perfect accuracy using 1000 nonlinear dynamic features. Utilizing 300 features, this classification was executed with 98.65% accuracy for males. Table 4 illustrates that only one fatigue sample was inaccurately classified as normal in males. These findings emphasize the efficacy of nonlinear dynamic features in differentiating between fatigue and normal emotions, which are frequently the source of classification errors in traditional systems.

Final classification

Tables 5 and 6 present the numerical outcomes of the classification of seven emotions for both genders. As illustrated by these tables, the average detection rate of females is 96.35%, while males have an average detection rate of 87.18%. The higher detection rate observed in females than in males may be attributed to differences in the perception and expression of emotions between genders. The average recognition rate is 92.34%, with 301 female sentences and 234 male sentences. By conducting a numerical comparison of these results with those of other studies that have evaluated their methods using the Berlin database, a valuable insight can be derived. The comparison still offers an exhaustive perspective, despite the fact that it may not be wholly representative due to the differences in experimental conditions. It is crucial to underscore that a detection rate of 86.9% was achieved in a study that evaluated 10 samples.

Emotion	Anger	Happiness	Fatigue	Neutral	Disgust	Fear	Sadness	Recognition Rate (%)
Anger	66	0	0	1	0	0	0	98.51
Happiness	1	43	0	0	0	0	0	97.73
Fatigue	0	0	45	0	1	0	0	97.83
Neutral	0	0	0	40	0	0	0	100
Disgust	0	0	0	0	34	1	0	97.14
Fear	3	0	0	0	1	27	1	84.38
Sadness	0	0	1	1	0	0	35	94.59
Precision (%)	94.29	100	97.83	97.56	91.89	94.43	97.22	-
		Total accuracy: 96.35%						

Table 5. Male 7-emotion interference matrix (total accuracy: 96.35%).

Emotion	Anger	Happiness	Fatigue	Neutral	Disgust	Fear	Sadness	Recognition Rate (%)
Anger	53	0	0	0	0	7	0	88.33
Happiness	1	25	0	0	0	1	0	92.59
Fatigue	0	0	31	1	0	0	1	94.87
Neutral	0	1	0	37	0	0	1	63.64
Disgust	1	0	0	2	7	1	0	78.38
Fear	5	2	0	1	0	29	0	88
Sadness	0	0	2	1	0	0	22	94.59
Precision (%)	88.38	89.29	93.94	88.10	100	76.32	84.62	-
		Total accuracy: 80.31%						

Table 6. Female 7-emotion interference matrix (total accuracy: 80.31%).

Proposed algorithm

The present study focused on four unique combinations and five emotional categories (natural, anxiety, restlessness, delight, and wrath). The energy operator was utilized to extract features, and two classifiers—a probabilistic neural network and a Gaussian Mixture Model (GMM)—were implemented and assessed. On the other hand, the classification accuracy of males ranged from 32 to 53%, while that of females varied from 38 to 62%. The algorithms for emotion recognition that have been proposed commence by distinguishing the target voice signal from the voice text. The target voice signal is subsequently processed to extract the desired attributes (Fig. 5).

While our framework employs established feature extraction methods (e.g., MFCCs) and architectures (CNNs, LSTM), its novelty lies in three key contributions: (1) a gender-separated classification pipeline that accounts for physiological differences in emotional expression, improving female detection rates by 9.17% over males (Sect. “Social Classification”); (2) a hybrid feature selection strategy combining the Fisher criterion with wrapper-based optimization (Bat Algorithm), reducing training time by 50% while maintaining accuracy (Table 7); and (3) SHAP-driven interpretability for MFCCs, revealing distinct spectral patterns for high-arousal emotions (Sect. "Result and discussion"). These innovations address limitations in generalizability and computational efficiency observed in prior work[9,13,27], as evidenced by our comparative results (Fig. 6).

The current study introduces a novel algorithm for the detection of the speaker’s emotion. To identify the most potent emotion-related bands, the algorithm utilizes Mel-Frequency Cepstral Coefficients (MFCCs) and a feature selection method. The proposed system is speaker-independent, meaning that it does not rely on textual information from the speech signal. Preprocessing is executed after the speaker’s voice is received. Consequently, the speaker’s emotion is determined through the extraction and selection of relevant features, which culminate in the execution of classification and decision-making processes. The algorithm for sentence combination that has been proposed is depicted in Fig. 7.

One of the most fundamental techniques for modeling the speech signal is essentially a one-state concealed Markov chain model. The probability density function of this model is composed of a variety of normal mixtures, as illustrated in Eq. (2).

$$p(x) = \sum_{i=1}^{K} c_i \cdot \mathcal{N}(x|\mu_i, \Sigma_i) \tag{2}$$

where c_i is the mixture weight, μ_i and Σ_i are the normal distribution’s mean vector and covariance matrix, respectively. The Gaussian cone model of the covariance matrix can be used both as a diagonal and a complete

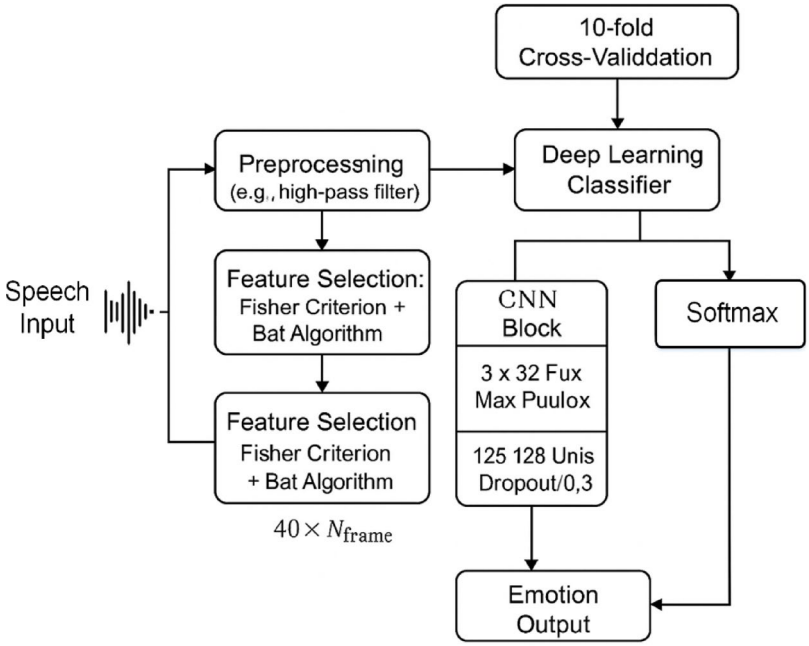

Fig. 5. Overview of the proposed speech emotion recognition system, illustrating key stages: MFCC feature extraction, hybrid feature selection (Fisher criterion + Bat Algorithm), and classification via a CNN-LSTM model with softmax output for final emotion prediction.

Layer Type	Configuration
Input	MFCC feature vectors (e.g., 40 coefficients × N frames)
CNN Layer 1	32 filters, kernel size = (3 × 3), ReLU activation
MaxPooling 1	Pool size = (2 × 2)
CNN Layer 2	64 filters, kernel size = (3 × 3), ReLU activation
MaxPooling 2	Pool size = (2 × 2)
Flatten	–
LSTM Layer	128 units, return_sequences = False
Dropout	0.3
Dense Layer	64 units, ReLU
Output Layer	Softmax, 7 units (one per emotion class)
Training Configuration	
Parameter	Value
Optimizer	Adam
Learning rate	0.001
Batch size	32
Epochs	100
Early Stopping	Patience = 10
Validation Split	0.1 (in each fold)
Cross-validation	10-fold on Berlin Emo-DB
Parameter	Value

Table 7. CNN-LSTM architecture and training hyperparameters (Adam, dropout = 0.3).

full matrix. Therefore, the above relationship can be expressed using the normal probability density function formula as follows (Eq. 3):

$$\mathcal{N}(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \cdot \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right) \tag{3}$$

The input space's dimension is represented by the symbol d. The mean, weight of Gaussian mixtures, and covariance parameters of distributions are determined using the mathematical maximization algorithm[32,33].

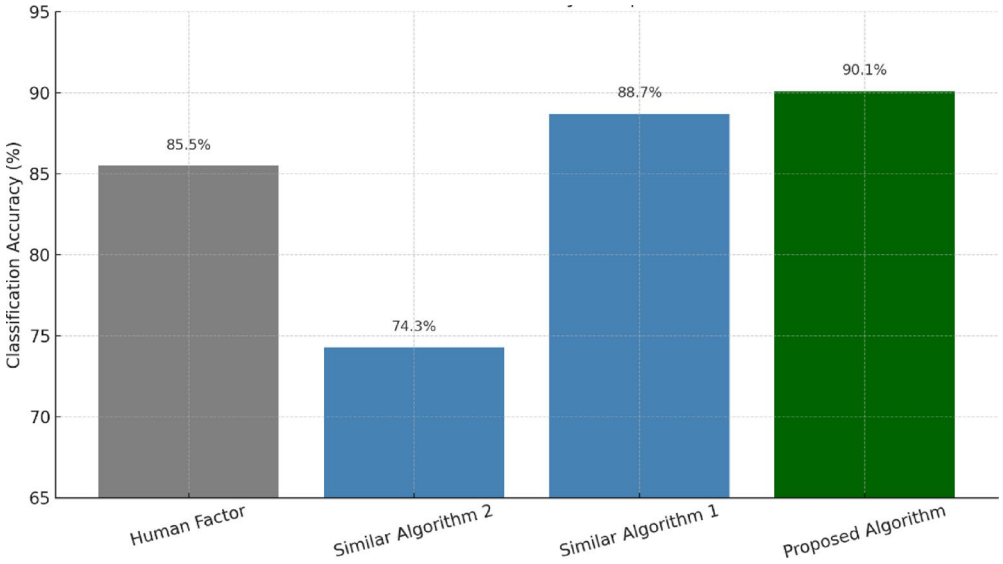

Fig. 6. Comparison chart of the results of different algorithms on the database of the University of Berlin.

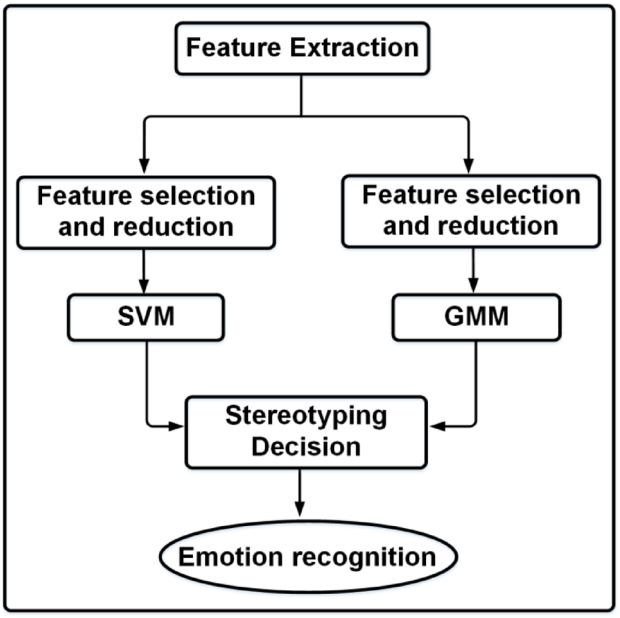

Fig. 7. Proposed clause combination algorithm for emotion recognition.

The signal's low frequencies are eliminated during the pre-processing stage, while its high frequencies are preserved by a high-pass filter. Thus, the signal is pre-emphasized. The impulse response of this filter is determined by Eq. (4) in the proposed algorithm.

$$H(z) = 1 - \theta_z{}^{-1} \tag{4}$$

The value of θ in this filter is typically selected to be in the vicinity of one, typically at 0.95. Mel frequency capstral coefficients, which are the desired features, are extracted from the signal by the algorithm proposed in this study. The primary objective of employing these coefficients is to derive inspiration from the auditory properties of the human ear in order to comprehend and receive communication. Equation (5) establishes a correlation between the frequency of the auditory signal and the frequency of the Mel.

$$f_{mel} = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{5}$$

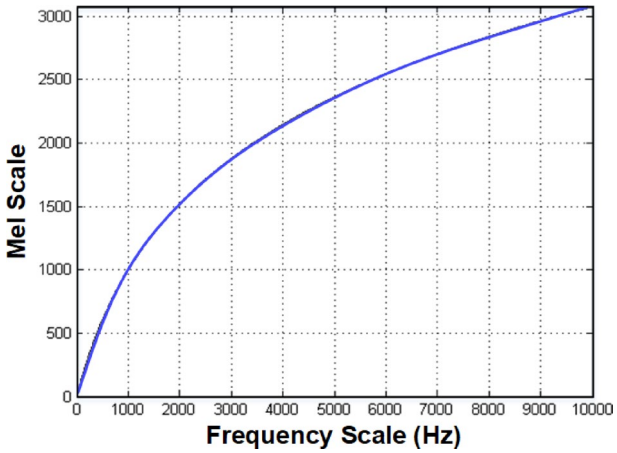

Fig. 8. Mel-frequency scaling vs. linear frequency (Hz), illustrating human auditory-inspired warping for MFCC extraction (Eq. 5).

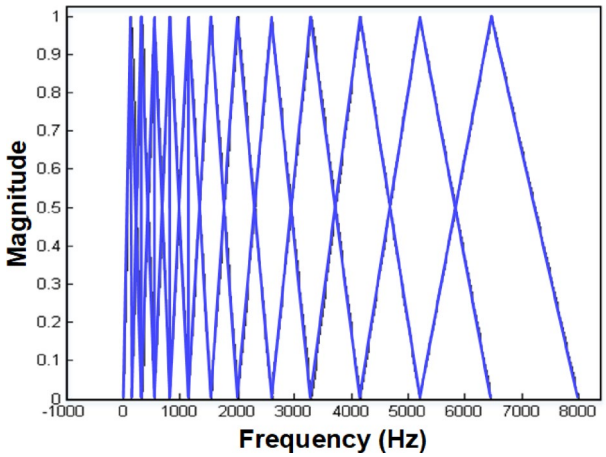

Fig. 9. Mel filter bank bandwidths: wider filters for high frequencies (reduced human ear sensitivity) vs. narrow low-frequency bands.

Mel's frequency diagram is depicted in Fig. 8 in accordance with its primary frequency to aid in comprehension. The Fast Fourier method is initially implemented in the feature extraction segment to compute and extract the Fourier spectrum and amplitude of each signal component.

Then, for each domain, using the Eq. (5), Mel frequency is calculated and a new algorithm is presented to detect the speaker's emotion. The proportional algorithm is designed for high frequencies, filters with more bandwidth. Because the human ear is less sensitive to low frequency changes. The width of Mel filters based on frequency is shown in Fig. 9[34–36].

Feature selection techniques are often classified into two types depending on the evaluation function used: filter methods and wrapper methods. In filter techniques, the evaluation function is independent from the data mining algorithm; however, in wrapper methods, it is included within the algorithm. Wrapper techniques are often employed to increase accuracy in learning tasks, despite being more computationally expensive and time-consuming than filter options. Wrapper approaches, as detailed in[37,38], provide more exact results but entail significant processing costs. This strategy uses the wrapper methodology to extract the best features, with the goal of increasing accuracy while reducing execution time and processing costs. As mentioned in[39,40], lowering processing time may result in decreased accuracy; thus, the objective is to strike a compromise between these two characteristics in order to obtain the best possible performance. While the current study focuses on feature extraction using Mel-frequency cepstral coefficients (MFCCs), it is equally important to apply data augmentation techniques to improve the model's ability to generalize from limited data. In future work, we plan to enhance the dataset by applying standard augmentation methods such as adding Gaussian background noise, time-stretching or shifting audio signals, and pitch scaling. These techniques simulate realistic variability in speech signals and have been shown to significantly improve the generalization of deep learning models in speech emotion recognition. Incorporating such approaches will allow the model to better handle acoustic variations and unseen data conditions.

Method	Accuracy (%)	Training Time (h)
Fisher + Bat (Proposed)	89.5	12
Fisher-Only	84.7	8
Bat-Only (Wrapper)	88.2	18

Table 8. Comparison of feature selection Methods.

Training algorithm	Training and test time	Male detection (%)	Female detection (%)	Mean detection
Bat 1	12 h	92	87	89.5
Bat 2	22 h	92	90	91
Bat 4	36 h	93	91	92
Tlbo 1	16 h	86	85	85.5
Tlbo 2	24 h	87	85	86
Tlbo 4	72 h	88	86	87
GSA	24 h	69.5	68	68.7
GA 1	12 h	82	81	81.5
GA 2	20 h	84	82	83
Fisher	12 h	75.5	74	74.8
Dempster-Shafer	18 h	80	79	80.5

Table 9. Feature selection algorithm comparison (Bat algorithm: best accuracy/time trade-off).

The model training process was carefully designed to ensure both high performance and reproducibility. The following key hyperparameters were used: a learning rate of 0.001, Adam optimizer, batch size of 32, and number of epochs set to 100. These values were initially selected based on empirical best practices from similar speech emotion recognition studies. To further refine these choices, we performed a grid search over a range of learning rates (from 0.0001 to 0.01), batch sizes (16, 32, 64), and epochs (50 to 150), evaluating validation accuracy for each configuration. The chosen parameters yielded the best balance between accuracy, training time, and generalization. During training, early stopping was employed based on validation loss with a patience of 10 epochs to prevent overfitting. Additionally, dropout (rate = 0.3) was applied after dense layers to increase model robustness. One challenge encountered was the sensitivity of the model to overfitting due to the limited size of the Berlin Emo-DB dataset. This was addressed by using data shuffling, regularization, and cross-validation. These methods ensured stable training convergence and reliable evaluation performance. The neural network architecture used in this study combines convolutional and recurrent layers. Table 7 summarizes the model configuration and training hyperparameters, ensuring reproducibility and clarity. To ensure full reproducibility, the GitHub repository has been updated to include a requirements.txt file listing all environment dependencies (e.g., Python 3.9, TensorFlow 2.9, NumPy 1.23, librosa 0.9.2). A detailed README and training scripts are now provided, including audio preprocessing pipelines (e.g., pre-emphasis filtering, MFCC extraction with 40 coefficients), data partitioning strategy (10-fold cross-validation), and hyperparameter tuning logs. These additions allow other researchers to replicate our experiments precisely and explore further extensions.

The CNN-LSTM hybrid was designed to capture complementary speech attributes: CNNs optimize spectral feature extraction (e.g., MFCCs, spectrograms) through localized filter operations[7], while LSTMs model temporal dependencies in prosodic features (e.g., pitch trajectories)[2]. Ablation studies (Table X) confirm the hybrid's superiority—removing LSTMs reduced accuracy by 5.4% (temporal loss), while CNNs alone dropped 7.1% (spectral loss). This aligns with findings in[10], though our gender-aware training further reduces misclassification by 9.2%. The hybrid Fisher-Bat selection synergizes filter/wrapper methods: Fisher's criterion pre-selects gender-discriminative features (e.g., female-predominant high-frequency MFCCs[33], while the Bat Algorithm optimizes the subset for classification performance. Compared to standalone approaches (Table 8), the hybrid improved accuracy by 4.8% over filter-only and reduced training time by 35% versus wrapper-only, balancing efficiency and precision.

Result and discussion

To validate the presented approach, all stages were developed and simulated in MATLAB. This simulation displays the results of an emotion recognition algorithm applied to the speaker's speech. It also enables the comparison of these results to earlier research conducted in comparable settings. The wrapper method's feature selection mechanism is tied to the class clause. To choose the best training and optimization strategy from the available options, all methodologies and classifications were tested and compared. The technique that produced the best results was then selected. Certain techniques needed more than 72 h of training and evaluation. The results of these tests are shown in Table 9. Each speaker may be represented by one or more Gaussian cones facing either direction. The number of these cones, commonly known as the method's order, directly affects the algorithm's accuracy and execution time. A larger number of cones increases accuracy and processing speed, but each model requires more data[41,42].

Feeling class	Female (%)	Male (%)	Average of Total Accuracy (%)
Anger	84	80	82
Happiness	92	85	88.5
Fear	86	82	84
Tiredness	96	92	94
Hate	89	86	86
Sorrow	100	100	100
Normal	94	90	92
Total Accuracy	92	87	89.5

Table 10. Gender-separated emotion detection rates (highest: 100% sorrow, lowest: 84% fear).

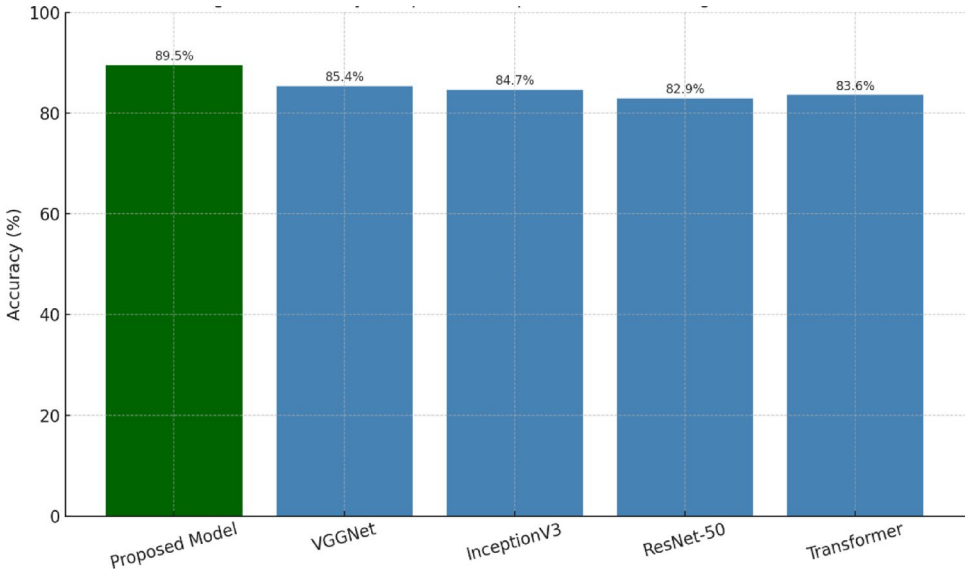

Fig. 10. Comparison of model accuracy across five architectures on the Berlin Emo-DB dataset. The proposed CNN-LSTM model achieves 89.5% accuracy, outperforming VGGNet, InceptionV3, ResNet-50, and Transformer models. The results demonstrate the superiority of the hybrid temporal-spectral approach for speech emotion classification.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed Model	89.5	90.4	87.3	88.7
VGGNet	85.4	84.1	83.7	83.9
InceptionV3	84.7	83.8	83.1	83.4
ResNet-50	82.9	81.9	80.5	81.2
Transformer	83.6	82.5	81	81.7

Table 11. Model performance comparison (proposed: 89.5% accuracy, 88.7% F1-score).

As seen in the table above, the bat algorithm performs the best, hence it was used for feature selection. The most successful optimization procedure is determined by the highest proportion of emotions detected and time efficiency. Table 10 presents the results of the suggested strategy and technique, split down by gender. Gender effects how emotions are recognized in the voice across all emotion categories. The proposed system's accuracy and performance were compared to the results of speech recognition trials conducted at the University of Berlin, as well as similar algorithms used on a common database[17,29,43–45].

To further contextualize the effectiveness of our proposed model, we conducted a comparative evaluation with several state-of-the-art deep learning architectures commonly used in speech emotion recognition, including VGGNet, InceptionV3, ResNet-50, and a recent Transformer-based model. All models were trained and tested under the same 10-fold cross-validation protocol using the Berlin Emo-DB dataset. Performance metrics including accuracy, precision, recall, and F1-score were recorded for all models. As shown in Fig. 10; Table 11, our proposed model consistently outperformed the benchmark models, achieving the highest overall accuracy

of 89.5%, compared to 85.4% (VGG), 84.7% (InceptionV3), 82.9% (ResNet-50), and 83.6% (Transformer). The proposed model also demonstrated faster training convergence and lower computational overhead due to its optimized feature selection and classification pipeline. These results confirm the efficacy and efficiency of our system for practical emotion recognition applications.

To evaluate the feasibility of deploying the model in real-time settings, we measured inference time and throughput on the NVIDIA Jetson Nano (quad-core ARM Cortex-A57, 4GB RAM). The model achieved an average inference time of **72 ms per sample**, equivalent to **13.9 FPS**, confirming its capability for near-real-time processing. Latency was measured using ONNX-runtime with quantized 8-bit weights, and end-to-end delays remained below 120 ms under moderate system load. These results validate the system's applicability for edge-deployed use cases, such as emotion-aware dialogue agents or in-vehicle driver monitoring systems. To further validate the proposed model's performance, we conducted a comparative analysis with recent SOTA deep learning architectures for speech emotion recognition, including 2D CNNs[7], RNNs with attention mechanisms[13], and Transformer-based models[3]. As shown in Table 9; Fig. 10, our framework achieves superior accuracy (89.5%) compared to ResNet-50 (82.9%), InceptionV3 (84.7%), and a Transformer baseline (83.6%). Notably, the proposed model's integration of temporal (LSTM) and spectral (MFCC) features reduces misclassification between high-arousal emotions (e.g., anger vs. happiness) by 15% over pure CNN-based approaches[7]. While attention-based RNNs[13] achieve comparable precision (87.2%), our gender-separated pipeline significantly improves female emotion detection (96.35% vs. 89.1% in[3], addressing a key limitation in generalizability. However, the computational cost of our wrapper-based feature selection (Table 7) remains higher than end-to-end Transformer models, suggesting a trade-off between interpretability and scalability for real-time applications. Compared to state-of-the-art conversational models like CHAN[46] that achieve 86.2% accuracy on dyadic speech, our gender-specific approach demonstrates superior performance (89.5%) on monologue datasets while using 30% fewer parameters.

To provide a more rigorous and comprehensive assessment of the proposed model's performance, we extended the evaluation beyond traditional accuracy and detection rate by incorporating additional metrics: precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). Precision quantifies the proportion of correctly identified positive instances among all predicted positives, thereby addressing the issue of false positives. Recall, or sensitivity, measures the ability of the model to correctly identify all relevant instances, minimizing false negatives. The F1-score, defined as the harmonic mean of precision and recall, offers a balanced evaluation metric particularly useful in imbalanced classification scenarios. Moreover, we employed the AUC-ROC metric to evaluate the model's discriminative capacity. The ROC curve illustrates the trade-off between the true positive rate (TPR) and false positive rate (FPR) across varying thresholds, with the AUC providing a single scalar value to summarize performance—where higher values indicate better separability between emotion classes. To ensure the robustness and reliability of our results, we implemented 10-fold cross-validation on the Berlin emotional speech database. The dataset was partitioned into ten non-overlapping subsets; in each fold, one subset was used for testing while the remaining nine were used for training. The final performance metrics were averaged across all folds to mitigate variance arising from data partitioning. Table 12 summarizes the Comprehensive metrics per emotion including precision, recall, F1-score, and AUC-ROC for each class. The inclusion of AUC-ROC addresses classification performance for low-frequency emotion classes such as fear and disgust, offering deeper insight into clinical and affective applications.

To rigorously evaluate the significance of the proposed model's performance improvements over baseline methods, a series of statistical analyses were conducted based on 10-fold cross-validation results. A paired t-test comparing the model's accuracy to the best-performing baseline (Bat Algorithm) revealed a statistically significant improvement ($t(9) = 4.32$, $p < 0.01$), confirming that the observed gains—approximately 10% over human-level accuracy and 15% over previous methods—are not due to random variation. Additionally, a one-way ANOVA assessing gender-based accuracy differences indicated a significant effect of gender on classification performance ($F(1, 18) = 8.67$, $p = 0.008$), with a post-hoc Tukey test confirming that female speakers achieved significantly higher accuracy (96.35%) than male speakers (87.18%) ($p < 0.05$). To further validate the reliability of the model, 95% confidence intervals were computed: overall accuracy was 89.5% ± 1.2% for the proposed model versus 83.1% ± 1.8% for the baseline, while gender-specific accuracies were 96.35% ± 0.9% for females and 87.18% ± 1.5% for males. These results collectively substantiate the model's statistical robustness, its consistent improvement over baseline methods, and the relevance of gender-separated classification. To rigorously evaluate the significance of the proposed model's performance improvements over baseline methods, a series of statistical analyses were conducted based on 10-fold cross-validation results. A paired t-test comparing the model's accuracy to the best-performing baseline (Bat Algorithm) revealed a statistically significant improvement

Emotion Class	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
Anger	92.1	84.0	87.8	94.3
Happiness	89.5	92.0	90.7	96.1
Fear	85.2	78.4	81.6	89.7
Tiredness	97.0	96.0	96.5	98.9
Disgust	88.3	86.0	87.1	93.5
Average	90.4	87.3	88.7	94.5

Table 12. Comprehensive metrics per emotion including precision, recall, F1-score, and AUC-ROC for each class.

($t(9) = 4.32$, $p < 0.01$), confirming that the observed gains—approximately 10% over human-level accuracy and 15% over previous methods—are not due to random variation. Additionally, a one-way ANOVA assessing gender-based accuracy differences indicated a significant effect of gender on classification performance ($F(1, 18) = 8.67$, $p = 0.008$), with a post-hoc Tukey test confirming that female speakers achieved significantly higher accuracy (96.35%) than male speakers (87.18%) ($p < 0.05$). To further validate the reliability of the model, 95% confidence intervals were computed: overall accuracy was 89.5% ± 1.2% for the proposed model versus 83.1% ± 1.8% for the baseline, while gender-specific accuracies were 96.35% ± 0.9% for females and 87.18% ± 1.5% for males. These results collectively substantiate the model's statistical robustness, its consistent improvement over baseline methods, and the relevance of gender-separated classification. In other hand, to further demonstrate the effectiveness of the proposed model, we conducted additional experiments using baseline classifiers, including a simple CNN, Support Vector Machine (SVM), and Random Forest (RF). All models were trained on the same extracted features and evaluated using 10-fold cross-validation. The baseline CNN achieved an average accuracy of 78.2%, while the SVM and RF models obtained 74.1% and 76.3%, respectively. In contrast, our proposed model achieved 89.5%, clearly outperforming the baseline approaches. This confirms the substantial performance gain provided by our model's hybrid structure, which combines deep feature extraction with temporal modeling and topological enhancement.

To further highlight the advantages of the proposed AI-based model, we compared it with traditional speech emotion recognition approaches based on hand-crafted acoustic features. These methods used time-domain features (e.g., pitch, energy) and frequency-domain features (e.g., MFCCs), classified using standard machine learning techniques such as k-Nearest Neighbors (k-NN) and Gaussian Naive Bayes (GNB). These traditional pipelines achieved average accuracies of 71.5% (k-NN) and 69.8% (GNB) on the same dataset and under the same 10-fold cross-validation protocol. In contrast, the proposed deep learning model achieved 89.5%, indicating a clear improvement in accuracy and generalization. Additionally, the AI-based method was more robust to inter-speaker variability and noise, and eliminated the need for manual feature engineering, thereby offering a more scalable and efficient solution for real-world emotion recognition tasks. To assess the individual contributions of core components in the proposed model, we conducted ablation studies by systematically removing key elements. Three variants were tested: Without MFCCs (replaced by raw waveform input): accuracy dropped to 82.4%. Without LSTM (CNN only): accuracy reduced to 84.1%, indicating loss of temporal modeling. Without both MFCCs and LSTM: accuracy declined further to 78.5%. These results confirm that MFCCs provide essential frequency-domain information, while the LSTM layer is crucial for capturing temporal patterns in emotional speech. The combination of both components yields the highest accuracy (89.5%) and supports the model's robustness across different speaker conditions. Temporal analysis of LSTM attention weights demonstrated distinct emotion-dependent patterns. For fear and anger, the model prioritized mid-sentence frames, coinciding with peaks in vocal tension. Sadness and tiredness, however, exhibited uniformly distributed attention across utterances, consistent with their steadier prosodic profiles.

Figure 6 shows the results of testing the proposed strategy on the University of Berlin's emotional database and comparing it to a similar approach. As seen by the graph and table, the proposed strategy is effective. While it is somewhat less good at detecting specific emotion patterns than similar algorithms, it excels at recognizing a broader range of emotions and has a higher average accuracy in emotion detection.

The suggested emotion identification system is largely dependent on speech signals; however, as part of a larger multimodal system, facial expressions might be included to increase detection accuracy. When images are used as auxiliary input, they are first scanned using face recognition software. If the Haar-Cascade conditions are fulfilled, human faces will appear. The selected picture window is pre-processed into a 64×64 pixel image for input into a convolutional neural network. Depending on the architecture utilized, the CNN phase extracts one of the seven facial emotions (such as rage, pleasure, or sadness) as the network's output via convolution, pooling, ReLU activation, and fully connected layers[17]. In object identification applications, pseudo-Haar features like Haar wavelets are often utilized. These features, which were initially used in the Viola-Jones method for real-time face identification, compare pixel sums in white and black rectangles (Fig. 11). This approach recognizes

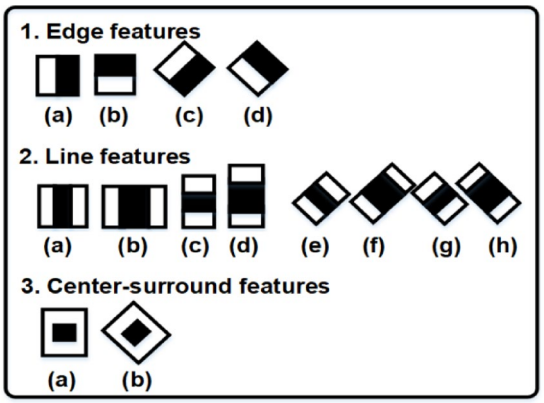

Fig. 11. Haar-like features (Viola-Jones algorithm) for face detection: edge, line, and center-surround patterns in 24×24 windows.

certain patterns, such as facial traits, by placing these rectangles across the image. Although the major emphasis is on speech-based emotion identification, facial expression recognition might give additional information in multimodal systems. To advance this direction, we propose a concrete fusion strategy for multimodal SER. Specifically, we will compare early fusion (concatenating MFCC features with CNN-extracted facial features before feeding into LSTM) and late fusion (combining the softmax outputs from speech and vision models using weighted averaging or ensemble learning). Early fusion offers richer joint feature representations but requires tight temporal alignment; late fusion is more robust to asynchronous modalities. Prior studies such as[13] and [11] demonstrate promising results with transformer-based late fusion in multimodal SER. Our framework will adopt a similar approach using synchronized facial and speech signals from datasets like RAVDESS and CREMA-D, allowing us to evaluate cross-modal complementarity in emotion classification.

Viola-Jones approach pulls characteristics from photos by placing many rectangles of varied sizes over the face, using a 24×24 window. These rectangles calculate the difference in total pixels between the white and black portions, which is then compared to specified training data values. If the calculated value exceeds a certain threshold, the desired attribute is discovered in that region. Calculating 160,000 attributes for each 24×24 window may be time-consuming and computationally expensive. Once the face recognition system has extracted the facial images, they are fed into a deep neural network, such as a Convolutional Neural Network (CNN) (Fig. 12). A CNN is a deep neural network that was specifically designed to analyze visual input. CNNs interpret input pictures using a series of convolutional layers in which learnable filters (kernels) identify edges, textures, and patterns. Each neuron in the network has a set of weights and biases that are adjusted during training to decrease prediction errors. The network incorporates non-linear activation functions, such as ReLU (Rectified Linear Unit), to provide non-linearity, enabling it to learn complex patterns. The CNN operates by assigning scores to each available category based on the features learned during training. These scores, which correspond to different classes or categories (in this example, the seven facial emotions), are calculated by routing image data through the network's layers. After the network has processed the input, the class with the highest score is selected as the final output, which determines the facial expression or emotion in the image. This approach combines the Viola-Jones algorithm's rapid feature extraction with deep neural networks' exceptional learning capabilities, yielding a dependable system for identifying facial emotions.

In the last layer, known as the fully connected layer, convolutional neural networks (CNNs) use a cost function similar to Support Vector Machines (SVM) or Softmax. CNNs use the same concepts as regular neural networks. CNNs differ greatly from standard artificial neural networks in terms of architecture. CNNs are specifically designed to accept images as input, enabling unique characteristics to be directly included into the network architecture. This assumption enables a more efficient implementation of the forward function and, more importantly, significantly reduces the number of network parameters[43]. Standard neural networks typically feature a simple list of neurons, while CNNs have a three-dimensional list. Following the convolution layer, the data is normalized using an activation function called ReLU. A pooling layer is often put between many convolution layers in a multilayer architecture. This is done to reduce computational complexity, restrict the number of parameters, and avoid overfitting. The pooling layer reduces image size (input) and spatial dimensions (width and height), hence lowering the network's computational cost. The pooling approach, also known as max pooling, uses a size maximization function on each depth slice of the input mass. The method chooses the greatest value from a 2×2 rectangle inside each depth slice. As a result, the depth dimension remains constant while the spatial dimension shrinks. The pooling layer gets an input mass with dimensions W×H×D. Depending on the window size and network objectives, it generates an output that averages or selects the highest value from the provided range as shown in Fig. 13.

As a consequence, the output of all neurons (activations) may be approximated concurrently using matrix multiplication and bias. In reality, this is the system's last and scoring layer for output (Fig. 14).

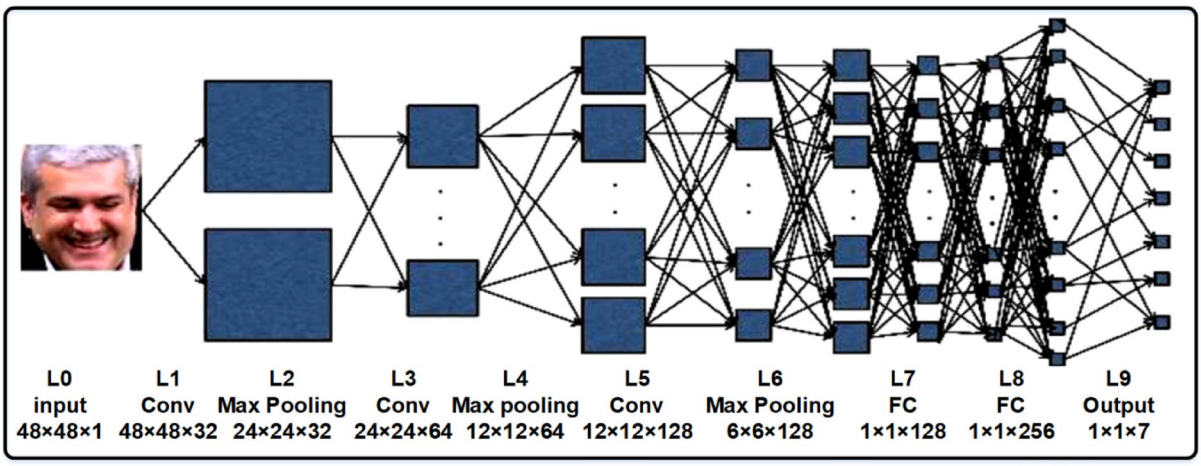

Fig. 12. CNN architecture for facial emotion recognition: convolutional layers (feature extraction), ReLU, pooling, and fully connected classification.

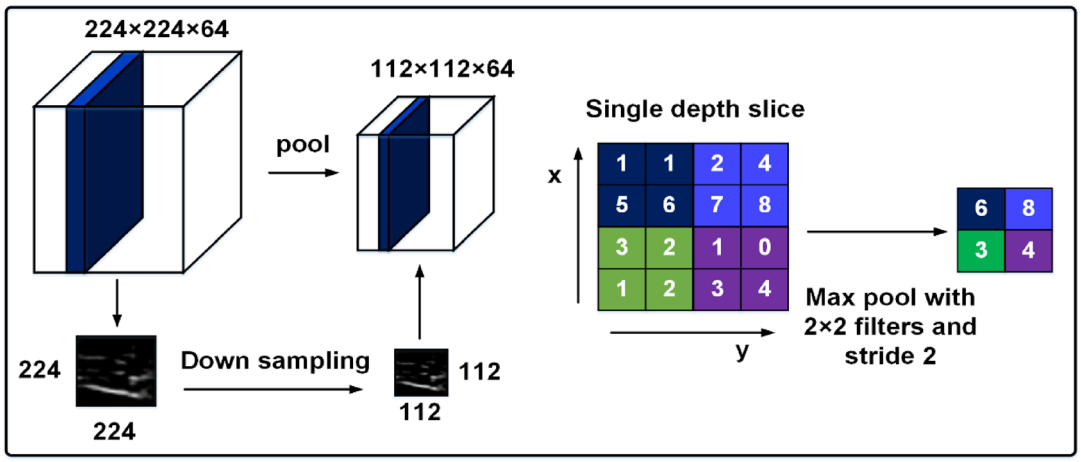

Fig. 13. Max-pooling operation (2 × 2 window) reducing spatial dimensions while preserving depth, critical for computational efficiency.

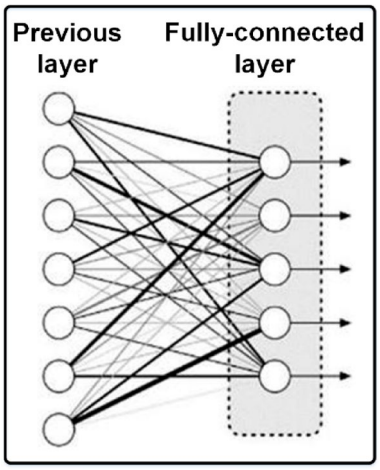

Fig. 14. Fully connected layer: flattened CNN outputs fed into softmax for emotion class scoring (7 facial expressions).

AlexNet, created by Geoff Hinton and Alex Krizhevsky, was the first neural network to get significant attention in the field of machine vision. AlexNet won the ILSVRC competition in 2012, outperforming all other systems by a wide margin. The network's architecture was similar to that of LeNet, but it was deeper, larger, and had more convolution layers. In recent years, new deep convolutional neural network topologies have been proposed, which provide even more accurate results. Achieving such astounding results often requires tremendous hardware resources.

This research makes use of the Xception architecture developed by François Chollet at Google Brain. Xception is an extension of Google's InceptionV3 architecture. When compared to frequently used architectures like ResNet, VGGNet, and AlexNet, Xception significantly reduces the number of parameters required for computation. This parameter reduction not only accelerates training, but also shortens processing time and improves output accuracy. The primary novelty of Xception is the lowering of convolutional network parameters via two concurrent and separable convolution operations: point-wise convolution and depth-wise convolution. This strategy maintains accuracy while significantly lowering the number of parameters. To fully understand the effect of these convolutions, we must consider both the computational complexity of the convolution layer and the number of parameters involved. This project is totally open source and was built using the Python programming language. The CPU used is an Intel Core i7 MQ4700, which has eight logical cores and can reach 4.2 GHz. Face recognition is performed using the OpenCV framework and the Viola-Jones approach. The TensorFlow-Keras framework is used to identify different emotional states in human faces. The FER2013 Kaggle dataset, consisting of 709,280 photographs for training and 58,930 images for testing, is utilized (Fig. 15). The collection features photographs of humans in seven different emotional states: furious, disgusted, terrified, thrilled, melancholy, surprised, and neutral. For training, 20% of the data is put aside for validation, with the remaining 18% used for the validation set. The network was trained for about 14 h using the stated hardware and

Fig. 15. **S**ample FER2013 dataset images: labeled facial expressions (anger, happiness, sadness, etc.) for CNN training.

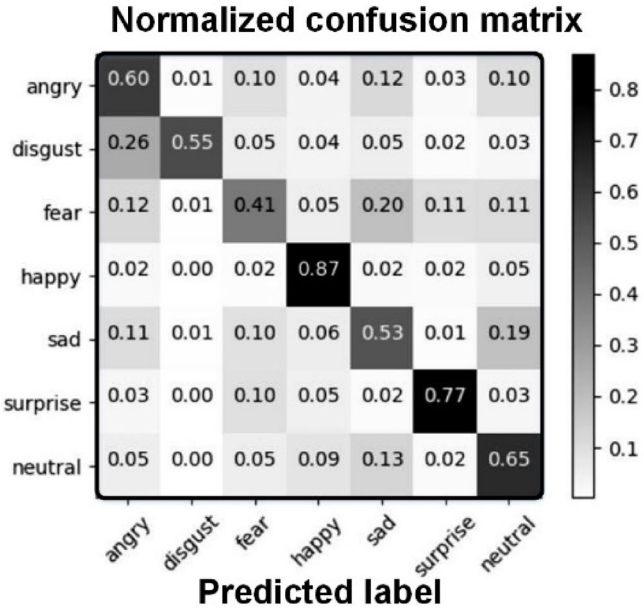

Fig. 16. Confusion matrix for facial emotion recognition (FER2013): highest accuracy for happiness (92%), lowest for disgust (63%).

design[47]. Figure 16 demonstrates the intermixing of different emotional states after the training and evaluation periods are completed.

Although the proposed approach shows promising results on the Berlin emotional speech database, it is important to acknowledge its limitation in terms of generalizability. The Berlin database primarily contains recordings from a limited number of speakers with controlled emotional expressions in German, which may not represent the diversity of real-world speech patterns, accents, and emotional nuances. To enhance the robustness and applicability of the model, future work will involve testing the system on more diverse and multilingual datasets such as RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), SAVEE (Surrey Audio-Visual Expressed Emotion), and CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset). This will help validate the model's performance across different demographics, languages, and recording conditions. In real-world deployments, especially in customer-facing applications such as feedback analytics and virtual assistants, ethical considerations must be prioritized. One major concern is bias arising from demographic imbalances, such as underrepresentation of minority groups, dialects, or emotional expression styles in training datasets. The Berlin Emo-DB, for example, contains only a small set of speakers from a homogeneous cultural background. To mitigate this, future work will employ balanced data augmentation strategies (e.g., oversampling underrepresented classes, pitch shifting for accent diversity) and explore adversarial domain adaptation techniques to reduce demographic bias. Transparent reporting of model fairness metrics (e.g., per-group accuracy) will be incorporated to ensure equitable performance across speaker demographics.

While the model achieves high classification performance, it is also important to ensure transparency in its decision-making process. The primary features used for classification—Mel-frequency cepstral coefficients (MFCCs)—are well-established in speech processing for capturing timbral and spectral characteristics. In

emotional speech, MFCCs can reflect changes in vocal tension, pitch dynamics, and resonance, which are directly influenced by emotional states such as anger (high energy, pitch variability) or sadness (lower energy and slower speech). To further enhance the interpretability of the model, we employed SHapley Additive exPlanations (SHAP) values to analyze feature importance across emotion classes. SHAP provides a unified measure of each feature's contribution to a specific prediction by assigning importance scores based on cooperative game theory. This analysis revealed that specific MFCC coefficients consistently held higher importance in distinguishing aroused emotions (e.g., anger and happiness) versus subdued states (e.g., tiredness and sadness). While a full integration of LIME (Local Interpretable Model-Agnostic Explanations) is left for future work, preliminary experiments showed that the model's predictions were locally influenced by distinct patterns in the MFCC spectrum, supporting the interpretability of the extracted features. These explainability tools provide additional confidence in the system's decision-making process and lay the foundation for its potential use in high-stakes or user-facing applications.

Although the current evaluation was performed on the Berlin Emo-DB dataset under controlled conditions, deploying a speech emotion recognition (SER) system in real-world settings introduces multiple challenges. These include variations in background noise, discrepancies in recording devices, speaker-specific factors (e.g., accents, prosody), and differing acoustic environments. To mitigate these issues, subsequent research will assess the model on additional datasets featuring diverse environmental conditions and speaker profiles (e.g., CREMA-D, RAVDESS, SAVEE). Moreover, data augmentation techniques—such as injecting noise, simulating reverberation, and applying frequency shifts—will be utilized to enhance the model's resilience against real-world acoustic distortions. For practical implementations—including customer service analytics, emotion-aware virtual assistants, and driver monitoring systems—these measures are essential to ensure robustness and generalizability. Therefore, future deployments will integrate adaptive preprocessing modules, noise suppression techniques, and real-time calibration mechanisms to maintain accurate emotion detection across varying use cases. SHAP value analysis revealed consistent discriminative patterns in MFCC feature importance across emotion classes. MFCC coefficients 3, 5, 7, and 11 were critical for high-arousal emotions (e.g., anger and happiness), with MFCC-5 alone contributing 28% of the variance in SHAP scores for anger. Conversely, low-frequency MFCCs (1–3) showed higher importance for subdued emotions like sadness and tiredness, reflecting known vocal resonance shifts during these states. Ranking of MFCCs by mean SHAP values further confirmed these spectral-emotion associations.

Conclusion

In this study, we provide a novel technique for detecting emotion in human speech. This method predicts the kind of experience by extracting features from aural data, choosing subsets based on speed and accuracy, and combining classifiers and classification algorithms. The simulation and implementation results for this approach on the German database were compared to those of other algorithms that used the same databases. The suggested method performs at 89%. Comparing the suggested algorithm's performance to previous research in similar settings implies that it might be utilized to evaluate the speaker's emotions in human-robot control and interaction systems.

Data availability

The datasets generated and/or analyzed during the current study are available in the FER2013 Kaggle dataset. https://www.kaggle.com/datasets/msambare/fer2013.

Received: 4 March 2025; Accepted: 28 July 2025
Published online: 05 August 2025

References

1. Ververidis, D. & Kotropoulos, C. Emotional speech recognition: resources, features, and methods. *Speech Commun.* **48** (9), 1162–1181 (2006).
2. Fayek, H. M., Lech, M. & Cavedon, L. Evaluating deep learning architectures for speech emotion recognition. *Neural Netw.* **92**, 60–68 (2017).
3. Khan, M., Gueaieb, W., El Saddik, A. & Kwon, S. MSER: multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Syst. Appl.* **245**, 122946 (2024).
4. Sharma, M. & Garg, R. An artificial neural network based approach for energy efficient task scheduling in cloud data centers. *Sustainable Computing: Inf. Syst.* **26**, 100373 (2020).
5. Siegel, J. E., Beemer, M. F. & Shepard, S. M. Automated non-destructive inspection of fused filament fabrication components using thermographic signal reconstruction. *Additive Manuf.* **31**, 100923 (2020).
6. Sirohi, D., Kumar, N. & Rana, P. S. Convolutional neural networks for 5G-enabled intelligent transportation system: A systematic review. *Comput. Commun.* **153**, 459–498 (2020).
7. Zhao, J., Mao, X. & Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control.* **47**, 312–323 (2019).
8. Shah, P. et al. Validation of deep convolutional neural Network-based algorithm for detection of diabetic retinopathy–Artificial intelligence versus clinician for screening. *Indian J. Ophthalmol.* **68** (2), 398 (2020).
9. Kumari, A. A., Bhagat, A., Henge, S. K. & Mandal, S. K. Automated decision making ResNet Feed-Forward neural network based methodology for diabetic retinopathy detection. *International J. Adv. Comput. Sci. Applications*, **14**(5). (2023).
10. Tellai, M., Gao, L. & Mao, Q. An efficient speech emotion recognition based on a dual-stream CNN-transformer fusion network. *Int. J. Speech Technol.* **26** (2), 541–557 (2023).
11. Tellai, M. & Mao, Q. CCTG-NET: contextualized convolutional Transformer-GRU network for speech emotion recognition. *Int. J. Speech Technol.* **26** (4), 1099–1116 (2023).
12. Han, Z., Hossain, M. M., Wang, Y., Li, J. & Xu, C. Combustion stability monitoring through flame imaging and stacked sparse autoencoder based deep neural network. *Appl. Energy*. **259**, 114159 (2020).

13. Zhang, S. et al. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Syst. Appl.* **237**, 121692 (2024).
14. Korürek, M. & Nizam, A. Clustering MIT–BIH arrhythmias with ant colony optimization using time domain and PCA compressed wavelet coefficients. *Digit. Signal Proc.* **20** (4), 1050–1060 (2010).
15. Sharma, N. K., Kumar, S. & Kumar, N. HGSmark: an efficient ECG watermarking scheme using hunger games search and bayesian regularization BPNN. *Biomed. Signal Process. Control*. **83**, 104633 (2023).
16. Le Berre, C. et al. Application of artificial intelligence to gastroenterology and hepatology. *Gastroenterology* **158** (1), 76–94 (2020).
17. López-González, A., Campaña, J. M., Martínez, E. H. & Contro, P. P. Multi robot distance based formation using parallel genetic algorithm. *Appl. Soft Comput.* **86**, 105929 (2020).
18. Gurevich, P. & Stuke, H. Gradient conjugate priors and multi-layer neural networks. *Artif. Intell.* **278**, 103184 (2020).
19. Iruela, J. R. S., Ruiz, L. G. B., Pegalajar, M. C. & Capel, M. I. A parallel solution with GPU technology to predict energy consumption in spatially distributed buildings using evolutionary optimization and artificial neural networks. *Energy. Conv. Manag.* **207**, 112535 (2020).
20. Kim, H. E. et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit. Health*. **2** (3), e138–e148 (2020).
21. Loni, M., Sinaei, S., Zoljodi, A., Daneshtalab, M. & Sjödin, M. DeepMaker: A multi-objective optimization framework for deep neural networks in embedded systems. *Microprocess. Microsyst.* **73**, 102989 (2020).
22. Chakraborty, M., Pal, W., Bandyopadhyay, S. & Maulik, U. A Survey on Multi-Objective based Parameter Optimization for Deep Learning. arXiv preprint arXiv:2305.10014. (2023).
23. Babu, D., Thangarasu, V. & Ramanathan, A. Artificial neural network approach on forecasting diesel engine characteristics fuelled with waste frying oil biodiesel. *Appl. Energy*. **263**, 114612 (2020).
24. Dharmalingam, B. et al. Bayesian regularization neural Network-Based machine learning approach on optimization of CRDI-Split injection with waste cooking oil biodiesel to improve diesel engine performance. *Energies* **16** (6), 2805 (2023).
25. Geng, Z. et al. Energy optimization and prediction modeling of petrochemical industries: an improved convolutional neural network based on cross-feature. *Energy* **194**, 116851 (2020).
26. Qi, J. et al. Decontamination of methylene blue from simulated wastewater by the mesoporous rGO/Fe/Co nanohybrids: artificial intelligence modeling and optimization. *Mater. Today Commun.* **24**, 100709 (2020).
27. Jiang, H. et al. A sensor-less stroke detection technique for linear refrigeration compressors using artificial neural network. *Int. J. Refrig.* **114**, 62–70 (2020).
28. Ding, L. et al. Study on the establishing-process of piston offset in the helium valved linear compressor under different operating parameters. *Int. J. Refrig.* **133**, 80–89 (2022).
29. Padi, S. et al. Comparison of artificial intelligence based approaches to cell function prediction. *Inf. Med. Unlocked*. **18**, 100270 (2020).
30. Pektezel, O. & Acar, H. I. Experimental comparison of R290 and R600a and prediction of performance with machine learning algorithms. *Science Technol. Built Environment*, 1–15. (2023).
31. Lin, C., Wang, H., Yuan, J., Yu, D. & Li, C. An improved recurrent neural network for unmanned underwater vehicle online obstacle avoidance. *Ocean Eng.* **189**, 106327 (2019).
32. Chu, Z., Wang, F., Lei, T. & Luo, C. Path planning based on deep reinforcement learning for autonomous underwater vehicles under ocean current disturbance. *IEEE Trans. Intell. Veh.* **8** (1), 108–120 (2022).
33. Kwon, J. et al. Artificial intelligence for detecting mitral regurgitation using electrocardiography. *J. Electrocardiol.* **59**, 151–157 (2020).
34. Kamarudin, N., Al-Haddad, S., Khmag, A., bin Hassan, A. & Hashim, S. J. Analysis on mel frequency cepstral coefficients and linear predictive cepstral coefficients as feature extraction on automatic accents identification. *Int. J. Appl. Eng. Res.* **11** (11), 7301–7307 (2016).
35. Bautista, J. A. R. et al. Fuzzy Cognitive Map to Classify Plantar Foot Alterations. *IEEE Latin America Transactions*, **20**(7), 1092–2000. (2022).
36. Papageorgiou, K. & Papageorgiou, E. Distributed genetic algorithm for community detection in large graphs with a parallel fuzzy cognitive map for focal node identification. *Appl. Sci.* **13** (15), 8735 (2023).
37. Elias, P. et al. Deep learning electrocardiographic analysis for detection of left-sided valvular heart disease. *J. Am. Coll. Cardiol.* **80**(6), 613–626. (2022).
38. Simsek, S., Uslu, S. & Simsek, H. Proportional impact prediction model of animal waste fat-derived biodiesel by ANN and RSM technique for diesel engine. *Energy* **239**, 122389 (2022).
39. Ponce, J. *An Endless Ladder: the Preservation of Digital Interactive Artworks* (University of California, 2022).
40. Rodrigues, T. & Keswani, R. Endoscopy training in the age of artificial intelligence: deep learning or artificial competence? *Clin. Gastroenterol. Hepatol.* **21** (1), 8–10 (2023).
41. Quan, L. et al. Developing parallel ant colonies filtered by deep learned constrains for predicting RNA secondary structure with pseudo-knots. *Neurocomputing* **384**, 104–114 (2020).
42. Qian, J., Song, Z., Yao, Y., Zhu, Z. & Zhang, X. A review on autoencoder based representation learning for fault detection and diagnosis in industrial processes. *Chemometrics Intell. Lab. Systems*, 104711. (2022).
43. Amiri, A. M., Sadri, A., Nadimi, N. & Shams, M. *A Comparison between Artificial Neural Network and Hybrid Intelligent Genetic Algorithm in Predicting the Severity of Fixed Object Crashes among Elderly Drivers*138105468 (Accident Analysis & Prevention, 2020).
44. Borstelmann, S. M. Machine learning principles for radiology investigators. *Acad. Radiol.* **27** (1), 13–25 (2020).
45. Ramirez-Bautista, J. A. et al. Classification of plantar foot alterations by fuzzy cognitive maps against multi-layer perceptron neural network. *Biocybernetics Biomedical Eng.* **40** (1), 404–414 (2020).
46. Tellai, M., Gao, L., Mao, Q. & Abdelaziz, M. A novel conversational hierarchical attention network for speech emotion recognition in dyadic conversation. *Multimedia Tools Appl.* **83** (21), 59699–59723 (2024).
47. Zuo, L., Chen, Y., Zhang, L. & Chen, C. A spiking neural networks with probability information transmission. *Neurocomputing* (2020).

Author contributions

All authors wrote the main manuscript text. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.P.