ADBA Chaos and Fractals
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8951
RESEARCH ARTICLE
Vol.3 / No.2 / 2026 / pp.117-124
https://doi.org/10.69882/adba.chf.2026075

Efficacy of Multi-Domain Acoustic Features in Speech Emotion Recognition: A Comparative Analysis of Machine Learning and Deep Learning Models

Sude Emine Baydar [*,1] **and Muhammed Ali Pala** [α,2]

*Biomedical Engineering Graduate Student, Sakarya University of Applied Sciences, Sakarya, Türkiye, αFaculty of Technology, Department of Electrical and Electronics Engineering, Sakarya University of Applied Sciences, Sakarya , Türkiye.

ABSTRACT Speech Emotion Recognition plays a pivotal role in advancing human-computer interaction; however, achieving high model generalization on limited and controlled datasets remains a persistent challenge. This study investigates the efficacy of multi-domain acoustic features and compares the classification performance of traditional machine learning algorithms with that of baseline deep neural architectures. Using the speech subset of the RAVDESS dataset, a comprehensive feature vector was constructed by extracting time-, frequency-, and cepstral-domain attributes, with a particular focus on Mel-Frequency Cepstral Coefficients (MFCCs) and mel-spectrogram properties. To mitigate overfitting and simulate real-world variability, acoustic data augmentation techniques, specifically noise injection, pitch shifting, and time stretching, were applied. Support Vector Machine, Random Forest, and Deep Neural Network (DNN) models were systematically trained and evaluated. Empirical results demonstrated that the SVM achieved the most robust performance, yielding an overall accuracy of 79.51%. In contrast, despite data augmentation, the high parametric complexity of the DNN resulted in significant overfitting; while its training accuracy reached 97%, its test accuracy degraded to 73.78%. The RF model recorded the lowest baseline performance at 71.18%. Error analysis revealed that misclassifications were predominantly concentrated among acoustically congruent emotion pairs, such as *happy–surprised* and *sad–calm*. The findings underscore that while MFCCs and mel-spectrograms are decisive descriptors for SER, well-regularized traditional algorithms like SVM offer superior generalization capabilities compared to standard feed-forward deep learning models when data is scarce.

KEYWORDS
Speech emotion recognition
SVM
Deep learning
MFCC
RAVDESS
Machine learning

INTRODUCTION

Communication is one of the most fundamental processes of human interaction, enabling individuals to communicate emotions, thoughts, and needs to each other (Juslin and Laukka 2003). In everyday interactions, people express themselves not only through spoken words but also through various acoustic and prosodic cues embedded within speech. The same sentence may convey entirely different meanings depending on intonation, stress, speaking rate, rhythm, and vocal intensity (Hsu *et al.* 2024). Therefore, the accurate interpretation of verbal communication depends not only on linguistic content but also on the manner in which speech is produced (Tao and Tan 2005). Acoustic characteristics such as pitch, duration, rhythm, intensity, stress, and pauses play a critical role in emotional expression (Akçay and Oğuz 2020).

Emotional states, including happiness, sadness, anger, fear, and surprise, can often be inferred not only from lexical content but also from variations in vocal patterns (Kane *et al.* 2024). Consequently, effective speech analysis requires the evaluation of both linguistic and emotional-acoustic information (Pala 2025). In addition to emotional expression, speech signals provide critical information regarding an individual's cognitive and psychological state. Variations in speech production may reflect emotional states such as confidence, anxiety, hesitation, or uncertainty (Jiang and Pell 2018). Accordingly, speech is considered a multidimensional biological signal that reflects both human behaviour and internal psychological processes (Brahmi *et al.* 2024) .This complex structure has made speech analysis increasingly important not only for interpersonal communication but also for computer-based systems and intelligent technologies (Tao and Tan 2005).

With the rapid advancement of human–computer interaction (HCI) technologies, enabling computer systems to recognise and interpret human emotions has become a pivotal research area (Cowie *et al.* 2001). In applications such as intelligent virtual assistants, call centre systems, educational technologies, healthcare applications, and social robots, the objective is not only to recognise spoken words but also to determine the speaker's emotional state (Hsu *et al.* 2024). However, despite significant technological progress, computer systems still struggle to achieve human-level performance in emotional interpretation. As a result, Speech Emotion Recognition (SER) has attracted considerable attention in recent

Manuscript received: 3 June 2026,
Revised: 6 July 2026,
Accepted: 7 July 2026.

[1]25500305001@subu.edu.tr (**Corresponding author**)
[2]pala@subu.edu.tr

years (Akçay and Oğuz 2020).

Speech Emotion Recognition systems aim to automatically identify emotional states using acoustic and prosodic information extracted from speech signals. In SER studies, various acoustic features are commonly used to represent the temporal, spectral, and energy-related characteristics of speech (Çolakoğlu *et al.* 2022). Among these features, Mel-Frequency Cepstral Coefficients (MFCC), mel-spectrograms, zero-crossing rate, energy-based parameters, and spectral descriptors are widely preferred (Durukal and Hocaoglu 2015). MFCC are particularly important because they model speech perception like the human auditory system. Similarly, mel-spectrograms provide detailed time–frequency representations that are especially suitable for deep learning-based approaches (Mustaqeem and Kwon 2019). The combined use of different acoustic representations contributes significantly to improving emotion classification performance (Dixit *et al.* 2024).

Various artificial intelligence methods have been employed for speech emotion recognition tasks. Among traditional machine learning approaches, Support Vector Machine and Random Forest algorithms are widely used because of their strong classification performance on high-dimensional datasets (Cortes and Vapnik 1995). In recent years, deep learning approaches such as Convolutional Neural Networks and Deep Neural Networks have become increasingly prominent due to their ability to automatically learn complex nonlinear relationships within data (Fayek *et al.* 2017). Furthermore, deep learning models reduce the need for manual feature engineering and can achieve high classification performance when sufficient data are available (Pala 2025).

In this study, the problem of speech emotion recognition was investigated using speech recordings from the RAVDESS dataset (Livingstone and Russo 2018). To improve model generalisation and reduce overfitting, data augmentation techniques, including noise addition, pitch shifting, and time stretching, were applied during preprocessing. Acoustic features representing emotional characteristics, including MFCC, mel-spectrograms, zero-crossing rate, energy-related parameters, and spectral features, were extracted from the speech recordings. Using these features, Support Vector Machine , Random Forest , and Deep Neural Network models were trained and comparatively evaluated. The primary objective of this study was to examine the effectiveness of different artificial intelligence approaches for speech emotion recognition and to determine suitable model architectures for this task, particularly in the context of constrained datasets (Yılmazcan and Pala 2026).

METHODOLOGY

Dataset

In this study, the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset was used for speech emotion recognition tasks (Livingstone and Russo 2018). RAVDESS is a standardised, publicly available database widely used in emotion recognition research for its high recording quality and well-structured emotional labelling (Serrano *et al.* 2025). The dataset consists of audio recordings collected in a controlled studio environment using professional recording equipment. During recording, environmental noise was minimised through standardised recording protocols, fixed microphone positioning, and balanced sound levels, yielding clean, consistent speech signals suitable for acoustic analysis. The dataset includes recordings from 24 professional actors, 12 female and 12 male. Each participant uttered predefined sentences using different emotional intonations. The complete database contains both speech and song recordings, enabling the investigation of emotional expression across different vocal production types (Juslin and Laukka 2003).

Each recording is labelled with one of eight emotional categories: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. The distribution of the speech samples across these emotional categories is illustrated in Figure 1. This balanced distribution across classes ensures that the models are trained on a representative range of emotional expressions, thereby mitigating potential biases (Akçay and Oğuz 2020). In addition, several emotions were recorded at different intensity levels, allowing the dataset to represent not only emotion categories but also variations in emotional intensity (Wu *et al.* 2002). This structure provides an important advantage for classification models by enabling the learning of subtle acoustic differences between similar emotional states. Furthermore, the dataset includes a systematic file-naming structure that contains information on modality, vocal channel type, emotion category, emotional intensity, spoken statement, repetition number, and speaker identity. This organisation facilitates preprocessing, labelling, and class-based segmentation procedures (Serrano *et al.* 2025).

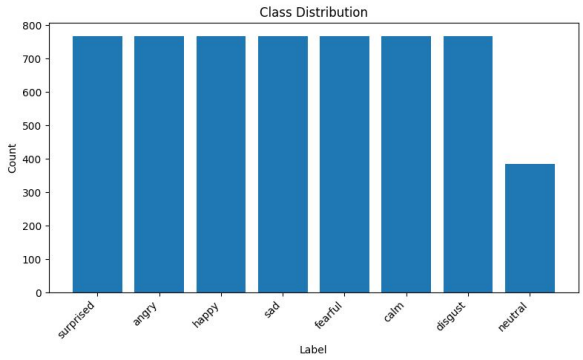

Figure 1 Distribution of speech samples in the original RAVDESS dataset

Data Augmentation

Data augmentation techniques were implemented in this study to enhance model generalisation, mitigate overfitting, and improve robustness to varying acoustic conditions. During the augmentation process, controlled modifications were applied to the original speech signals while strictly preserving their fundamental emotional characteristics (Mustaqeem and Kwon 2019). This approach facilitated greater diversity in the training dataset and prevented the models from memorising specific samples, which is particularly critical when operating with relatively constrained datasets such as RAVDESS (Livingstone and Russo 2018). Specifically, three distinct augmentation methods were applied to each speech signal: noise addition, pitch shifting, and time stretching. In the noise-addition procedure, low-level random background noise was added to the original recordings to simulate realistic environmental conditions encountered in practical applications. This process aimed to improve model robustness under low signal-to-noise ratio (SNR) conditions, ensuring stable performance across both pristine and degraded speech samples (McLaren *et al.* 2013).

Subsequently, the pitch-shifting procedure was employed to shift the fundamental frequency components of the speech signals upward or downward by specific ratios, while maintaining the original signal duration. This transformation simulated acoustic

variability associated with diverse speaker characteristics, such as gender, age, and vocal timbre (Ullah *et al.* 2023). Similarly, the time-stretching procedure was used to increase or decrease the speaking rate without altering the pitch structure, thereby accounting for temporal variations related to individual speech patterns and varying speaking speeds (Ming *et al.* 2023). By simulating these variations in a controlled manner, the augmentation techniques significantly increased acoustic diversity without compromising the underlying emotional content. In real-world scenarios, speech signals are inevitably influenced by environmental noise, microphone characteristics, and speaker-dependent variability (Pala 2025). Consequently, the integration of these synthetic variations served as a pivotal preprocessing step, enhancing the models' adaptation to real-world conditions and improving the overall classification performance and robustness of the proposed architectures (Fayek *et al.* 2017).

Feature Extraction

In speech emotion recognition systems, feature extraction is a critical stage in which raw speech signals are transformed into meaningful numerical representations suitable for classification algorithms. Raw audio signals consist of time-varying amplitude values; however, they do not directly provide structured information about emotional states. Therefore, various acoustic representations are required to characterise the temporal, spectral, and perceptual properties of speech (Akçay and Oğuz 2020; Çolakoğlu *et al.* 2022). Effective feature extraction is particularly vital because emotional characteristics are primarily reflected through subtle changes in speech acoustics. In this study, a hybrid set of time-domain, frequency-domain, cepstral, and time–frequency-based features was utilised to provide a comprehensive representation of emotional speech signals (Jha *et al.* 2022).

Time-domain features, specifically Zero Crossing Rate (ZCR) and Root Mean Square (RMS) energy, were extracted to capture the temporal dynamics of the signals. ZCR quantifies the number of sign changes within a signal over a given interval, providing insights into the dynamic structure and articulation patterns of speech. While higher ZCR values are generally associated with high-frequency components and rapid changes, lower values indicate smoother, more stationary structures (Durukal and Hocaoglu 2015). Concurrently, RMS energy represents the average physical intensity of the signal within a specific time window. Since emotional states such as anger or happiness typically exhibit higher vocal energy than calmer states, the integration of ZCR and RMS features enables the joint representation of temporal fluctuations and energy-related speech properties (Mustaqeem and Kwon 2019).

Frequency-domain features, including spectral centroid, spectral bandwidth, spectral roll-off, and spectral contrast, were also incorporated to describe the spectral distribution. The spectral centroid, associated with the perceptual "brightness" of sound, indicates the centre of mass of the frequency spectrum. Spectral bandwidth describes the spread of frequency components around this centroid, while spectral roll-off and spectral contrast provide information about the energy concentration and the distinctiveness of the spectral peaks and valleys, respectively (Gales *et al.* 1999). Given that emotional states significantly influence voice quality, articulation, and vocal intensity, these frequency-domain features play a pivotal role in distinguishing among emotion categories Dixit *et al.* (2024). Finally, cepstral features—comprising Mel-Frequency Cepstral Coefficients, along with their delta and delta-delta coefficients—were extracted. MFCC are a cornerstone of speech analysis, as they model frequency perception in a manner consistent with the human auditory system (Tao and Tan 2005).

This process represents the spectral envelope of speech in a compact, discriminative form by filtering the signal using the Mel scale. To capture the dynamic nature of speech, first-order (delta) and second-order (delta-delta) temporal variations were included, allowing the models to represent intonation, rhythm, and temporal transitions more effectively (Ullah *et al.* 2023). The temporal variation of the MFCC coefficients obtained from a representative emotional speech signal is illustrated in Figure 2.

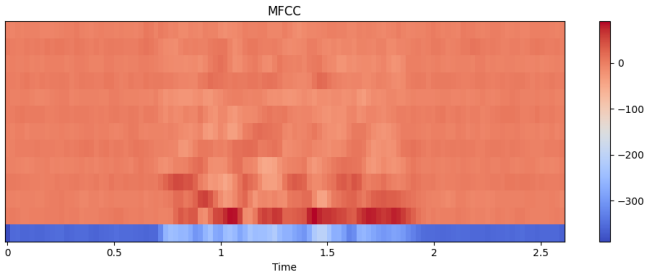

Figure 2 Display of MFCC coefficients for an angry speech signal

In addition to conventional acoustic features, mel-spectrogram-based representations were extracted to provide a more granular time–frequency analysis. A Mel-spectrogram offers a detailed visualisation of a speech signal based on the Mel scale, which closely aligns with the frequency-perception characteristics of the human auditory system (Tao and Tan 2005). In this representation, the horizontal axis corresponds to time, the vertical axis represents Mel-scaled frequency bands, and the colour intensity indicates the energy level in each band. Consequently, the temporal and spectral distribution of speech energy can be analysed both visually and numerically (Cui *et al.* 2018).

Mel-spectrogram-based features are particularly effective for capturing energy distribution, spectral transitions, and temporal intensity variations inherent in emotional speech (Mustaqeem and Kwon 2019). While MFCC provide a compact cepstral representation derived from mel-scaled spectral information, mel-spectrograms preserve more detailed time–frequency characteristics that might otherwise be lost during the cepstral transformation (Fayek *et al.* 2017). Therefore, the integrated use of MFCCs and mel-spectrograms enables the models to capture both the global spectral structure and localised temporal energy variations more comprehensively (Ullah *et al.* 2023). An examination of these feature sets demonstrates that MFCC efficiently represent the discriminative cepstral characteristics, whereas mel-spectrograms provide a more nuanced visualisation of spectral evolution over time (Dixit *et al.* 2024). The mel-spectrogram representation of the example speech signal is shown in Figure 3.

All features extracted and utilised in this study are summarised in Table 1, these features collectively represent the temporal, spectral, cepstral, and perceptual characteristics of the speech signals, thereby providing a comprehensive and multidimensional input space for the emotion classification models.

Implementation Details

All speech recordings were converted to mono signals and resampled to a uniform sampling rate of 16 kHz utilizing the librosa library in Python, followed by an amplitude normalization step to ensure consistency across the dataset. To capture a comprehensive representation of the acoustic characteristics, a high-dimensional feature vector was constructed by extracting temporal, spectral,

Table 1 Features extracted from speech signals

Feature Group	Feature	Description	Importance of Emotion Recognition
Time Domain	Zero Crossing Rate	The number of times the signal crosses the zero line	Speech intensity/dynamism
Time Domain	RMS Energy	Average energy level	Volume, the distinction between anger and happiness
Frequency Domain	Spectral Centroid	The centre of mass of the spectrum	Bright/crisp sound profile
Frequency Domain	Spectral Bandwidth	Frequency bandwidth	Variety of sounds
Frequency Domain	Spectral Roll-off	The cutoff frequency that accounts for most of the energy	High frequency density
Frequency Domain	Spectral Contrast	The difference between peaks and troughs across bands	The difference between peaks and troughs across bands
Cepstral	MFCC (13)	A spectral representation suited to human hearing	One of its strongest features
Cepstral	Delta MFCC	Temporal change	Changes in intonation
Cepstral	Delta-Delta MFCC	Acceleration	Emotional shifts
Time-Frequency	Mel Spectrogram	Time-dependent frequency energy	Powerful representations for deep learning

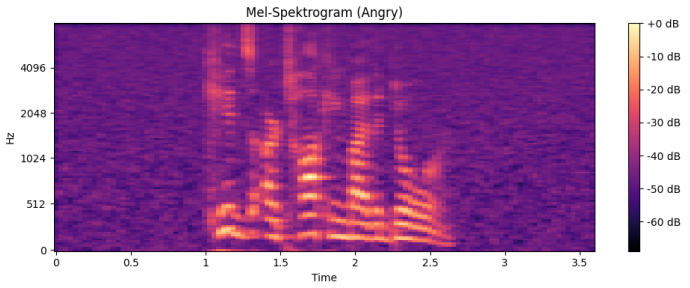

Figure 3 Mel-spectrogram representation of an angry speech signal

cepstral, and harmonic descriptors from each normalized audio signal. Temporal-domain features included the mean and standard deviation of both the Zero Crossing Rate (ZCR) and Root Mean Square (RMS) energy to capture time-domain signal dynamics. Spectral-domain features comprised the spectral centroid, spectral bandwidth, spectral roll-off (configured at an 85?% threshold), and spectral contrast to characterize the frequency distribution. Cepstral representation was established by extracting 13-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) along with the mean and standard deviation of their first-order (delta) and second-order (delta-delta) derivatives to capture dynamic temporal transitions. Furthermore, a 12-dimensional chroma vector, the mean and standard deviation of a 64-band Mel-spectrogram, and Tonnetz harmonic descriptors were computed to provide complementary harmonic and tonal information. Finally, all extracted acoustic descriptors were concatenated into a single, unified multi-dimensional feature vector prior to classification.

Classification

In speech emotion recognition studies, the classification stage is a critical process in which extracted acoustic features are mapped to their corresponding emotional categories. The performance of the selected classification algorithm directly influences the overall robustness and success of the system (Akçay and Oğuz 2020). Given the complexity of acoustic patterns, this study employs a comparative analysis using both traditional machine learning approaches and deep learning-based methods (Jha *et al.* 2022). Specifically, Support Vector Machines, Random Forest, and Deep Neural Network architectures were evaluated to assess their respective strengths in processing multidimensional acoustic features. The Support Vector Machine was utilised as a supervised learning algorithm to determine the optimal hyperplane that maximises the margin between different classes (Cortes and Vapnik 1995).

To address the nonlinear nature of the feature space, a Radial Basis Function (RBF) kernel was employed, enabling the model to learn complex decision boundaries. Concurrently, the Random Forest model was implemented as an ensemble learning approach. By aggregating the outputs of multiple decision trees trained on randomised subsets of the data, the RF model enhances generalisation capability and effectively mitigates the risk of overfitting through its inherent bagging mechanism (Jha *et al.* 2022). To capture more intricate, nonlinear relationships within the high-dimensional feature set, a Deep Neural Network architecture was developed. The proposed DNN consists of multiple fully connected (dense) layers, utilising nonlinear activation functions to expand its representational capacity (Kim and Kwak 2024). To ensure stable convergence and prevent overfitting, regularisation techniques, including dropout layers and early stopping criteria were integrated into the training process (Ullah *et al.* 2023). For model validation, the dataset was partitioned into training, validation, and testing subsets. Classification performance was rigorously evaluated using a suite of metrics, including accuracy, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). This multi-metric approach allowed for a nuanced assessment of model performance, particularly in terms of class-specific discriminative power. Consequently, this comparative framework provides a comprehensive understanding of how different algorithmic structures adapt to the challenges of speech emotion recognition.

Deep Neural Network (DNN) Architecture and Implementation

Deep Neural Networks are robust classification models capable of learning intricate nonlinear relationships through their multi-layered architectures. These models process input data through multiple hidden layers, generating increasingly abstract and discriminative feature representations at each stage, which allows for the successful modelling of complex data patterns (Fayek *et al.* 2017). In a DNN architecture, each layer processes the output from the preceding layer to produce a higher-level representation, thereby enabling hierarchical learning. This characteristic offers a significant advantage, particularly in problems involving high-

dimensional and nonlinear relationships. In speaker-independent emotion recognition tasks, features extracted from acoustic signals are typically high-dimensional and exhibit complex structures. Consequently, DNN-based approaches have been widely adopted in recent years due to their superior ability to learn nonlinear dependencies in such data (Brahmi *et al.* 2024). Particularly in studies using time-frequency representations such as mel-spectrograms, deep learning models have demonstrated an enhanced capacity to capture emotional patterns (Begazo *et al.* 2024). In the present study, the extracted multidimensional acoustic features were utilised as inputs to the DNN model, allowing it to learn the complex interrelationships among these variables (Dixit *et al.* 2024). Within a neural network, each neuron applies a linear transformation by weighting the input data, then generates a nonlinear output via an activation function. This process is formally expressed as follows:

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)} \tag{1}$$

$$\mathbf{a}^{(l)} = \sigma\left(\mathbf{z}^{(l)}\right) \tag{2}$$

During the training process, the loss function was minimised using the backpropagation algorithm, with weights updated iteratively (Fayek *et al.* 2017). To mitigate the risk of overfitting and ensure the model's generalisation capability, regularisation techniques, including dropout, early stopping, and the utilisation of a dedicated validation dataset were strictly employed. This rigorous approach ensures that the model maintains high performance not only on the training data but also on previously unseen test samples (Kim and Kwak 2024).

Specifically, the proposed Deep Neural Network (DNN) architecture consisted of six fully connected hidden layers with 2048, 1024, 512, 256, 128, and 64 neurons, respectively. To mitigate the vanishing gradient problem and ensure smooth nonlinear mapping, the Swish activation function was applied to all hidden layers. Batch Normalization was employed after each dense layer to stabilize the internal covariate shift and accelerate convergence, whereas Dropout layers with rates ranging from 0.20 to 0.45 were strategically inserted to reduce overfitting.

The output layer consisted of eight neurons with a Softmax activation function, producing a probability distribution over the eight emotion classes. Model optimization was performed using the Adam optimizer with an initial learning rate of 5×10^{-4} and the Categorical Cross-Entropy loss function. A label smoothing factor of 0.05 was applied to improve generalization. Training was conducted for a maximum of 600 epochs using a mini-batch size of 64. An EarlyStopping callback with a patience of 50 epochs was employed to restore the model weights corresponding to the lowest validation loss. In addition, the learning rate was automatically reduced by a factor of 0.5 using the ReduceLROnPlateau scheduler whenever the validation loss plateaued, with a minimum learning rate of 1×10^{-6}.

Support Vector Machine: The Support Vector Machine is a robust supervised learning algorithm that identifies the optimal decision boundary (hyperplane) that maximises the margin between classes. It is extensively utilised in speech emotion recognition tasks due to its effectiveness in handling high-dimensional datasets (Cortes and Vapnik 1995). The primary objective of SVM is to minimise the norm of the weight vector, which is equivalent to maximising the margin between the support vectors of different classes. This optimisation problem is formally defined as:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{3}$$

subject to the following constraint for all training samples:

$$y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1, \qquad i = 1, \ldots, N \tag{4}$$

where $\mathbf{w} \in \mathbb{R}^d$ represents the weight vector (normal vector) that defines the orientation of the optimal hyperplane, $b \in \mathbb{R}$ denotes the bias (intercept) term, which determines the offset of the hyperplane from the origin, $\mathbf{w}^T\mathbf{x}_i + b$ represents the linear decision function, where the superscript T denotes the transpose operator, ensuring a proper dot product between vectors, and $\|\mathbf{w}\|$ represents the Euclidean norm of the weight vector.

However, since emotional acoustic features are rarely linearly separable in real-world datasets, kernel functions are employed to model nonlinear decision boundaries. In this study, the Radial Basis Function (RBF) kernel was utilised to map data points into a higher-dimensional space, enabling the SVM to effectively capture complex relationships and distinguish between emotion categories with high precision.

In this study, the Support Vector Machine (SVM) classifier was configured with a Radial Basis Function (RBF) kernel to effectively handle the nonlinear boundaries within the high-dimensional speech feature space. The regularization parameter was set to $C = 10$ to achieve an optimal balance between maximizing the decision margin and minimizing training errors. The kernel coefficient (γ) was dynamically computed using the "scale" heuristic, defined as

$$\gamma = \frac{1}{n_{\text{features}} \times \text{Var}(X)} \tag{5}$$

to adapt to the variance of the extracted acoustic features. Furthermore, to mitigate the influence of class imbalance and prevent the classifier from biasing toward dominant classes, cost-sensitive learning was incorporated by applying class weights inversely proportional to class frequencies during model training.

Random Forest: Random Forest is a robust ensemble learning method that integrates multiple decision trees based on the bootstrap aggregating (bagging) principle. In this approach, each decision tree is trained on independent bootstrap samples of the dataset. By exposing each tree to different data distributions, the model's overall generalisation capability is significantly enhanced. Furthermore, at each node during the splitting process, a randomly selected subset of features is utilised rather than the entire feature set. This mechanism reduces the correlation between individual trees, effectively mitigating the risk of overfitting. While individual decision trees classify data by hierarchically partitioning the feature space, using criteria such as Information Gain or Gini Impurity, they are often prone to high variance and overfitting. The Random Forest architecture addresses these limitations by aggregating the outputs of numerous trees to yield more stable and robust predictions. Consequently, RF models exhibit superior performance, particularly in high-dimensional feature spaces and when dealing with noisy data (Jha *et al.* 2022).

The final classification result in a Random Forest is determined by majority voting, combining the predictions from all constituent trees. This process is formally defined as:

$$\hat{\mathbf{y}} = \text{mode}\left(T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_n(\mathbf{x})\right) \tag{6}$$

where $\mathbf{x}$ represents the input feature vector, $T_i(\mathbf{x})$ denotes the class prediction generated by the i-th decision tree in the forest, n is the total number of decision trees (bootstrap samples) in the forest, and $\hat{\mathbf{y}}$ represents the final predicted class label.

A significant advantage of the Random Forest model is its inherent ability to calculate feature importance, which identifies the acoustic features that are most discriminative for emotion classification. This interpretability is particularly valuable in speech emotion recognition, where understanding the contribution of specific spectral and temporal features is essential. Consistent with the literature, the Random Forest algorithm was selected because of its reliability and stable performance on complex, high-dimensional datasets24.

For the ensemble-based evaluation, the Random Forest (RF) classifier was implemented using 800 decision trees ($n_{\text{estimators}} = 800$) to ensure prediction stability and reduce variance. To address class imbalance during tree construction, the `balanced_subsample` strategy was employed, which automatically computes class weights from the bootstrap sample at each tree. Finally, to ensure reproducibility and enable consistent comparisons across different experimental runs, the pseudo-random number generator was initialized with a fixed seed (random_state = 42).

Evalutation Metrics

In this study, a comprehensive set of evaluation metrics, including accuracy, precision, recall, and F1-score was utilised to assess the performance of the speech-based emotion recognition models. These metrics facilitate a detailed analysis of both the overall system efficacy and class-specific performance (Jha *et al.* 2022). Accuracy represents the ratio of correct predictions to the total number of samples. While it provides a preliminary assessment of overall performance, it is often insufficient for evaluating multi-class tasks with imbalanced datasets. Consequently, Precision, Recall (Sensitivity), and F1-score were integrated to examine class-specific discriminative power (Serrano *et al.* 2025). Precision measures the reliability of the model's positive predictions, whereas recall evaluates the model's ability to identify all relevant instances within a specific emotional category correctly. To provide a balanced evaluation of these two metrics, the F1-score, the harmonic mean of precision and recall was employed as a more robust indicator of performance under class-imbalance conditions (Shah *et al.* 2023).

For the multi-class emotion recognition problem, both macro and weighted averages were calculated. The macro average assigns equal importance to all emotion classes regardless of their frequency, whereas the weighted average adjusts each class's contribution based on its sample size in the dataset. Furthermore, the Receiver Operating Characteristic curves and the Area Under the Curve metric were utilised to evaluate the models' discriminative performance (Brahmi *et al.* 2024). The ROC curve illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different threshold values. At the same time, the AUC quantitatively reflects the model's ability to distinguish a target emotion from other classes. In this multi-class framework, ROC-AUC values were calculated for each emotion category individually, providing a granular analysis of the model's strengths and limitations across different emotional states.

Experimental Results

In the scope of this research, the performance of the Support Vector Machine (SVM), Random Forest (RF), and Deep Neural Network (DNN) models was systematically evaluated. The comparative analysis indicates that each architecture exhibited distinct behaviours depending on the acoustic complexity of the emotion classes.

The learning dynamics of the DNN model, as depicted in the accuracy and loss curves in Figure 4, demonstrated a highly efficient optimisation process. Although the model achieved a training accuracy of 97%, the validation accuracy stabilised at 87%, highlighting a trade-off between learning capacity and generalisation. In contrast, the SVM model achieved the highest overall classification accuracy of 79.51%, indicating that its margin-maximisation principle is particularly effective for high-dimensional acoustic feature vectors used in speech emotion recognition. The Random Forest model achieved an overall accuracy of 71.18%. Despite its ensemble-based structure, its performance was relatively lower, which may be attributed to its greater sensitivity to noise present in prosodic features.

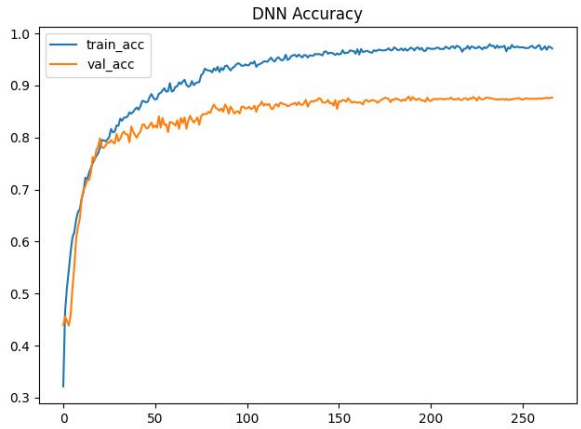

Figure 4 Training and validation performance of the DNN model

The detailed classification performance of the DNN model was further investigated using the confusion matrix (Figure 5) and the comparative performance chart (Figure 6). The model, evaluated on 1,152 test samples, achieved varying levels of performance across the emotion categories. As illustrated in Figure 6, the highest recognition performance was obtained for the Angry (0.85) and Disgust (0.81) classes. From a mathematical perspective, the high precision and recall achieved for these categories indicate that their acoustic signatures, characterised by elevated energy levels and pronounced frequency variations, occupy relatively distinct regions within the feature space.

In contrast, the Sad class presented the greatest challenge for the DNN model, yielding an F1-score of 0.59. The confusion matrix demonstrates that sad speech samples were frequently misclassified as Calm or Neutral, primarily because of the substantial overlap among low-arousal emotional speech patterns. A notable observation was obtained for the Neutral class, where a high recall (0.81) was accompanied by a comparatively low precision (0.50). This result indicates a false-positive tendency, suggesting that the model frequently assigned ambiguous low-energy speech signals to the Neutral category. Furthermore, this behaviour was intensified by the class imbalance present in the dataset, which biased the decision boundaries toward classes with a larger number of training samples and consequently reduced the classification performance for the underrepresented Neutral class.

The discriminative performance of the models was further evaluated using Receiver Operating Characteristic (ROC) analysis. SVM model demonstrated consistently superior performance, with Area Under the Curve (AUC) values ranging from 0.98 to 0.99 across all emotion classes. These high AUC values indicate an excellent ability to discriminate between positive and negative classes over a wide range of decision thresholds.

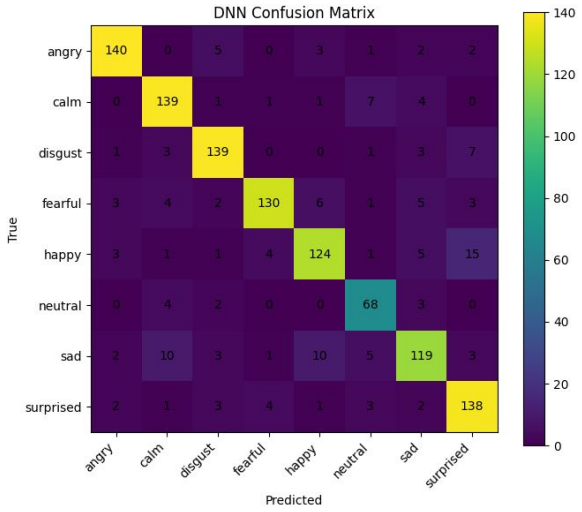

Figure 5 Confusion matrix of the DNN model for speech emotion recognition across eight emotional categories

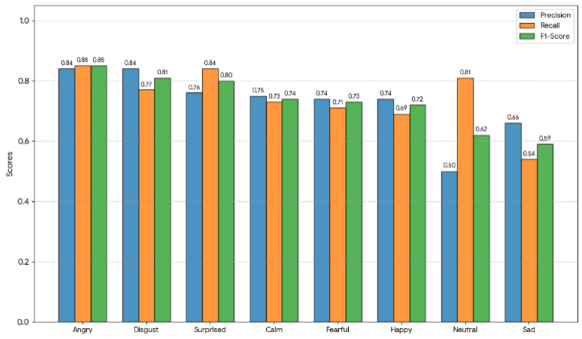

Figure 6 Comparative analysis of class-specific precision, recall, and F1-score for the DNN model

For the Random Forest model, although the Calm and Disgust classes achieved AUC values of 0.97, the AUC for the Sad class decreased to 0.91, indicating a reduced ability to distinguish acoustically similar emotional categories. The DNN model also achieved high overall AUC values; however, its discriminative performance for the Happy and Surprised classes was comparatively lower because of the similar high-arousal vocal characteristics shared by these emotions, resulting in increased inter-class confusion.

Overall, the experimental findings indicate that the SVM model provides the most consistent and balanced performance for speech emotion recognition. Although the DNN exhibited substantial learning capacity, its lower validation performance suggests that additional data augmentation or regularization strategies may be required to further improve generalization. The Random Forest model remained a competitive alternative but achieved lower overall performance for the selected acoustic feature set.

Ultimately, the results confirm that inter-class acoustic similarities, particularly between the Happy–Surprised and Sad–Calm emotion pairs, represent the primary limitation in achieving perfect classification accuracy. These findings suggest that the selection of an appropriate classification model should be guided by the acoustic characteristics and distribution of the target emotion classes to develop robust and reliable speech emotion recognition systems.

CONCLUSION

In this study, a comprehensive comparative evaluation was conducted to assess the performance of various machine learning and deep neural network architectures in speech emotion recognition. Specifically, SVM, RF, and DNN models were rigorously trained and analysed using a multidimensional feature set extracted from time, frequency, and cepstral domains. The experimental findings indicate that features derived from Mel-Frequency Cepstral Coefficients and Mel-spectrograms play a quintessential role in determining classification accuracy. The synergistic use of these spectral and cepstral representations allowed the models to more effectively characterise the complex interplay between the temporal dynamics and frequency components of emotional speech signals.

Regarding model performance, the SVM model achieved the best overall accuracy and F1-score. While the DNN model exhibited strong predictive performance, it showed limited generalisation, particularly when validation set results were compared with those from the training phase. In contrast, the Random Forest model consistently underperformed its counterparts, struggling to capture the non-linear relationships in the acoustic feature space. A granular analysis of the ROC curves on a class-by-class basis further validates these conclusions. The SVM model achieved exceptionally high and balanced Area Under the Curve values across all emotional categories, representing a robust discriminative threshold. While the DNN model also demonstrated substantial discriminative performance, it was notably lower for the happy and sad classes than for other emotional states. This performance degradation in specific categories can be theoretically attributed to the inherent acoustic similarities and overlapping pitch/energy profiles shared by these emotions. Furthermore, the simultaneous evaluation of the DNN model's training and validation metrics revealed a significant performance gap: although the model achieved near-perfect accuracy on the training data, it failed to maintain this level during validation, indicating a persistent overfitting problem.

This finding underscores that the model's architecture, while high in learning capacity, requires further refinement to improve generalisation to unseen data. Detailed confusion matrix analyses further corroborated this, showing a heightened frequency of confusion between happy-surprised and sad-calm pairs. These misclassifications suggest that certain emotional states remain difficult to isolate due to their closely related acoustic manifestations in the vocal signal. In conclusion, the strategic combination of diverse acoustic features has been identified as a critical factor in optimising speech emotion recognition systems. The study highlights that while the SVM model stands out for its robustness, the DNN model remains a high-capacity tool that requires advanced regularisation techniques, data augmentation, and more balanced datasets to overcome its current generalisation constraints. Future research will focus on integrating more sophisticated deep learning architectures and novel feature combinations to further refine the sensitivity and specificity of emotion recognition from speech (Kim and Kwak 2024).

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Funding

This research received no external funding.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

LITERATURE CITED

Akçay, M. B. and K. Oğuz, 2020 Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication **116**: 56–76.

Begazo, R., A. Aguilera, I. Dongo, and Y. Cardinale, 2024 A combined cnn architecture for speech emotion recognition. Sensors **24**.

Brahmi, Z., M. Mahyoob, M. Al-Sarem, J. Algaraady, K. Bousselmi, *et al.*, 2024 Exploring the role of machine learning in diagnosing and treating speech disorders: A systematic literature review. Psychology Research and Behaviour Management **17**: 2205–2232.

Çolakoğlu, E., S. Hızlısoy, and R. S. Arslan, 2022 Konuşmadan duygu tanıma üzerine detaylı bir inceleme: Özellikler ve sınıflandırma metotları. European Journal of Science and Technology .

Cortes, C. and V. Vapnik, 1995 Support-vector networks. Machine Learning **20**: 273–297.

Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, *et al.*, 2001 Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine **18**: 32–80.

Cui, W., F. Jiang, X. Gao, S. Zhang, and D. Zhao, 2018 An efficient deep quantized compressed sensing coding framework of natural images. In *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 1777–1785.

Dixit, S., D. M. Low, G. Elbanna, F. Catania, and S. S. Ghosh, 2024 Explaining deep learning embeddings for speech emotion recognition by predicting interpretable acoustic features. arXiv preprint arXiv:2409.09511 .

Durukal, M. and A. K. Hocaoglu, 2015 Performance analysis of mfcc features on emotion recognition from speech. International Journal of Scientific and Technological Research **1**.

Fayek, H. M., M. Lech, and L. Cavedon, 2017 Evaluating deep learning architectures for speech emotion recognition. Neural Networks **92**: 60–68.

Gales, M. J. F., K. M. Knill, and S. J. Young, 1999 State-based gaussian selection in large vocabulary continuous speech recognition using hmms. IEEE Transactions on Speech and Audio Processing **7**: 152–161.

Hsu, C.-C., J. Krajewski, and J. Felfe, 2024 How speech prosody affects leadership perceptions: A machine learning approach. International Journal of Organizational Leadership **13**: 432–450.

Jha, T., R. Kavya, J. Christopher, and V. Arunachalam, 2022 Machine learning techniques for speech emotion recognition using paralinguistic acoustic features. International Journal of Speech Technology **25**: 707–725.

Jiang, X. and M. D. Pell, 2018 Predicting confidence and doubt in accented speakers: Human perception and machine learning experiments. In *Proceedings of the International Conference on Speech Prosody*, pp. 269–273.

Juslin, P. N. and P. Laukka, 2003 Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin **129**: 770–814.

Kane, J., M. N. Johnstone, and P. Szewczyk, 2024 Voice synthesis improvement by machine learning of natural prosody. Sensors **24**: 1624.

Kim, T. W. and K. C. Kwak, 2024 Speech emotion recognition using deep learning transfer models and explainable techniques. Applied Sciences **14**.

Livingstone, S. R. and F. A. Russo, 2018 The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PLOS ONE **13**: e0196391.

McLaren, M., A. Lawson, Y. Lei, and N. Scheffer, 2013 Adaptive gaussian backend for robust language identification. In *Interspeech 2013*, pp. 84–88.

Ming, Y., B. Lyu, and Z. Li, 2023 Cast: Context-association architecture with simulated long-utterance training for mandarin speech recognition. Speech Communication **155**: 102985.

Mustaqeem and S. Kwon, 2019 A cnn-assisted enhanced audio signal processing for speech emotion recognition. Sensors **20**: 183.

Pala, M. A., 2025 Specedgenet: A hybrid graph neural network architecture integrating information from spectrum to bond for drug property prediction. In *Proceedings of the 2025 Innovations in Intelligent Systems and Applications Conference (ASYU 2025)*.

Serrano, S., O. Serghini, G. Esposito, S. Carbone, C. Mento, *et al.*, 2025 Review and comparative analysis of databases for speech emotion recognition. Data **10**.

Shah, N., K. Sood, and J. Arora, 2023 Speech emotion recognition for psychotherapy: an analysis of traditional machine learning and deep learning techniques. In *Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 718–723.

Tao, J. and T. Tan, 2005 Affective computing: A review. In *International Conference on Affective Computing and Intelligent Interaction*.

Ullah, R., M. Asif, W. A. Shah, F. Anjam, I. Ullah, *et al.*, 2023 Speech emotion recognition using convolution neural networks and multi-head convolutional transformer. Sensors **23**: 6212.

Wu, C.-H., J.-F. Yeh, Z.-J. Chuang, and C. Yi, 2002 Emotion perception and recognition from speech. LDC Catalog.

Yılmazcan, D. S. and M. A. Pala, 2026 Exploring the chemical space of bace-1 inhibitors: Structure-based prediction with deep learning and machine learning. Computers and Electronics in Medicine **3**: 36–41.

How to cite this article: Baydar, S. E., and Pala M. A. Efficacy of Multi-Domain Acoustic Features in Speech Emotion Recognition: A Comparative Analysis of Machine Learning and Deep Learning Models. *Chaos and Fractals*, 3(2), 117-124, 2026.