Prosodic Spatio-Temporal Feature Fusion with Attention Mechanisms for Speech Emotion Recognition

by Kristiawan Nugroho

Submission date: 04-Aug-2025 07:27PM (UTC+0700)
Submission ID: 2725153263
File name: Kris-Computers.docx (690.78K)
Word count: 6076
Character count: 38035

computers

MDPI

Article

Prosodic Spatio-Temporal Feature Fusion with Attention Mechanisms for Speech Emotion Recognition

Kristiawan Nugroho[1,*], Imam Husni Al Amin[2], Nina Anggraeni Noviasari[3], and De Rosal Ignatius Moses Setiadi[4]

[1] Department of Information Technology, Faculty of Information Technology and Industry, Universitas Stikubank, Semarang 50241, Indonesia; kristiawan@edu.unisbank.ac.id
[2] Department of Industrial Engineering, Faculty of Information and Industrial Technology, Universitas Stikubank, Semarang 50241, Indonesia; imam@edu.unisbank.ac.id
[3] Faculty of Medicine, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia; ninaanggraeni@unimus.ac.id
[4] Research Center for Quantum Computing and Materials Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, 50131, Indonesia; moses@dsn.dinus.ac.id
* Correspondence: Kristiawan Nugroho

Abstract

Speech Emotion Recognition (SER) plays a vital role in supporting applications such as healthcare, human–computer interaction, and security. However, many existing approaches still face challenges in achieving robust generalization and maintaining high recall, particularly for emotions related to stress and anxiety. This study proposes a dual-stream hybrid model that combines prosodic features with spatio-temporal representations derived from the Multitaper Mel-Frequency Spectrogram (MTMFS) and the Constant-Q Transform Spectrogram (CQTS). Prosodic cues, including pitch, intensity, jitter, shimmer, HNR, pause rate, and speech rate, were processed using dense layers, while MTMFS and CQTS features were encoded with CNN and BiGRU. A Multi-Head Attention mechanism was then applied to adaptively fuse the two feature streams, allowing the model to focus on the most relevant emotional cues. Evaluations conducted on the RAVDESS dataset with subject-independent 5-fold cross-validation demonstrated an accuracy of 97.64% and a macro F1-score of 0.9745. These results confirm that combining prosodic and advanced spectrogram features with attention-based fusion improves precision, recall, and overall robustness, offering a promising framework for more reliable SER systems.

Keywords: Speech Emotion Recognition; Prosodic Features; Multitaper Mel-Frequency Spectrogram; Constant-Q Transform; Attention Mechanism.

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

Citation: To be added by editorial staff during production.

1. Introduction

Speech Emotion Recognition (SER) has emerged as a significant research area over the last two decades. Emotions play a crucial role in human communication, influencing social interactions, decision-making, and mental health conditions. SER systems offer broad opportunities in various applications, ranging from human-computer interaction (HCI) and technology-based education to mental health services, security systems, and intelligent vehicles [1–3]. Previous research has shown that speech signals contain two main components: linguistic information that conveys literal meaning, and paralinguistic information that reflects emotional aspects [1]. Paralinguistic information is considered

more relevant in identifying emotional states such as stress and anxiety because it is relatively independent of lexical and linguistic content. This makes paralinguistic-based SER very promising for detecting psychological conditions, especially those related to stress and anxiety.

Prosody, which encompasses variations in pitch, intensity, pauses, syllable duration, and rhythm of speech, has been recognized as an important indicator in emotion analysis [4]. Kuuluvainen et al. [5] asserted that prosody plays a role in facilitating the understanding of statistical patterns in the speech stream, thereby enhancing language learners' ability to capture hidden structures. In the context of SER, prosodic cues have been shown to significantly contribute to the model's ability to distinguish emotions, particularly those related to stress and anxiety. Shan [6] noted that intonation, stress, and rhythm have a direct impact on the interpretation of a speaker's intent and emotions, making prosody a crucial element in understanding conversational dynamics. Furthermore, research [7] shows that prosodic variations are often the most consistent nonverbal cues when someone is experiencing emotional distress. For example, higher pitch, increased speech rate, jitter, and unstable shimmer are often associated with stress or anxiety.

SER methodology has evolved rapidly, from classical machine learning (ML) approaches to deep learning. Traditional approaches typically rely on hand-crafted features such as Mel-Frequency Cepstral Coefficients (MFCC) [8–11], Linear Predictor Coefficients (LPC) [12,13], and prosodic features [14,15] processed through classification models such as Support Vector Machine (SVM) [16], Hidden Markov Models (HMM) [17,18], or Random Forest (RF) [19,20]. While quite effective on small and clean datasets, these approaches often fail to generalize to real-world environments with varying speakers, accents, and background conditions. With the advent of deep learning approaches, such as CNNs [21–23], RNNs [22,24], and Transformers [25,26], the accuracy of SER has significantly improved. CNNs are effective in extracting spectral patterns from speech spectrograms, while RNNs, such as GRUs and LSTMs, are capable of capturing long-term temporal dependencies [27–29]. Furthermore, the integration of attention mechanisms has been shown to help models focus more on emotion-relevant signal regions, thus improving performance [29–32].

With the development of deep learning-based methods, the use of spectrograms has become a popular approach for extracting speech acoustic patterns. One commonly used spectrogram is the Log-Mel Spectrogram; however, this representation still faces limitations in adaptive resolution and sensitivity to spectral leakage, which can reduce the model's ability to distinguish emotions with similar acoustic characteristics [33,34]. Therefore, this study hypothesizes that the Multitaper Mel-Frequency Spectrogram (MTMFS) and Constant-Q Transform Spectrogram (CQTS) can provide richer representations than conventional spectrograms [1,35,36]. The MTMFS is expected to produce a more stable and detailed spectrum, while the CQTS is believed to capture low-frequency variations with higher resolution and fast temporal dynamics at high frequencies, thereby improving the accuracy of emotion detection, particularly in low-arousal and high-arousal classes.

Although various approaches have made significant progress, SER still faces several key challenges. First, the problem of overfitting on small datasets, which often occurs due to the limited number of available samples [37]. Second, the lack of in-depth exploration of prosodic features, despite prosody having long been recognized as a key indicator of stress. Third, research tends to prioritize overall accuracy over recall, which can result in missing cases of mental health-related emotions. Finally, the lack of multi-modal feature integration, as most studies still rely on a single feature type, results in an incomplete representation of complex emotional information.

Recall is one of the most crucial metrics in SER, reflecting the model's ability to detect all relevant positive cases. Achieving high recall is crucial to ensure that no emotional

samples are missed, especially for emotions that are difficult to recognize. In the context of stress and anxiety detection, recall becomes even more vital because incorrectly detecting individuals experiencing stress (false negatives) can have a direct impact on a person's psychological well-being, risking more than false positives [38–40]. Unfortunately, most SER research still focuses on overall accuracy rather than maintaining optimal recall [29–32]. Several studies, such as [2,27,28] report that recall is often lower than accuracy, especially for emotion classes related to stress and anxiety.

To address this challenge, this study proposes a hybrid dual-stream architecture that combines prosodic modulation and spatio-temporal features based on MTMFS and CQTS. The prosodic branch is designed to capture the temporal dynamics of prosody through dense layers, while the spectrogram branch uses CNN and BiGRU to extract spatio-temporal patterns from MTMFS and CQTS. The results from both branches are then fused through a Multi-Head Attention Fusion mechanism, which adaptively weights the most relevant features. With this design, this study contributes to improving recall without compromising overall accuracy, thereby providing more reliable emotion detection, especially under stressful and anxious conditions.

2. Related Works

Currently, research related to SER has been conducted using various feature and model approaches. Previously, many studies relied on conventional acoustic features such as MFCC, LPC, and prosodic cues, which were then processed using classical ML classification algorithms. Research [2] employed MFCC on the RAVDESS dataset with SVM, achieving an accuracy of approximately 82%. However, this method is generally limited in its applicability and sensitive to noise, resulting in low recall performance for high-arousal emotions.

Advances in deep learning allow for richer feature representation. Research [41] proposed a Multi-Scale Feature Pyramid Network (MSFPN), which combines a Multi-Scale CNN (MSCNN) with Convolutional Self-Attention (CSA) and BiLSTM to capture temporal context. Evaluations were conducted on the IEMOCAP and RAVDESS datasets. The results showed an Unweighted Accuracy (UA) of 86.5% on RAVDESS. This approach excels because it preserves multi-granularity information while improving local correlations between features through CSA. Furthermore, despite the relatively good accuracy performance, these studies still focused on UA and WA without an in-depth analysis of recall within specific emotional classes.

Multi-feature fusion-based approaches were also developed. Bhangale et al. [1] introduced the Parallel Emotion Network (PEmoNet), which combines MTMFS, Gammatonegram (GS), and CQTS. Evaluations on EMODB and RAVDESS demonstrated accuracy of up to 97% with an average F1-score of 0.97. Ablation studies demonstrated that MTMFS and CQTS contributed significantly to performance improvements compared to using a single spectrogram type alone. However, using all three spectrograms simultaneously increases computational complexity, making the integration more challenging.

Other studies have begun to emphasize the importance of prosodic features. Kuuluvainen et al. [5] showed that variations in pitch, intonation, and pauses can facilitate the learning of statistical dependencies in continuous speech. Meanwhile, Guo et al. [7] applied a prosody- and spectrogram-based dual-stream architecture to a Mandarin dataset, and successfully improved the sensitivity of negative emotion detection. However, recall results still varied across classes, so the risk of false negatives for stress and anxiety emotions remained high.

The application of attention mechanisms has also been shown to improve performance. Makhmudov et al. [32] developed a hybrid CNN–LSTM model with attention, using RMS, ZCR, and MFCC features. Evaluation on TESS and RAVDESS yielded

accuracies of up to 99.8% and 95.7%, respectively. The attention mechanism allows the model to focus on significant emotional segments, while the CNN handles spectral representation and the LSTM captures long-term temporal dependencies. However, this research is still limited to conventional acoustic features and tends to emphasize global accuracy.

Bhanbhro et al. [29] compared CNN-LSTM with Attention-Enhanced CNN-LSTM on the RAVDESS dataset. The use of attention improved accuracy by more than 2% compared to standard CNN-LSTM. The attention mechanism was demonstrated to enhance the separation between similar emotion classes while maintaining performance under noisy conditions. However, this research was still limited to Mel-spectrogram-based spectral features, without the integration of prosodic cues, which are important for detecting stress and anxiety.

Although various SER studies have achieved high accuracy, significant limitations remain. Conventional feature-based approaches (MFCC, LPC) are sensitive to noise and less effective in capturing the dynamics of high-arousal emotions, such as fear and anger. Multi-scale deep learning methods enhance representation but often prioritize global accuracy over recall. Multi-feature fusion studies have demonstrated the dominant contribution of MTMFS and CQTS through ablation studies. Meanwhile, attention mechanism-based models improve accuracy and class separation, but are still limited to spectral features without the integration of prosodic cues. Therefore, methods that integrate prosodic cues, spatio-temporal fusion of MTMFS and CQTS, and adaptive attention mechanisms are needed to improve recall and robustness, especially for stress and anxiety detection.

3. Proposed Method

The method proposed in this study aims to integrate prosodic and spectral features into a richer spatio-temporal representation, thereby improving emotion detection performance. The RAVDESS dataset was chosen as the experimental benchmark because it provides a corpus of emotional conversations under controlled conditions with professional actors. The dataset comprises eight distinct emotion classes, expressed by 24 speakers with both male and female voices, ensuring a diverse range of gender and vocal characteristics.

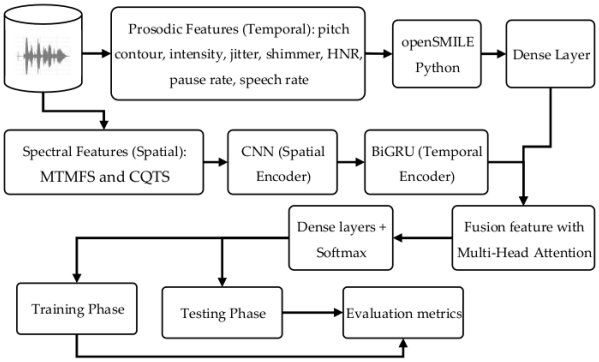

Figure 1. Proposed dual-stream hybrid.

Importantly, RAVDESS offers emotions with both low and high arousal, making it well-suited for testing methods designed to improve recall in stress and anxiety detection scenarios. To address this challenge, the system is designed with a hybrid dual-stream architecture comprising a prosodic feature branch and a spectrogram feature branch, which are then combined through an attention-based fusion mechanism.

3.1 Prosodic Features

Prosodic features were extracted from the RAVDESS dataset (48 kHz, mono) using the openSMILE Python library. The extracted features included pitch contour, intensity, jitter, shimmer, Harmonics-to-Noise Ratio (HNR), pause rate, and speech rate. Each segment yielded approximately 40–60 prosodic features. These features were processed through stacked dense layers (128 → 64 units, ReLU activation, dropout 0.3), resulting in a 64-dimensional representation vector.

3.2 Spatio-Temporal Features

Spatial features were extracted using a parallel representation of the Multitaper Mel-Frequency Spectrogram (MTMFS) and the Constant-Q Transform Spectrogram (CQTS) to enhance the discriminative power of emotional cues. MTMFS employed multiple orthogonal tapers instead of a single Hamming window, reducing spectral leakage and capturing subtle variations in pitch, timbre, and intonation. A 64-point Mel filterbank with a 25 ms frame size, 10 ms hop length, and a 2048-point FFT was applied, providing a stable and high-resolution spectral envelope. CQTS offered an adaptive time–frequency resolution, providing superior frequency resolution at lower frequencies (suitable for low-arousal emotions, such as sadness and calmness) and higher temporal resolution at higher frequencies (critical for high-arousal emotions, such as anger and surprise).

Each spectrogram was processed independently using a CNN-based spatial encoder with 3×3 kernels to capture local time–frequency patterns. Each CNN comprised three convolutional blocks with filters of [32, 64, 128], each followed by Batch Normalization, ReLU activation, and Max Pooling. Instead of full flattening, the CNN outputs were retained as sequence embeddings (time steps × feature dimension) to preserve temporal information. The sequence embeddings from MTMFS and CQTS were then individually passed through a BiGRU-based temporal encoder (2 layers, 128 hidden units each, dropout 0.3), allowing each branch to capture its long-term temporal dependencies.

The BiGRU outputs were subsequently concatenated and normalized using Layer Normalization to align feature scales and stabilize recurrent learning. This design enables the model to leverage both stable spectral envelopes from MTMFS and adaptive time–frequency dynamics from CQTS in a balanced manner. Formally, the spatio-temporal representation is given by:

$$H_t = \text{LayerNorm}\left(\text{Concat}\left(\text{BiGRU}\left(\text{CNN}\left(X_{MTMFS}\right)\right), \text{BiGRU}\left(\text{CNN}\left(X_{CQTS}\right)\right)\right)\right) \quad (1)$$

where X_{MTMFS} and X_{CQTS} denote the MTMFS and CQTS matrices of dimension $T \times F$.

3.3 Fusion Configuration

The prosodic embedding and the spatio-temporal embedding from the BiGRU were combined through a Multi-Head Attention (MHA) layer to weight the most relevant features for emotion recognition adaptively. Prior to fusion, the prosodic embedding was normalized using Layer Normalization, aligning its scale with the spatio-temporal embedding to ensure balanced contribution from both streams. This step is particularly crucial since prosodic features and spectrogram-derived embeddings differ in dimensionality and statistical distribution. An MHA module then processed the normalized embeddings with eight attention heads, which provided the best balance between recall improvement and computational efficiency. Formally, the fused vector can be expressed as:

$$Z = \text{Concat}\left(\text{Attention}(Q,K,V)\left(\text{LayerNorm}(E_{prodody})\right), \text{Attention}(Q,K,V)\left(E_{spectogram}\right)\right) \quad (2)$$

Where $E_{prodody}$ and $E_{spectogram}$ denote the normalized prosodic embedding and the spatio-temporal embedding, respectively, and Q,K,V denote the query, key, and value vectors of the respective feature branches.

3.4 Classifier

The fused representation was passed through fully connected layers [256, 128, 64] with ReLU activation and dropout 0.5. The final layer employed a Softmax function to classify eight emotional classes as defined by RAVDESS. The loss function used was Categorical Cross-Entropy:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (3)$$

where C is the number of classes, y_i the true label, and $\hat{y}_i$ the predicted probability.

3.5 Training Strategy

The model was trained using the Adam optimizer with a learning rate of 1e-4, a batch size of 32, and a maximum of 100 epochs. Early stopping was applied with a patience of 10 epochs, monitoring validation macro F1-score and validation loss to prevent overfitting and ensure balanced recall and precision across classes. This strategy allows the model to halt training when further improvements on the validation set become marginal, reducing the risk of overfitting while maintaining generalization. To further enhance robustness, a 5-fold subject-independent cross-validation was conducted, ensuring that speakers in the training set did not appear in the testing set. This evaluation protocol provides a more reliable estimate of the model's performance across different speakers.

Table 1. Configuration for the proposed model.

Component	Proposed Configuration
Prosodic Features	40–60 features (pitch, jitter, shimmer, intensity, HNR, pause rate, speech rate)
Prosody Encoder	Dense layers (128 → 64), ReLU activation, Dropout 0.3, Layer Normalization
Spectrogram Input	Parallel MTMFS (64 Mel filters, 25 ms frame size, 10 ms hop, FFT=2048) and CQTS (84 bins, 12 bins)
CNN Encoder	3 convolutional blocks (filters [32, 64, 128], kernel size 3×3, BatchNorm, ReLU, MaxPooling)
Temporal Encoder	BiGRU per branch (2 layers, 128 hidden units, Dropout 0.3) → Concatenation → Layer Normalization
Fusion Layer	Multi-Head Attention (8 heads); prosodic embedding normalized prior to fusion
Classifier	Dense layers [256, 128, 64], ReLU activation, Dropout 0.5, Softmax output (8 classes)
Optimizer	Adam (learning rate = 1e-4)
Training Strategy	100 epochs, Batch size 32, Early Stopping (patience=10, monitoring validation loss & macro F1), 5-fold subject-independent CV
Evaluation Metrics	Precision, Recall, F1-score, Accuracy

3.6 Evaluation Metrics

The evaluation metrics included precision, recall, specificity, F1-score, AUC, and accuracy.

- Accuracy measures the proportion of correctly classified samples but may be misleading for imbalanced emotional classes.
- Precision indicates the fraction of correctly predicted positive samples among all predicted positives, ensuring the reliability of stress predictions.
- Recall (sensitivity) measures the proportion of correctly detected positive cases; in stress and anxiety detection, recall is critical since missing stressed cases (false negatives) can be more harmful than false positives.
- F1-score, the harmonic mean of precision and recall, balances sensitivity and reliability.

4. Results and Discussion

The dataset used in this study is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), a widely used resource in Speech Emotion Recognition (SER) research. This dataset comprises 24 actors (12 men and 12 women) who voiced speech in eight emotion categories: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. Each emotion was recorded at two intensity levels: normal and strong, resulting in a total of 1,440 records. The data distribution consisted of seven classes, with one other class being a minority (see Figure 2 for further details).

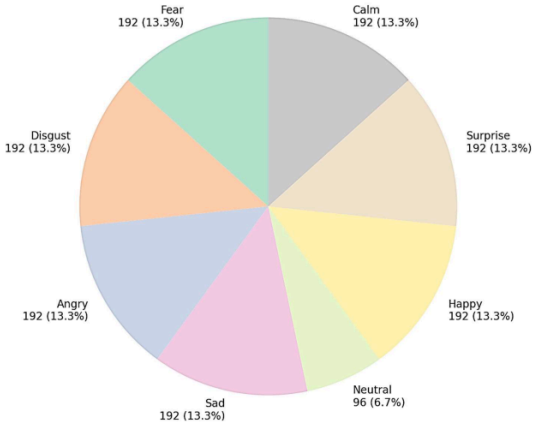

Figure 2. Class distribution of the RAVDESS dataset.

4.1. Prosodic Features

Figure 3 displays a sample of prosodic extraction results, including pitch, intensity, HNR, loudness, jitter, shimmer, pause rate, and speech rate. The observed patterns illustrate acoustic dynamics over time. Pitch and intensity exhibit more fluctuating contours in segments with high arousal, such as fearful and angry emotions. HNR decreases in certain sections, which is typically associated with decreased voice quality resulting from vocal tension. Jitter and shimmer are relatively higher in segments with low vocal stability, which often occur in states of anxiety. Pause and speech rate reflect the rhythm of speech; fewer pauses and a faster speech rate are seen in emotionally intense sections.

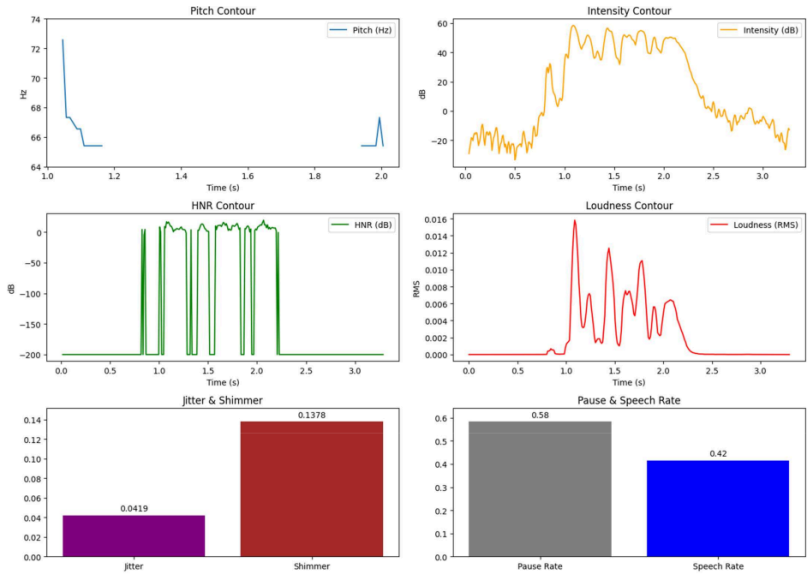

Figure 3. Example of extracted prosodic features from a speech segment, including pitch contour, intensity contour, HNR contour, loudness contour, jitter, shimmer, pause rate, and speech rate.

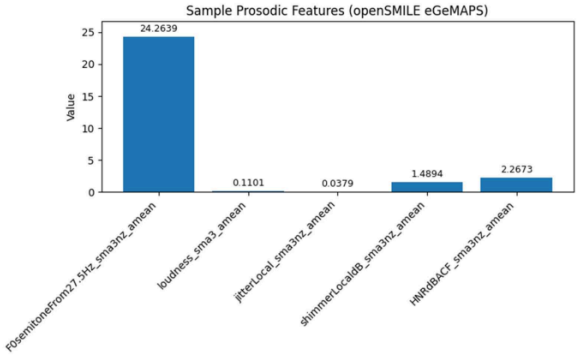

Figure 4. Sample plot Prosodic Features using openSMILE.

This relationship is reinforced by Figure 4, which presents a statistical summary of prosodic features with openSMILE eGeMAPS. Loudness values are relatively higher in

segments with high arousal, while jitter, shimmer, and HNR show variations that are in line with the temporal pattern in Figure 3. Thus, the values and patterns shown by Figures 3 and 4 reinforce each other, indicating that prosodic features can be consistent markers in distinguishing certain emotional states, especially those related to stress and anxiety.

4.2. Spectrogram Features

Figure 5 shows an example of MTMFS and CQTS for a voice sample in RAVDESS. MTMFS, shown on the left, produces a relatively smooth and stable representation with a clear energy distribution across the frequency range. This multitaper approach effectively reduces spectral leakage and improves the accuracy of identifying subtle variations in pitch, timbre, and intonation, which are highly relevant for distinguishing similar emotions. Meanwhile, CQTS on the right displays a more adaptive frequency pattern, characterized by high frequency resolution in low tones and high temporal resolution in high tones. This enables CQTS to capture the subtle changes associated with low-arousal emotions, such as sadness or calm, while also representing the rapid dynamics common to high-arousal emotions, like anger or surprise. The combination of these two representations is expected to enrich the mapping of emotional features in speech signals, thereby improving the model's accuracy in emotion recognition compared to using a single spectrogram.

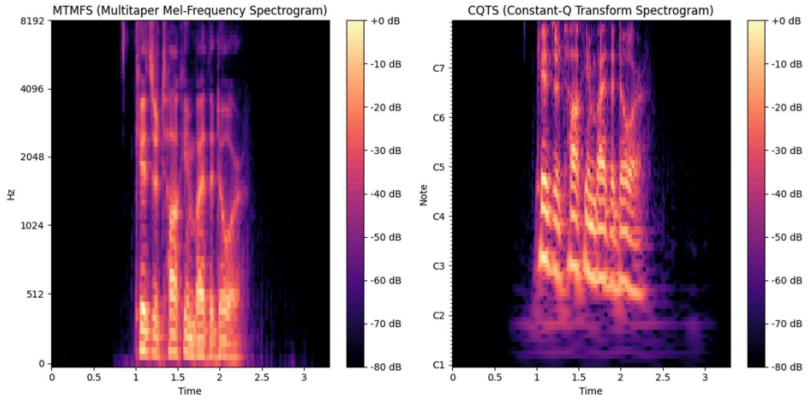

Figure 5. Sample plot of Spectrogram (**a**) MTMFS; (**b**) CQTS.

4.3. Results

To comprehensively evaluate the performance of the proposed model, a confusion matrix representing the classification results of eight emotion classes on the RAVDESS dataset was used. The values displayed in this confusion matrix are the aggregated results of 5-fold cross-validation, providing a more stable picture and reducing bias caused by certain data splits. With this approach, each fold alternates as test data, while the other folds serve as training data, ensuring that all samples contribute to both the training and testing phases. The confusion matrix in Figure 5, the result of this aggregation, provides detailed information on the distribution of correct predictions and misclassifications for

each emotion class, allowing for an in-depth analysis of the model's strengths and weaknesses in each emotion category.

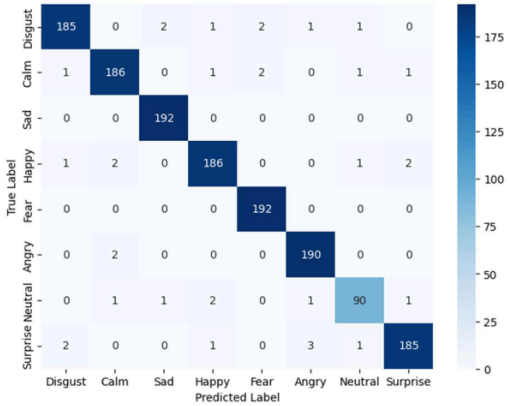

Figure 5. Confusion matrix proposed method.

In general, the prediction distribution exhibits a dominant diagonal, indicating high accuracy across almost all classes. The Sad, Fear, and Calm classes show nearly perfect predictions with few misclassifications. However, some confusion is observed in the Neutral class, which is sometimes predicted as Happy or Calm. Similarly, a small number of Disgust and Surprise cases are swapped, although the numbers are relatively small. This pattern suggests that the model is quite reliable in recognizing emotions with explicit acoustic expressions, while confusion still occurs in classes with more subtle prosodic features, such as the Neutral class. For more details, see Table 2 for differences in precision, recall, and f1-score values for each class.

Table 2. Classification Report of the Proposed Model on the RAVDESS Dataset.

Class	Precision	Recall	F1-Score	Support
Disgust	0.9788	0.9635	0.9711	192
Calm	0.9738	0.9688	0.9713	192
Sad	0.9846	1.0000	0.9922	192
Happy	0.9738	0.9688	0.9713	192
Fear	0.9796	1.0000	0.9897	192
Angry	0.9744	0.9896	0.9819	192
Neutral	0.9574	0.9375	0.9474	96
Surprise	0.9788	0.9635	0.9711	192
Accuracy	-	-	0.9764	1440
Macro Avg	0.9752	0.9740	0.9745	1440
Weighted Avg	0.9763	0.9764	0.9763	1440

Table 2 shows that the model's performance demonstrates high consistency across classes, with an average F1-score approaching 0.98. Two classes, Sad and Fear, achieved perfect recall (1.0000), indicating the model's ability to detect these emotions without

losing relevant samples. This indicates the model's reliability in recognizing both low- and high-intensity emotions, which are often challenging in SER. Meanwhile, the Neutral class achieved the lowest recall value, which, while still a good performance considering its minority, suggests that its more subtle prosodic characteristics make it relatively difficult to distinguish compared to other classes. These findings confirm that integrating prosodic features with MTMFS and CQTS successfully strengthens the model's generalization, especially for classes that are acoustically susceptible to confusion, while maintaining a balance between precision and recall across emotion categories.

Furthermore, to assess the contribution of each component in the proposed architecture, an ablation study was conducted by removing or replacing specific parts of the model. The primary goal of this ablation was to ensure that performance improvements stem not solely from model complexity but from the integration of the designed features and mechanisms. Specifically, this study focuses on four aspects: (i) the role of prosodic features in increasing sensitivity to stress and anxiety emotions, (ii) the contribution of MTMFS and CQTS spectral representations compared to using only one type of spectrogram, and (iii) the effectiveness of the attention mechanism in balancing the contributions of prosody and spatio-temporal features. The results of this study are expected to clarify the relative role of each component in achieving increased recall and robustness of the model.

Table 3. Ablation study of the proposed method.

Study	Accuracy	Precision	Recall	F1
Without prosody features	93.78	94.01	93.69	93.85
Without MTMFS features	93.23	93.42	93.23	93.32
Without CQTS features	95.18	95.25	95.18	95.21
Without attention mechanism	95.53	95.67	95.57	95.61
Proposed (full)	97.64	97.63	97.64	97.63

Table 3 presents the results of an ablation study to assess the contribution of each component in the proposed architecture. It can be seen that removing prosodic features decreased performance to 93.78% accuracy and 93.85% F1-score, indicating that prosody plays a significant role in detecting both high- and low-arousal emotions. This is consistent with the hypothesis that prosodic cues provide additional sensitivity to stress and anxiety dynamics. Removing MTMFS features had the most significant impact, resulting in a 93.32% decrease in F1-score. This finding confirms MTMFS's dominant role in enriching spectral representation and reducing spectral leakage, consistent with previous studies that identified MTMFS as the most stable representation in the RAVDESS dataset. Meanwhile, removing CQTS features resulted in a more moderate performance reduction (95.21% F1-score), indicating a significant but less significant contribution than MTMFS. This can be explained by CQTS being more prominent at low-frequency resolution, while RAVDESS is relatively rich in prosodic and mid-spectral variations.

Removing the attention mechanism also had a significant impact (F1-score 95.61%). Although CNN and BiGRU were still able to capture spatial and temporal patterns, the lack of attention caused an imbalance in the contributions between features, resulting in decreased recall in the minority class. Overall, the ablation results showed that:

- MTMFS makes the largest contribution to classification stability and accuracy,
- Prosodic features directly improve recall in the stress and anxiety classes,
- CQTS adds depth to the representation but with a more moderate impact,
- The attention mechanism ensures adaptive integration between features, maintaining a balance between precision and recall.

The best performance of the proposed method (full) shows that the combination of prosodic cues, MTMFS, CQTS, and the attention mechanism synergistically contributes to achieving optimal generalization.

After confirming the contribution of each component through ablation studies, the next step was to compare the performance of the proposed model with previous research on the RAVDESS dataset. Table 4 summarizes the comparison results with several state-of-the-art models.

Table 4. Results of proposed model and comparison with prior works.

Study	Accuracy	Precision	Recall	F1
SVM [16]	72.40	72.20	72.10	-
HuBERT + DPCNN + CAF [30]	81.86	-	-	82.84
K-SVM + GWO [28]	87.00	88.00	85.00	86.00
1D CNN + Feature Fusion [2]	91.90	90.50	91.10	90.80
CNN+LSTM [32]	95.70	93.49	94.99	94.20
MTMFS + GS + CQTS+ PEmoNet [1]	97.41	97.53	97. 53	97.26
Ours	97.64	97.63	97.64	97.63

The results in Table 4 show the performance improvement of SER on the RAV-DESS dataset using various approaches. The SVM-based methods [16] and K-SVM + GWO [28] performed quite well in the classical machine learning category, with accuracies of 72.40% and 87.00%, respectively. The HuBERT + DPCNN + CAF approach [30] achieved an F1-score of 82.84%, confirming the potential of self-supervised learning-based representations. Model [2] achieved an F1-score of 90.80%, demonstrating the effectiveness of feature fusion. Research [32] further improved recall to 94.99% with an F1-score of 94.20, thanks to the LSTM's ability to capture long-term temporal dependencies.

A study of MTMFS + GS + CQTS with PEmoNet [1] demonstrated competitive results, achieving an F1-score of 97.26, which highlights the power of multi-spectrogram fusion in enriching emotion representation. The model proposed in this study achieved the highest performance. This achievement indicates that the integration of prosody features, MTMFS and CQTS fusion, and attention mechanisms can provide a better balance between precision and recall, particularly since recall values are superior to other key metrics in supporting more reliable emotion detection, especially for stress and anxiety.

5. Conclusions

This study proposes a dual-stream hybrid architecture that integrates prosodic features with spatio-temporal representations from the MTMFS and the CQTS, combined through a Multi-Head Attention mechanism. The experimental results on the RAVDESS dataset using 5-fold subject-independent cross-validation demonstrated that the proposed model consistently outperformed state-of-the-art approaches, achieving an overall accuracy of 97.64% and a macro F1-score of 0.9745. More importantly, the model achieved a recall of 97.64%, higher than previous studies that often prioritized accuracy at the expense of recall. This improvement confirms the effectiveness of integrating prosodic cues with multi-spectrogram fusion to enhance sensitivity, particularly in detecting stress- and anxiety-related emotions, where missing positive cases is highly detrimental.

Author Contributions: Conceptualization, K.N. and D.R.I.M.S.; methodology, K.N.; software, K.N. and I.H.A.; validation, K.N., I.H.A. and N.A.N.; formal analysis, K.N. and I.H.A.; investigation, K.N. and N.A.N.; resources, K.N.; writing—original draft preparation, K.N.; writing—review and editing, D.R.I.M.S.; visualization, I.H.A.; supervision, K.N.; project administration, D.R.I.M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this study are publicly available. The experiments in this research were conducted using the RAVDESS Emotional Speech Audio dataset, which can be accessed at https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio. No additional new data were created or analyzed in this study.

Acknowledgments: The authors would like to express their sincere gratitude to the Ministry of Higher Education, Science, and Technology (Kemdiktisaintek) of the Republic of Indonesia for providing financial support for this research with grant number 127/C3/DT.05.00/PL/2025, 026/LL6/AL.04/2025, and 081/DPPMP/UNISBANK/UM/VI/2025. The authors also extend their appreciation to the Research and Community Service Institute (LPPM) of Universitas Stikubank for the administrative and technical assistance. During the preparation of this manuscript, the authors used generative AI to assist in improving the clarity and structure of the writing. The ideas, study design, data analysis, and interpretation remain the sole responsibility of the authors, who have thoroughly reviewed and edited the final content.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bhangale, K.B.; Kothandaraman, M. Speech Emotion Recognition Using the Novel PEmoNet (Parallel Emotion Network). *Appl. Acoust.* **2023**, *212*, 109613, doi:10.1016/j.apacoust.2023.109613.
2. Waleed, G.T.; Shaker, S.H. Speech Emotion Recognition on MELD and RAVDESS Datasets Using CNN. *Information* **2025**, *16*, 518, doi:10.3390/info16070518.
3. Maulisa Liztio, L.; Atika Sari, C.; Ignatius Moses Setiadi, D.R.; Hari Rachmawanto, E. Gender Identification Based on Speech Recognition Using Backpropagation Neural Network. In Proceedings of the 2020 International Seminar on Application for Technology of Information and Communication (iSemantic); IEEE, September 19 2020; pp. 88–92.
4. Guo, X.; Mai, G.; Mohammadi, Y.; Benzaquén, E.; Yukhnovich, E.A.; Sedley, W.; Griffiths, T.D. Neural Entrainment to Pitch Changes of Auditory Targets in Noise. *Neuroimage* **2025**, *314*, 121270, doi:10.1016/j.neuroimage.2025.121270.
5. Kuuluvainen, S.; Kaskivuo, S.; Vainio, M.; Smalle, E.; Möttönen, R. Prosody Enhances Learning of Statistical Dependencies from Continuous Speech Streams in Adults. *Cognition* **2025**, *262*, 106169, doi:10.1016/j.cognition.2025.106169.
6. Shan, Y. Prosodic Modulation of Discourse Markers: A Cross-Linguistic Analysis of Conversational Dynamics. *Speech Commun.* **2025**, *173*, 103271, doi:10.1016/j.specom.2025.103271.
7. Guo, P.; Huang, S.; Li, M. DDA-MSLD: A Multi-Feature Speech Lie Detection Algorithm Based on a Dual-Stream Deep Architecture. *Information* **2025**, *16*, 386, doi:10.3390/info16050386.
8. Ayvaz, U.; Gürüler, H.; Khan, F.; Ahmed, N.; Whangbo, T.; Akmalbek Bobomirzaevich, A. Automatic Speaker Recognition Using Mel-Frequency Cepstral Coefficients Through Machine Learning. *Comput. Mater. Contin.* **2022**, *71*, 5511–5521, doi:10.32604/cmc.2022.023278.
9. Prabakaran, D.; Sriuppili, S. Speech Processing: MFCC Based Feature Extraction Techniques- An Investigation. *J. Phys. Conf. Ser.* **2021**, *1717*, 012009, doi:10.1088/1742-6596/1717/1/012009.
10. Sood, M.; Jain, S. Speech Recognition Employing MFCC and Dynamic Time Warping Algorithm. In *Innovations in Information and Communication Technologies (IICT-2020)*; 2021; pp. 235–242.
11. Wijaya, N.N.; Setiadi, D.R.I.M.; Muslikh, A.R. Music-Genre Classification Using Bidirectional Long Short-Term Memory and Mel-Frequency Cepstral Coefficients. *J. Comput. Theor. Appl.* **2024**, *1*, 243–256, doi:10.62411/jcta.9655.
12. Saleem, N.; Gao, J.; Khattak, M.I.; Rauf, H.T.; Kadry, S.; Shafi, M. DeepResGRU: Residual Gated Recurrent Neural Network-Augmented Kalman Filtering for Speech Enhancement and Recognition. *Knowledge-Based Syst.* **2022**, *238*, 107914, doi:10.1016/j.knosys.2021.107914.
13. Li, Y.; Kang, S. Deep Neural Network-based Linear Predictive Parameter Estimations for Speech Enhancement. *IET Signal Process.* **2017**, *11*, 469–476, doi:10.1049/iet-spr.2016.0477.
14. Karapiperis, S.; Ellinas, N.; Vioni, A.; Oh, J.; Jho, G.; Hwang, I.; Raptis, S. Investigating Disentanglement in a Phoneme-Level Speech Codec for Prosody Modeling. In Proceedings of the 2024 IEEE Spoken Language Technology Workshop (SLT); IEEE, December 2 2024; pp. 668–674.

15. Sivasathiya, M.G.; D, A. kumar; AR, H.R.; R, K. Emotion-Aware Multimedia Synthesis: A Generative AI Framework for Personalized Content Generation Based on User Sentiment Analysis. In Proceedings of the 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT); IEEE, January 4 2024; pp. 1344–1350.
16. Colunga-Rodriguez, A.A.; Martínez-Rebollar, A.; Estrada-Esquivel, H.; Clemente, E.; Pliego-Martínez, O.A. Developing a Dataset of Audio Features to Classify Emotions in Speech. *Computation* **2025**, *13*, 39, doi:10.3390/computation13020039.
17. Wang, N.; Zhang, X.; Sharma, A. A Research on HMM Based Speech Recognition in Spoken English. *Recent Adv. Electr. Electron. Eng. (Formerly Recent Patents Electr. Electron. Eng.* **2021**, *14*, 617–626, doi:10.2174/2352096514666210413122517.
18. Srivastava, D.R.K.; Pandey, D. Speech Recognition Using HMM and Soft Computing. *Mater. Today Proc.* **2022**, *51*, 1878–1883, doi:10.1016/j.matpr.2021.10.097.
19. Turki, T.; Roy, S.S. Novel Hate Speech Detection Using Word Cloud Visualization and Ensemble Learning Coupled with Count Vectorizer. *Appl. Sci.* **2022**, *12*, 6611, doi:10.3390/app12136611.
20. Hao, C.; Li, Y. Simulation of English Speech Recognition Based on Improved Extreme Random Forest Classification. *Comput. Intell. Neurosci.* **2022**, *2022*, 1–10, doi:10.1155/2022/1948159.
21. Dua, S.; Kumar, S.S.; Albagory, Y.; Ramalingam, R.; Dumka, A.; Singh, R.; Rashid, M.; Gehlot, A.; Alshamrani, S.S.; AlGhamdi, A.S. Developing a Speech Recognition System for Recognizing Tonal Speech Signals Using a Convolutional Neural Network. *Appl. Sci.* **2022**, *12*, 6223, doi:10.3390/app12126223.
22. Shashidhar, R.; Patilkulkarni, S.; Puneeth, S.B. Combining Audio and Visual Speech Recognition Using LSTM and Deep Convolutional Neural Network. *Int. J. Inf. Technol.* **2022**, *14*, 3425–3436, doi:10.1007/s41870-022-00907-y.
23. Hema, C.; Garcia Marquez, F.P. Emotional Speech Recognition Using CNN and Deep Learning Techniques. *Appl. Acoust.* **2023**, *211*, 109492, doi:10.1016/j.apacoust.2023.109492.
24. Oruh, J.; Viriri, S.; Adegun, A. Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition. *IEEE Access* **2022**, *10*, 30069–30079, doi:10.1109/ACCESS.2022.3159339.
25. Orken, M.; Dina, O.; Keylan, A.; Tolganay, T.; Mohamed, O. A Study of Transformer-Based End-to-End Speech Recognition System for Kazakh Language. *Sci. Rep.* **2022**, *12*, 8337, doi:10.1038/s41598-022-12260-y.
26. Song, Q.; Sun, B.; Li, S. Multimodal Sparse Transformer Network for Audio-Visual Speech Recognition. *IEEE Trans. Neural Networks Learn. Syst.* **2023**, *34*, 10028–10038, doi:10.1109/TNNLS.2022.3163771.
27. Gondohanindijo, J.; -, M.; Noersasongko, E.; -, P.; Setiadi, D.R.M. Multi-Features Audio Extraction for Speech Emotion Recognition Based on Deep Learning. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, doi:10.14569/IJACSA.2023.0140623.
28. Tyagi, S.; Szénási, S. Optimizing Speech Emotion Recognition with Deep Learning and Grey Wolf Optimization: A Multi-Dataset Approach. *Algorithms* **2024**, *17*, 90, doi:10.3390/a17030090.
29. Bhanbhro, J.; Memon, A.A.; Lal, B.; Talpur, S.; Memon, M. Speech Emotion Recognition: Comparative Analysis of CNN-LSTM and Attention-Enhanced CNN-LSTM Models. *Signals* **2025**, *6*, 22, doi:10.3390/signals6020022.
30. Yu, S.; Meng, J.; Fan, W.; Chen, Y.; Zhu, B.; Yu, H.; Xie, Y.; Sun, Q. Speech Emotion Recognition Using Dual-Stream Representation and Cross-Attention Fusion. *Electronics* **2024**, *13*, 2191, doi:10.3390/electronics13112191.
31. Wei, Z.; Ge, C.; Su, C.; Chen, R.; Sun, J. A Deep Learning Model for Speech Emotion Recognition on RAVDESS Dataset. *Int. J. Adv. Comput. Sci. Appl.* **2025**, *16*, 316–323, doi:10.14569/IJACSA.2025.0160531.
32. Makhmudov, F.; Kutlimuratov, A.; Cho, Y.-I. Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition. *Appl. Sci.* **2024**, *14*, 11342, doi:10.3390/app142311342.
33. Kim, J.-Y.; Lee, S.-H. Accuracy Enhancement Method for Speech Emotion Recognition From Spectrogram Using Temporal Frequency Correlation and Positional Information Learning Through Knowledge Transfer. *IEEE Access* **2024**, *12*, 128039–128048, doi:10.1109/ACCESS.2024.3447770.
34. Huang, Z.; Ji, S.; Hu, Z.; Cai, C.; Luo, J.; Yang, X. ADFF: Attention Based Deep Feature Fusion Approach for Music Emotion Recognition. In Proceedings of the Interspeech 2022; ISCA: ISCA, September 18 2022; Vol. 2022-Septe, pp. 4152–4156.
35. de Souza, D.B.; Bakri, K.J.; Ferreira, F. de S.; Inacio, J. Multitaper-Mel Spectrograms for Keyword Spotting. *IEEE Signal Process. Lett.* **2022**, *29*, 2028–2032, doi:10.1109/LSP.2022.3208411.
36. McAllister, T.; Gambäck, B. Music Style Transfer Using Constant-Q Transform Spectrograms. In *Artificial Intelligence in Music, Sound, Art and Design*; 2022; pp. 195–211.
37. Raju, V.V.N.; Saravanakumar, R.; Yusuf, N.; Pradhan, R.; Hamdi, H.; Saravanan, K.A.; Rao, V.S.; Askar, M.A. Enhancing Emotion Prediction Using Deep Learning and Distributed Federated Systems with SMOTE Oversampling Technique. *Alexandria Eng. J.* **2024**, *108*, 498–508, doi:10.1016/j.aej.2024.07.081.

38. Ding, Z.; Wang, Z.; Zhang, Y.; Cao, Y.; Liu, Y.; Shen, X.; Tian, Y.; Dai, J. Trade-Offs between Machine Learning and Deep Learning for Mental Illness Detection on Social Media. *Sci. Rep.* **2025**, *15*, 14497, doi:10.1038/s41598-025-99167-6.
39. Modi, N.; Kumar, Y.; Mehta, K.; Chaplot, N. Physiological Signal-Based Mental Stress Detection Using Hybrid Deep Learning Models. *Discov. Artif. Intell.* **2025**, *5*, 166, doi:10.1007/s44163-025-00412-8.
40. Pathirana, A.; Rajakaruna, D.K.; Kasthurirathna, D.; Atukorale, A.; Aththidiye, R.; Yatiipansalawa, M.; Yatipansalawa, M. A Reinforcement Learning-Based Approach for Promoting Mental Health Using Multimodal Emotion Recognition. *J. Futur. Artif. Intell. Technol.* **2024**, *1*, 124–142, doi:10.62411/faith.2024-22.
41. Wang, Y.; Huang, J.; Zhao, Z.; Lan, H.; Zhang, X. Speech Emotion Recognition Using Multi-Scale Global–Local Representation Learning with Feature Pyramid Network. *Appl. Sci.* **2024**, *14*, doi:10.3390/app142411494.

Prosodic Spatio-Temporal Feature Fusion with Attention Mechanisms for Speech Emotion Recognition

ORIGINALITY REPORT

16%	12%	13%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	3%
2	Submitted to King Abdulaziz University Student Paper	2%
3	Kishor B. Bhangale, Mohanaprasad Kothandaraman. "Speech emotion recognition using the novel PEmoNet (Parallel Emotion Network)", Applied Acoustics, 2023 Publication	1%
4	arxiv.org Internet Source	1%
5	www.medrxiv.org Internet Source	<1%
6	Submitted to Arts, Sciences & Technology University In Lebanon Student Paper	<1%
7	edupij.com Internet Source	<1%
8	Keiji Nakamura, Norihiro Itsubo. "Lifecycle Assessment of Monosodium Glutamate Made from Non-Edible Biomass", Sustainability, 2021 Publication	<1%
9	Submitted to University of Hull Student Paper	<1%
10	www.researchsquare.com Internet Source	<1%
11	Submitted to Budapest University of Technology and Economics	<1%

	Student Paper	
12	lettersinhighenergyphysics.com Internet Source	<1%
13	Poonam Nandal, Mamta Dahiya, Meeta Singh, Arvind Dagur, Brijesh Kumar. "Progressive Computational Intelligence, Information Technology and Networking", CRC Press, 2025 Publication	<1%
14	Samarjeet Borah, Ratna Raja Kumar Jambi, Sharifah Sakinah Syed Ahmad, Mahendra Prabhakar Deore. "Applied Soft Computing Techniques - Theoretical Principles and Practical Applications", Apple Academic Press, 2025 Publication	<1%
15	Yuhua Wang, Jianxing Huang, Zhengdao Zhao, Haiyan Lan, Xinjia Zhang. "Speech Emotion Recognition Using Multi-Scale Global–Local Representation Learning with Feature Pyramid Network", Applied Sciences, 2024 Publication	<1%
16	napier-repository.worktribe.com Internet Source	<1%
17	Kishor B. Bhangale, Mohanaprasad Kothandaraman. "A novel two-way feature extraction technique using multiple acoustic and wavelets packets for deep learning based speech emotion recognition", Multimedia Tools and Applications, 2024 Publication	<1%
18	Mehdi Ghayoumi. "Generative Adversarial Networks in Practice", CRC Press, 2023 Publication	<1%
19	Submitted to Universidade do Porto Student Paper	<1%
20	artemis.cslab.ece.ntua.gr:8080 Internet Source	<1%

21	www.opastpublishers.com Internet Source	<1%
22	Jun Yang, Liyan Wang, Yong Qi, Haifeng Chen, Jian Li. "Multimodal Information Fusion and Data Generation for Evaluation of Second Language Emotional Expression", Applied Sciences, 2024 Publication	<1%
23	Mohammed H. Abdulwahhab, Parosh Aziz Abdulla, Karwan Jacksi. "A Novel Approach to Efficiently Verify Sequential Consistency in Concurrent Programs", Computers, 2025 Publication	<1%
24	Sudi Murindanyi, Kyamanywa Hamza, Sulaiman Kagumire, Ggaliwango Marvin. "Responsible Music Genre Classification Using Interpretable Model-Agnostic Visual Explainers", SN Computer Science, 2024 Publication	<1%
25	Ting Guo, Nurmemet Yolwas, Wushour Slamu. "Efficient Conformer for Agglutinative Language ASR Model Using Low-Rank Approximation and Balanced Softmax", Applied Sciences, 2023 Publication	<1%
26	dl.futuretechsci.org Internet Source	<1%
27	www.frontiersin.org Internet Source	<1%
28	"Speech and Computer", Springer Science and Business Media LLC, 2023 Publication	<1%
29	cis.temple.edu Internet Source	<1%
30	dx.doi.org Internet Source	<1%
	fastercapital.com	

31	Internet Source	<1%
32	pure.southwales.ac.uk Internet Source	<1%
33	Afaq Khattak, Pak-wai Chan, Feng Chen, Abdulrazak H. Almaliki. "Deep ResNet Strategy for the Classification of Wind Shear Intensity Near Airport Runway", Computer Modeling in Engineering & Sciences, 2025 Publication	<1%
34	Astha Tripathi, Poonam Rani. "Multilingual speech emotion recognition using IGRFXG – Ensemble feature selection approach", Applied Acoustics, 2025 Publication	<1%
35	D. Thiripurasundari, Kishor Bhangale, V. Aashritha, Sisira Mondreti, Mohanaprasad Kothandaraman. "Speech emotion recognition for human–computer interaction", International Journal of Speech Technology, 2024 Publication	<1%
36	Krisha S. Patel, Kenneth L. McCall, Brian J. Piper. "From Peak to Plunge: A Multi-Database Analysis of State Level Disparities in Hydromorphone use in the U.S. (2010-2023)", Cold Spring Harbor Laboratory, 2025 Publication	<1%
37	Tan Guangwen, Li Mengshan, Hou Biyu, Zhu Jihong, Guan Lixin. "Achieving accurate trajectory predicting and tracking for autonomous vehicles via reinforcement learning-assisted control approaches", Engineering Applications of Artificial Intelligence, 2024 Publication	<1%
38	Xin Qi, Qing Song, Guowei Chen, Pengzhou Zhang, Yao Fu. "Acoustic Feature Excitation-	<1%

	and-Aggregation Network Based on Multi-Task Learning for Speech Emotion Recognition", Electronics, 2025 Publication	
39	aclanthology.org Internet Source	<1%
40	ijercse.com Internet Source	<1%
41	orbi.uliege.be Internet Source	<1%
42	pmc.ncbi.nlm.nih.gov Internet Source	<1%
43	sistemasi.ftik.unisi.ac.id Internet Source	<1%
44	vdoc.pub Internet Source	<1%
45	www.acadlore.com Internet Source	<1%
46	www.nature.com Internet Source	<1%
47	"Advances in Social Networks Analysis and Mining", Springer Science and Business Media LLC, 2025 Publication	<1%
48	"Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017", Springer Nature, 2017 Publication	<1%
49	Alvaro A. Colunga-Rodriguez, Alicia Martínez-Rebollar, Hugo Estrada-Esquivel, Eddie Clemente, Odette A. Pliego-Martínez. "Developing a Dataset of Audio Features to Classify Emotions in Speech", Computation, 2025 Publication	<1%

50	Dong Liu, Qirong Mao, Lijian Gao, Gang Wang. "Leveraging Contrastive Language–Image Pre-Training and Bidirectional Cross-attention for Multimodal Keyword Spotting", Engineering Applications of Artificial Intelligence, 2024 Publication	<1%
51	Zhanyi Ding, Zhongyan Wang, Yeyubei Zhang, Yuchen Cao, Yunchong Liu, Xiaorui Shen, Yexin Tian, Jianglai Dai. "Trade-offs between machine learning and deep learning for mental illness detection on social media", Scientific Reports, 2025 Publication	<1%
52	"Artificial Intelligence and Robotics", Springer Science and Business Media LLC, 2025 Publication	<1%
53	"Cognitive Informatics and Soft Computing", Springer Science and Business Media LLC, 2019 Publication	<1%
54	Michela Balconi. "The Neuropsychology of Nonverbal Communication: The Facial Expressions of Emotions", Neuropsychology of Communication, 2010 Publication	<1%
55	Soila Kuuluvainen, Saara Kaskivuo, Martti Vainio, Eleonore Smalle, Riikka Möttönen. "Prosody enhances learning of statistical dependencies from continuous speech streams in adults", Cognition, 2025 Publication	<1%
56	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		