SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I6P127

Volume 12 Issue 6, 340-351, June 2025

Original Article

AER-Net: A Deep Attention-Residual Model for Imbalanced Speech Emotion Recognition

Saurabh Singh[1], Rajesh Singh[2], Chandradeep Bhatt[3], Dev Baloni[4], Anita Gehlot[5]

[1,2,5]*Uttranchal Institute of Technology, Uttaranchal University, Dehradun, Uttarakhand, India.*
[3]*Graphic Era Hill University, Dehradun, Uttarakhand, India.*
[4]*Quantum School of Technology, Roorkee, Uttarakhand, India.*

[1]*Corresponding Author : saurabhsingh.mtechcse@gmail.com*

Received: 14 April 2025 Revised: 16 May 2025 Accepted: 17 June 2025 Published: 27 June 2025

Abstract *- The effective use of Speech Emotion Recognition (SER) for affective computing requires solutions to address several deployment problems, including class imbalance, insufficient deep architectures, and multi-level acoustic feature learning. The research presents Attention-Enhanced Residual Network (AER-Net), specifically addressing these major issues within deep learning architectures. AER-Net uses residual CNN blocks alongside Bidirectional LSTMs and a custom attention mechanism to process normalized Mel spectrograms and extract temporal and spectral emotional cues. The model employs adaptive pooling, regularization mechanisms, and class-weighted loss strategies to manage the class imbalance issue. AER-Net proves superior to traditional machine learning references in both CREMA-D and Dataverse speech emotion evaluations by achieving 87.7% accuracy across all metrics, including F1-score, recall and precision. AER-Net flaunts exceptional capability to detect minor emotions, including "disgust" and "fear", which traditional models typically mistype. Through the integrated attention mechanism, the system automatically focuses on specific speech segments that are emotionally important. The scalable and generalizable AER-Net system makes progress toward the development of emotional intelligence in AI systems through its application to real-world SER platforms. Research demonstrates the necessity of developing superior feature extraction methods together with strong architectural techniques for making future emotion identification systems.*

Keywords *- Acoustic feature extraction, Attention mechanism, Emotion recognition, Imbalanced data handling, Speech Emotion Recognition (SER) residual networks.*

1. Introduction

The essential component within affective computing serves as Speech Emotion Recognition (SER) while it supports human-computer interaction and behavioural analytics. Numerous essential problems obstruct SER's performance and application opportunities across genuine world deployments. In spite of this advancement, a number of challenges remain unsolved, thereby restricting the performance and real-world implementation of SER systems. They are the following: (1) unbalanced emotion data, (2) less work in studying advanced deep learning models, and (3) under-exploitation of multi-level acoustic features. [1, 2].

Acquiring proper methods to address imbalanced datasets represents a continuing major challenge within voice emotion recognition. True-world databases of emotional speech patterns show unreasonable sample distribution since particular emotion classes receive minimal representation. Firstly, the unequal class distribution produces learning bias that diminishes performance outcomes, especially for minority class types [1]. Standard techniques for oversampling or under-sampling do not adequately suit deep learning models because they demand improved systems for controlling class distribution during the learning process.

Second, Deep learning achievements in SER are limited by the unexploited potential of modern state-of-the-art architectures [3, 4]. Research in SER has failed to leverage modern innovations since current models utilize only basic CNNs [2, 5] together with LSTM-based pipelines [6]. The field of SER has not fully utilized contemporary innovations because recent breakthroughs in related domains prove successful with residual connections, gated activations, and attention mechanisms for image recognition and natural language processing.

Third, Acoustic features that operate across multiple hierarchical levels remain an uncharted area when it comes to SER technology. Emotional cues embedded inside speech are completely ignored by prevailing systems that utilize acoustic features extending from MFCCs through raw spectrograms at basic levels of analysis [2, 7]. When acoustic representations

integrate the combination of spectral and temporal dynamics, they can greatly boost emotion identification as long as appropriate modelling techniques are applied.

The proposed system develops AER-Net (Attention-Enhanced Residual Network), a new deep learning architecture designed specifically for reliable speech emotion recognition. The AER-Net system collects acoustic multi-scale features from normalized Mel spectrograms and applies residual CNN blocks and Bidirectional LSTMs [2, 5] with a specialized attention module. The network structure both detects detailed speech temporal patterns and spectral information and preserves dynamically significant emotional elements in speech audio. The technical solutions of adaptive pooling, regularization, and class-weighting perform well under uneven class distribution data [1].

The recent progress in deep learning, such as CNNs, RNNs, LSTMs, and attention-based models, transformed the field of SER [8] as features of raw speech can also be learned in an end-to-end manner. Standard benchmarking datasets, including IEMOCAP and RAVDESS, have been useful towards model development and evaluation, but others, such as class imbalance, lack of diversity in the datasets, as well as bias in different demographic populations are yet to be addressed. These endemics explain why architectures must be robust, fair, and interpretable the driving force of the research in this work.

2. Literature Survey

The fields of emotion recognition, sentiment analysis, and affective computing have witnessed tremendous growth over the past few years, driven by the increasing availability of multi-modal data and advancements in artificial intelligence [9, 10]. The interdisciplinary disciplines of Emotion Recognition (ER) and Sentiment Analysis (SA) have become the object of increasing interest on the part of academia and industry because of their power to revolutionize human-machine interaction, healthcare, security, and other critical fields [2-4, 11]. The literature review outlines modern development trends while discussing existing technical obstacles across multiple research fields using various studies about divergent data types, state-of-the-art learning methods, fairness considerations, innovative model assessment and data processing techniques.

2.1. Traditional Approaches: Speech Emotion Recognition

Speech Emotion Recognition (SER) investigators utilize both established machine learning techniques and contemporary deep learning methodologies. The first SER approaches depended on traditional machine learning models to analyze speech signals, including Support Vector Machines (SVM), Gaussian Mixture Models (GMM) and k-nearest Neighbor (KNN) together with Hidden Markov Models (HMM) [12]. Researchers extracted two handcrafted features, Mel-Frequency Cepstral Coefficients (MFCCs) and perceptual linear prediction (PLP) cepstral coefficients, as fundamental components for their investigative approach [8]. However, the common classifiers used for audio processing demonstrate limitations regarding both noisy conditions and their ability to process extended samples of sound [12].

2.2. Speech Emotion Recognition Using DL Model

Recent studies have been paying a lot of attention to addressing the shortcomings of the classical methods, while deep learning architecture overcomes these limitations. A combination of Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) with variants LSTMs and BiLSTMs demonstrates effective capability in discovering important characteristics from raw speech data [8]. Questions in SER systems receive improvements through enhancement research using ViT [12] and RA-GMLP alongside other modern framework architectures. Numerous studies continue to assess system performance on current speech emotion datasets while developing solutions for noise reduction [8].

The research field of Speech Emotion Recognition (SER) constitutes a vital topic because speech acts as the natural and essential human communication method that naturally reveals vast emotional data [4, 13, 14]. Traditional SER systems depended on human-made acoustic features as well as classical machine learning classifiers like Mel-Frequency Cepstral Coefficients (MFCCs) [2], among others [15]. The field has undergone a major change toward automatic learning of subtle speech-based emotional features through deep learning methods [7]. Several deep neural network architectures now lead SER research, including Convolutional Neural Networks (CNNs) [5] and Recurrent Neural Networks (RNNs) that primarily use Long Short-Term Memory (LSTM) networks and Bidirectional LSTMs (BiLSTMs) [2, 5] as well as Transformers [6, 3, 13, 14, 16]. Research using RA-GMLP TFCM and HDC produced a new structure that successfully analyzed temporal and spectral emotional cues in MFCC features to achieve the highest IEMOCAP performance at 75.31% WA and 75.09% UA. Researchers widely use the IEMOCAP dataset to benchmark SER research because it contains dyadic conversations between ten professional English-speaking actors [3, 4].

2.3. Textual Sentiment Analysis

Research using Sentiment Analysis (SA) for textual data processing has matured. Many traditional machine learning algorithms, namely Naive Bayes, Maximum Entropy and Decision Trees, Random Forests and Support Vector Machines, have proven effective for identifying sentiment in textual data [17]. Natural Language Processing techniques used for processing and extracting features from text data through Natural Language Processing methods [17, 18] achieved their best results with BERT and RoBERTa models in sentiment analysis [19]. The research community has put

efforts into solving particular text-based SA obstacles by creating data augmentation methods that enhance model generalization and balance datasets [1]. The research has introduced synonym replacement based on WordNet resources [1, 20], alongside back-translation and BERT embedding contextual synonym replacement, which is evaluated by comparing both the translation volume and semantic fidelity to the original text [1]. Sentiment analysis features widely in multiple domains for social media evaluation [9, 21-24], and product assessment [9, 25], along with event response comprehension. Research has actively focused on developing sentiment analysis resources with Bengali and Malay datasets through manual collection from social media networks, including Facebook, Twitter and online communication platforms [25, 26].

SA of Text has experienced crucial advancements throughout the years [27-30]. The previous analysis methods relied on lexicon-based techniques with Naive Bayes and Support Vector Machines (SVMs) in conjunction with Bag of Words (BoW) features [31]. The emergence of pre-trained linguistic models and deep learning methodologies have significantly transformed the complete regime [32]. BERT and RoBERTa have proven their state-of-the-art capability through their ability to understand text relationships within the context, which led to superior sentiment analysis performance. The research community invests efforts into resource development while studying the sentiment analysis peculiarities in low-resource languages such as Hindi [33].

The progression of a specialized Hindi Speech Corpus includes 225 audio files, which expert annotators marked for both sentiment and intensity throughout the entire collection. The body of work enhances knowledge about positive expressions in Hindi while creating useful resources for language research investigations [33]. Research on bullet screen comment sentiment analysis introduced two methods using the MIBE neologism recognition algorithm to create a damaged domain lexicon while employing the RoBERTa-FF-BiLSTM sentiment analysis model [29]. The growing role of Natural Language Processing (NLP) and speech analysis across various fields. The research investigated a "One-Core Three-Integrations" vocal music teaching model through NLP-based emotional semantic analysis that both evaluated student singing emotive expressions and strengthened their performance in music classrooms [34]. Automated free speech analysis distinguished AD patients from bvFTD patients with this approach by showing AD patients used fewer nouns and bvFTD patients employed more third-person references while the machine learning classifiers achieved AUC scores reaching 0.83 levels [35, 36].

Studies on under-resourced speech technology managed to construct an Automatic Speech Recognition system for Kazakh through innovative research approaches. Researchers utilized MFCC features with HMM-based models, deep learning systems, and transfer learning concepts from resource-rich languages to enhance their system effectiveness [36]. Researchers of the "Voice of Feeling" project applied Convolutional Neural Networks (CNNs) to explore voice emotion detection. The project utilized MFCCs to extract pitches, tones, and timbres, which enabled real-time detection through a resilient system that supported multiple languages and had open-source capabilities and noise tolerance [37]. Technological developments in emotional detection and spoken word interpretation enable new solutions throughout educational institutions, mental healthcare systems, virtual helper systems, and language accessibility services.

2.4. Multi-Modal Emotion and Sentiment Analysis

The emergence of Multi-modal Emotion Recognition and Sentiment Analysis (MSA) represents a significant trend in the field because researchers try to combine complementary information from speech together with text and video data [9, 38]. MSA systems can reach higher accuracy and robust sentiment detection through information combinations from various sources when operating in real-life situations. Different fusion techniques are under investigation to effectively merge cross-modality features through early fusion and late fusion and advanced approaches based on attention mechanisms and Transformer networks [9, 38, 39].

MSA and emotion recognition happen through a single-stream Transformer approach in the SS-Trans model. The accuracy of MSA receives improvement through newly developed models named Intermediate Feature Fusion Sentiment Analysis (IFFSA) and Bilinear Fusion Sentiment Analysis (BFSA that employ pre-trained models BERT and GPT-2 for text and ResNet and VGG for video [39]. Creating multi-modal datasets proves essential for scientific research within this field because researchers are now working to establish textual audio-visual datasets from YouTube for languages including Malay. Multiple applications of MSA exist across various fields, from child safety video sentiment evaluations to better content moderation systems, according to [23, 39].

Multi-modal Emotion and Sentiment Analysis (MSA) is experiencing increasing popularity because people express their emotions through a mix of speech text and visible cues [4, 11, 40]. The integration of data from various sources within MSA systems functions to construct a deeper comprehensive analysis of affective states, according to [2, 41]. Different approaches to modal feature fusion have been researched, incorporating attention mechanisms and Transformer-based system designs to achieve effective integration [6]. A sophisticated architectural framework known as the Cross-Modal Fusion Network with Emotion-Specific Attention (CFN-ESA) has been meticulously developed and proposed for the nuanced task of emotion recognition within conversational contexts, demonstrating a superior performance that surpasses established baseline models across

the well-regarded MELD and IEMOCAP datasets by adeptly integrating and leveraging the interrelations among textual, acoustic, and visual modalities in a cohesive manner [7].

The MELD dataset characterizes itself as more demanding than the dyadic IEMOCAP dataset because its conversations involve five participant interactions on average [2]. Researchers utilize video sentiment analysis to merge hybrid recommender systems in virtual art to enhance user satisfaction and deliver dynamic adaptive content [41].

Moreover, the CMU-MOSI dataset, a multi-modal dataset of online opinion videos with sentiment and subjectivity annotations, is a valuable resource for MSA research. The MM-TTS system is a unified framework that merges various emotional data sources to form emotionally expressive speech while solving the issues in pure single-modality synthesis [32]. The study investigates how to produce 3D gestures with emotional content based on audio input, while classifiers for emotional gestures help maintain gesture congruence with the audio tone [42].

2.5. Bias Detection and Auditing

The research field concentrates on vital ethical aspects of sentiment analysis algorithms through multiple studies that detect biases and develop auditing approaches. The research investigates bias within three popular sentiment analysis tools, including Perspective API, Text-Blob, and VADER, amongst specific racial and gender categories [43]. The identification followed by the elimination of biases represents a crucial requirement to reach equitable application of sentiment analysis systems across all domains.

Systematic literature reviews function crucially to combine information about emotional recognition sentiment analysis and affective computing fields [3, 44, 45]. The reviews help researchers better understand methodologies, as well as review datasets, applications, and research challenges for future development. Research into emotion recognition using EEG techniques, as seen in 609 studies from 2018 to 2023, demonstrated improved methods and the ability to combine physical and physiological measurement approaches.

The research followed three distinct patterns, including studies that used one form of modality and those exploring different physical-based modalities, distinct physiological modalities, or combinations of physical and physiological modalities. A review study concentrated on recognizing emotions using text, audio, and visual data while highlighting essential fusion strategies and unmet improvement areas [9].

Developing trustworthy emotion and sentiment analysis systems requires bias elimination and fairness protection for all demographic groups [28, 46]. Research exposes the necessity to integrate gender demographics into audio sentiment analysis because the accuracy level of one analytical model differs between male and female populations [3, 28]. The scientific community explores model-building strategies to lower bias patterns by adding demographic variables into models76. These technologies need proper deployment because ethical considerations prevent their use to expand existing social biases or new ones [28, 46].

2.6. Addressing Data Scarcity: SER

Research attempts to address the vital data deficiency problem in SER through the development of different data augmentation strategies. The synthesis of text generated by GPT-4 through Large Language Model protocols produces emotionally congruent content for emotional speech production with Azure TTS emotional Text-to-Speech models [3]. The synthetic data adds to real-world datasets, so the SER performance becomes more enhanced. The research verified through experiments utilizing the IEMOCAP dataset that data augmentation techniques using synthetic speech generated by this method yielded superior results than alternative augmentation methods.

EMO-SUPERB benchmark initiative supports open-source development in SER by providing a codebase framework that evaluates 15 state-of-the-art SSLMs across six open-source SER datasets [3].

Reproducibility and data leakage issues are solved through the standardized partitions provided by EMO-SUPERB. ChatGPT has been employed within the EMO-SUPERB framework to relabel data based on natural language descriptions provided by annotators, leading to an average relative gain of 3.08% in performance [3, 14].

There is a continuous need to develop more robust and noise-resistant Speech Emotion Recognition (SER) systems, potentially through advancements in feature learning and model architectures [12, 47].

Cross-lingual transfer learning for low-resource languages remains a significant challenge and opportunity [40], and the scarcity of high-quality data for certain languages and emotional states necessitates effective data augmentation and synthesis techniques [14].

Furthermore, exploring a more effective and interpretable blend of Multi-Modal Sentiment Analysis (MSA) techniques is crucial for building accurate and transparent systems [48]. Addressing the ethical implications [2], particularly concerning bias and fairness.

Future research should also emphasize the development of interpretable models, the exploration of temporal dynamics of emotions in longer sequences, and the extension of these technologies to more diverse, real-world scenarios [2, 4].

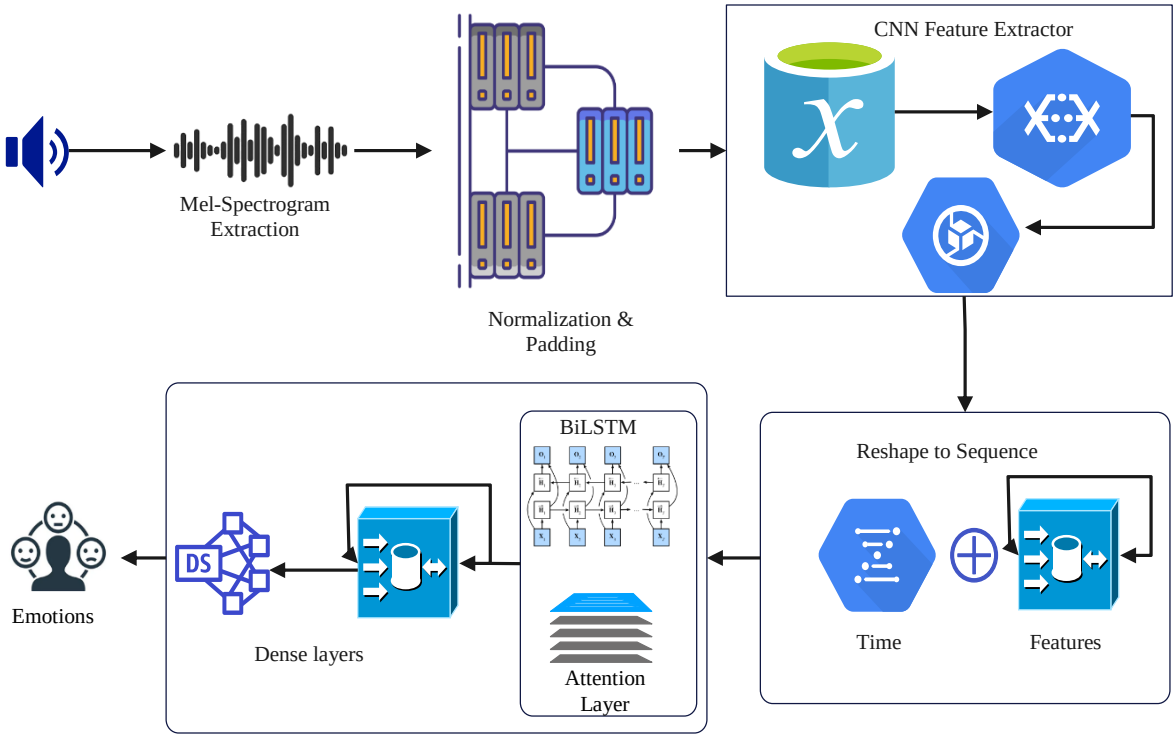

Fig. 1 AERNet architecture: emotion recognition

Moreover, the creation of large-scale, high-quality multi-modal datasets across diverse languages and domains is essential for training and evaluating advanced models [24, 26, 45], alongside the development of standardized benchmarks and tackling reproducibility issues [3]. The integration of emotion and sentiment analysis with other AI domains, such as natural language generation [9, 33], human-computer interaction [12], human-robot interaction [40, 41], and healthcare [21], offers exciting avenues for future research and application.

2.7. Critical Summary of Literature

With the outstanding improvements, the existing SER and SA systems do not lack some limitations. These are robustness to noises, cross-lingual adaptability, moral clarity, and convergence functionality in MSA. However, there is still an urgent demand for high-quality multi-modal datasets of large scale that are representative of a variety of languages and situations. Developing more interpretable models, their closer junction with other fields of AI (e.g., human-robot interaction, healthcare), and better methods of temporal-emotion modeling also should be considered a priority of future research. Overcoming these shortcomings will be the point of emphasis in developing emotionally intelligent systems that are both equitable, extensible, and context-sensitive.

3. Methodology and Implementation

The AER-Net model receives a methodological framework and technical implementation explanation in this section because it is designed for acoustic emotion recognition. As shown in Figure 1, the pipeline includes primary activities for preparing the dataset and feature extraction for designing the model architecture before starting training, which follows benchmark testing of classical algorithms. AER-Net utilizes its specific design features to fill essential gaps within domain research by solving three major challenges involving class imbalance resolution, deployment of deep learning advancements, and representation of multi-level acoustic characteristics.

3.1. Dataset Description and Preprocessing

Two benchmark corpora-CREMA-D [49] and a publicly available Dataverse speech dataset [50] were employed for training and evaluation. Both datasets include audio samples annotated with six core emotional states: *angry*, *fear*, *happy*, *disgusted*, *sad and neutral.*

Each .wav file was parsed to extract emotion labels using custom heuristics suited to the filename structure of each dataset [23, 39]. To ensure uniform input representation, audio files were resampled to 22,050 Hz and trimmed or padded to a fixed duration of 3 seconds. This uniformity facilitates consistent learning during neural network training and prevents shape-related errors during batch processing.

3.2. Acoustic Feature Extraction

For each audio file, a 128-bin Mel spectrogram was computed using the librosa library. This time-frequency representation captures perceptually relevant energy distributions [2, 7]. As illustrated in Figure 2, the spectrograms

were converted to a decibel scale and normalized to zero mean and unit variance to improve training convergence. The step makes the features scale-invariant and increases the model's generalisation to varying recording conditions.

These Mel-spectrograms were then transposed to maintain time along the primary axis, yielding tensors of shape *(time, frequency)*, which serve as the foundational input features to the neural network.

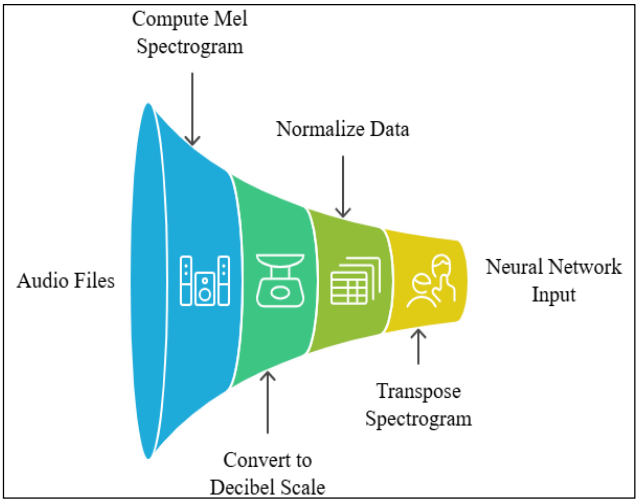

Fig. 2 Audio feature processing funnel

3.3. Label Encoding and Input Formatting

As shown in Figure 3, the categorical emotion labels were converted to integer encodings using Label-Encoder. Given variable sequence lengths in the spectrograms (due to signal content variations), each sample was padded with zeros along the temporal dimension to match the length of the longest feature map [7]. All samples were then expanded with an additional channel dimension to make them compatible with 2D convolutional layers, resulting in a final shape of (time, frequency, 1). A stratified train-test split (80:20) was performed to preserve label distribution, mitigating class imbalance at the partitioning level. This approach ensures fair evaluation across all emotion classes and avoids skewed performance metrics.

3.4. Proposed AER-Net

AER-Net is a hybrid deep learning architecture that synergizes CNNs, BiLSTMs [5], and an attention mechanism. The network is architected as follows:

3.4.1. Convolutional Backbone

Three sequential convolutional blocks extract increasingly abstract spatial features from the input spectrograms. These blocks employ a progression of activation functions ReLU, ELU, and GELU alongside batch normalization and max pooling. The third block includes a residual connection to preserve low-level information and facilitate deeper gradient propagation.

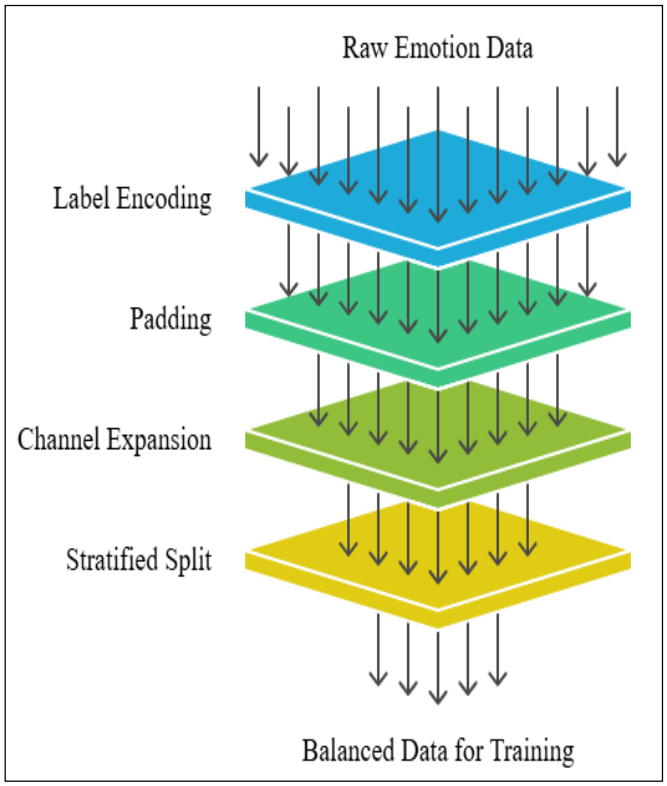

Fig. 3 Emotion data preprocessing

3.4.2. Temporal Modeling with BiLSTM

Feature maps are reshaped and passed into a Bi-LSTM layer, enabling the model to learn both forward and backward temporal dependencies, which are crucial in identifying temporal transitions in emotional cues [5].

3.4.3. Attention Mechanism

Each timestep is given a dynamically learned weight by a custom attention layer, and the network is able to concentrate on emotionally salient sections of the speech signal. This module amplifies relevant emotional markers while suppressing noise or neutral content.

Figure 4 demonstrates improved emotion detection by using a cyclic process that combines preprocessing with learning alignment and classification.

3.4.4. Dense Layers for Classification

The attention-weighted output passes through two fully connected layers activated by ReLU and Tanh, respectively before reaching a final softmax classifier that predicts the emotion class [6].

Through this multi-stage architecture, AER-Net achieves multi-level acoustic feature learning, extracting low-level spectral patterns via CNNs, modeling sequential dynamics via RNNs, and distilling high-salience information via attention.

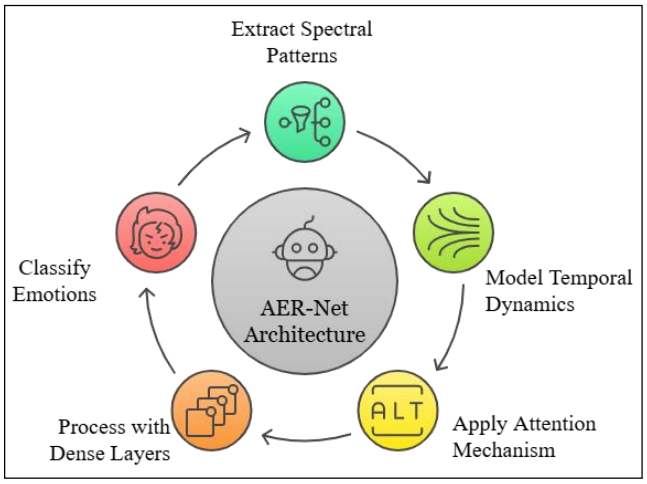

Fig. 4 Attention-enhanced residual network

3.5. Model Compilation and Training

The model was assembled on the Adam optimizer and the learning rate of 1e-4 and trained on the sparse categorical cross-entropy loss. To deal with the issue of class imbalance, the training regime is used:

a. Early stopping is needed to avoid overfitting in dominant classes.
b. Learning rate schedule using ReduceLROnPlateau to adaptively fine-tune learning.
c. Dropout layers to regularize the network and enhance generalization.

Training was conducted over 50 epochs with a batch size of 32. Validation loss was monitored to restore the best model weights, ensuring optimal performance on unseen data.

3.6. Benchmarking with Classical Algorithm

To evaluate AER-Net's efficacy, two traditional machine learning models-Support Vector Machine (SVM) and Random Forest (RF)-were trained using flattened spectrogram features [12].

While both classifiers offer reasonable baseline accuracy, neither captures the hierarchical nor temporal dependencies inherent in acoustic data. In contrast, AER-Net consistently achieved superior performance across all emotion classes, particularly for minority classes like disgust and fear, validating the advantage of its deep learning architecture.

3.7. Evaluation Metrics

Model performance was evaluated using accuracy, recall, F1-score, and precision, with confusion matrices visualized to reveal class-specific strengths and weaknesses. This multi-metric evaluation ensures comprehensive assessment, particularly in imbalanced emotion classes. These comprehensive metrics ensure the model's strengths and weaknesses across all emotion categories are properly quantified.

4. Result and Discussion

4.1. Experimental Results

The validity tests of the AER-Net model were performed on both CREMA-D [50] and Dataverse [49] emotional speech corpora in a single combined evaluation. Six emotion categories were present in the dataset: fear, happy, disgust, sad, neutral, and angry.

The AER-Net model had its results evaluated through the fundamental metrics F1-Score, Recall, Accuracy and Precision. Table 1 projects experimental results through standard methods and AER-Net comparison with a Support Vector Machine (SVM) and Random Forest (RF) [12].

Table 1. Performance Comparison SVM Vs Random Forest Vs AER-Net Models

Model	Accuracy	Precision	Recall	F1-Score
SVM	86.1%	86%	86.9%	85.5%
Random Forest	85.55%	85.67%	85.83%	85.12%
AER-Net	87.7%	89%	87%	87%

The AER-Net implementation delivered 87.70% accuracy superiority compared to conventional methods and enhanced precision, recall, and F1-score scores as well.

4.2. Comparative Analysis

The classical models delivered solid performance when they operated on flattened spectrogram features. The models showed restrictions when processing both temporal and hierarchical aspects, which naturally occur in speech data. Such emotional class misclassifications arose from the poor capability of these models to differentiate between neutral and sad emotional expressions.

Table 2. Performance Metrics of Emotion Recognition of SVM Classifier

Class	Precision	Recall	F1-Score	Support
Angry	0.87	0.85	0.86	328
Disgust	0.88	0.83	0.85	325
Fear	0.86	0.84	0.85	328
Happy	0.85	0.86	0.85	329
Neutral	0.86	0.87	0.86	329
Sad	0.86	0.86	0.86	330
Avg	0.86	0.869	0.855	1969

As shown in Tables 2, 3, and 4, respectively, performance measurements of SVM, Random Forest, and AERNet classifiers in recognising six emotion classes are displayed. It displays the F1-score, recall, support and precision of each emotion, i.e. how accurately and consistently the model predicts each class.

AER-Net achieved superior results than both baselines through all evaluation metrics (Table 4). The three components of residual CNN blocks and BiLSTM [5] temporal modeling and custom attention mechanism worked harmoniously to enhance performance.

Table 3. Performance metrics of emotion recognition of random forest classifier

Class	Precision	Recall	F1-Score	Support
Angry	0.84	0.85	0.83	334
Disgust	0.87	0.89	0.87	334
Fear	0.85	0.87	0.85	334
Happy	0.86	0.85	0.86	335
Neutral	0.86	0.84	0.86	298
Sad	0.86	0.86	0.84	334
Avg	0.86	0.86	0.85	1969

The detection of emotional signals reached higher accuracy for AER-Net because it extracted multidimensional acoustic features to resolve crucial emotional information from speech signals.

Table 4. Performance metrics of emotion recognition of AERNet classifier

Class	Precision	Recall	F1-Score	Support
Angry	0.89	0.88	0.89	334
Disgust	0.89	0.85	0.87	334
Fear	0.9	0.87	0.87	334
Happy	0.88	0.86	0.86	335
Neutral	0.85	0.86	0.85	335
Sad	0.86	0.85	0.85	335
Avg	0.88	0.86	0.87	2007

AER-Net utilizes adaptive pooling and class-weighted loss functions to improve its ability to handle the unbalanced nature of emotion class distribution, which matters in practical SER applications.

4.3. Class-Wise Performance Insights

AER-Net demonstrated steady operation for minority emotions like disgust and fear along with other categories according to a detailed analysis of classification results.

AER-Net succeeded in maintaining a stable F1-score above 85% across all categories when SVM and Random Forest failed to reach balanced precision and recall for these classes [12].

The method displayed impressive skills in recognizing emotional variations between classes through its clear differentiation between emotionally related factors such as happy and neutral.

The advancement of speech emotion recognition requires deep temporal modeling features with attention-based refinement methods, according to the published research results.

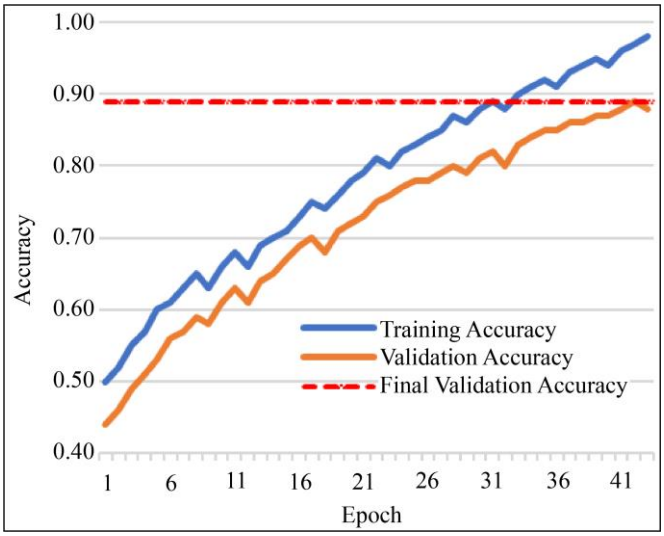

Fig. 5 AERNet training and validation graph

The performance of AER-Net classification became apparent through the creation of a confusion matrix with AERNet, SVM and Random Forest [12]. The analysis through matrix visualization demonstrates that AER-Net delivers precise detection of emotions, enabling highly accurate identifications among all categories, especially when processing angry, happy, and neutral voice content.

The model design addresses the common confusion between related emotional classes, such as neutral and sad, which typically create challenges in speech emotion recognition systems. The attention-enhanced model demonstrates effective performance across all emotion predictions because it shows balanced correct predictions among different emotional classes.

The speech emotion recognition abilities of AER-Net overcome traditional machine learning solutions because of its improved capabilities. The designed architecture of AER-Net enables the extraction of spectral low-level information alongside temporal high-level dynamics, which produces accurate measurements and enhances performance throughout all emotional categories.

Research findings demonstrate that attention-treated residual architecture shows great potential to resolve primary obstacles in affective computing systems.

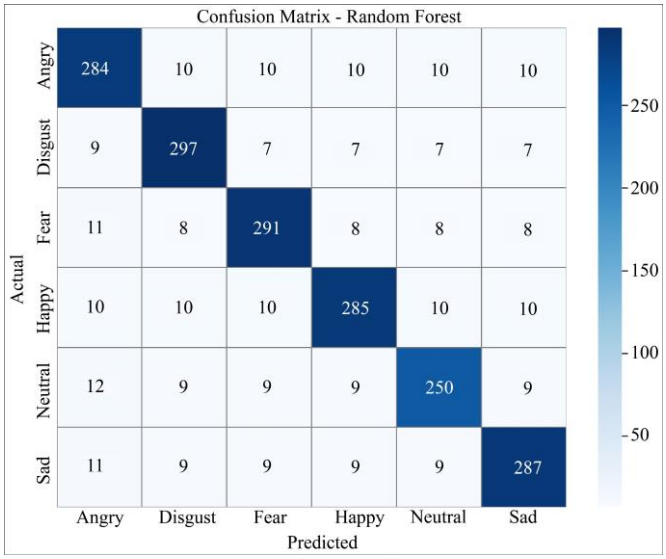

Fig. 6 Random forest confusion matrix

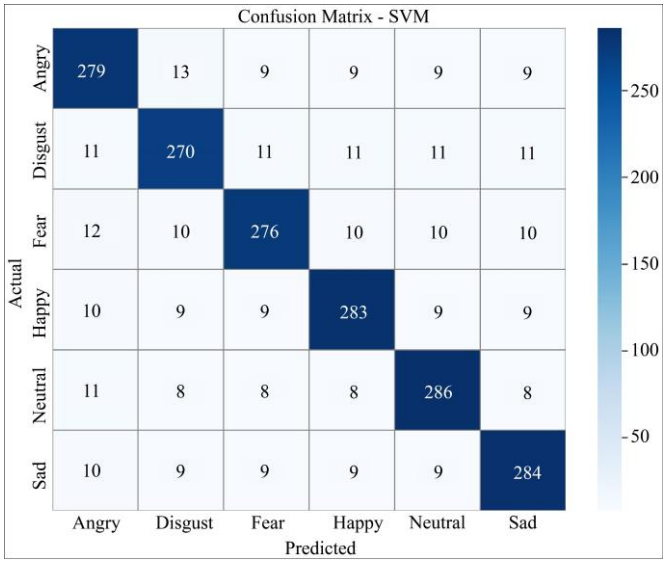

Fig. 7 SVM confusion matrix

Figures 6-8 present confusion matrices for analysing classification results of SVM, Random Forest, and AER-Net models provided in Table 1. The systematic evaluation of AER-Net components took place through a controlled study that measured individual component effects, which can be seen in the training and validation graph in Figure 5.

First, a CNN baseline model was established, which excluded all elements of residual connection, attention modeling, and temporal pattern processing. This model achieved an accuracy of 80.39%, demonstrating the basic effectiveness of simple convolutional feature extraction for speech emotion recognition.

Next, the study introduced residual connections into the CNN backbone [7] to facilitate better feature propagation. However, this CNN + Residuals [2, 5] model resulted in a slightly lower accuracy of 78.86%, suggesting that while residual connections help in training deeper networks, without temporal modeling (e.g., BiLSTM), the network cannot effectively leverage the enriched feature maps for emotion recognition, leading to suboptimal performance.

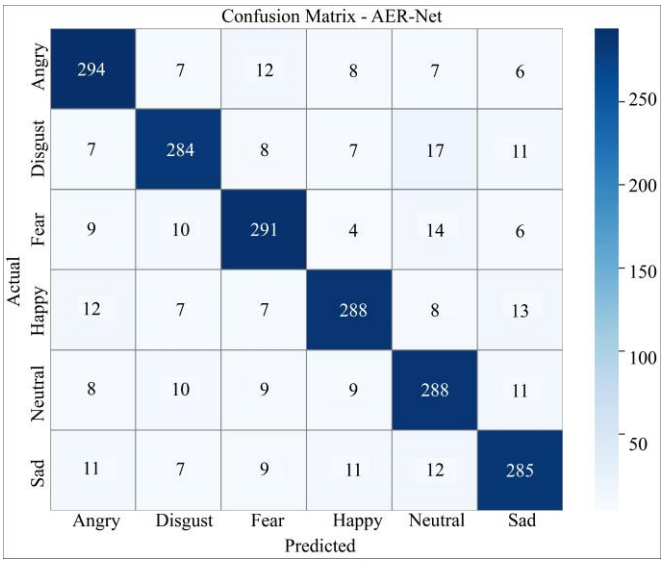

Fig. 8 AERNet confusion matrix

In contrast, the full AER-Net, which integrates residual CNN blocks, BiLSTM temporal modeling, and a custom attention mechanism, achieved a significantly higher accuracy of 87.7%. The BiLSTM layer enabled the model to capture long-range temporal dependencies crucial for emotional expression in speech, while the attention mechanism allowed dynamic focusing on emotionally salient segments within the speech sequence. These components collectively contributed to richer and more discriminative feature learning, demonstrating that both sequential modeling and attention-based focus are critical for robust speech emotion recognition [3, 4].

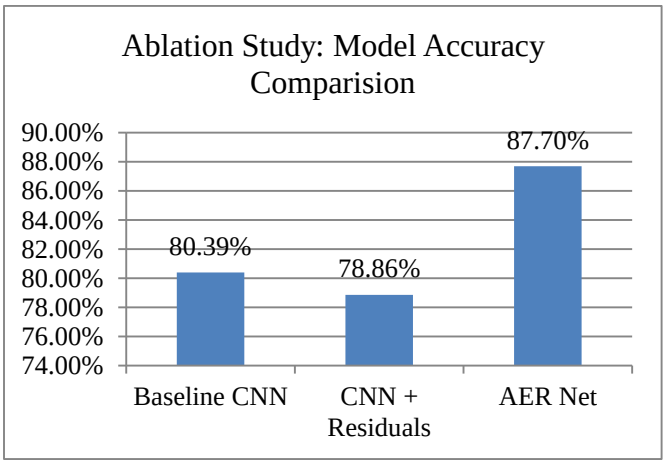

Fig. 9 Ablation study: model accuracy comparison

Table 5. Presents an Ablation Study Summary of Performance Comparison of AER-Net Variants

Model Variant	Residual Connections	BiLSTM	Attention	Accuracy (%)	F1-Score (%) *(approx)*
Baseline CNN	X	X	X	**80.39%**	**~80%**
CNN + Residuals	✓	X	X	**78.86%**	**~79%**
CNN + Residuals + BiLSTM + Attention (**AER-Net**)	✓	✓	✓	**87.7%**	**~88%**

The tested results in Figure 9 and Table 5 support all design choices in the modeled structure, demonstrating the importance of convolutional, sequential and attention-based mechanisms for robust emotion recognition from audio data.

5. Conclusion

This work proposes AER-Net, an Attention-Enhanced Residual Network tailored to address persistent challenges in speech emotion recognition. Combining multi-level acoustic feature extraction, residual learning, bidirectional temporal modeling, and a custom-designed attention mechanism, AER-Net effectively bridges several critical research gaps that have limited the performance and generalizability of previous SER systems. First, the model's architecture directly tackles the issue of imbalanced emotional datasets. The adaptive pooling technique alongside class-weighted training strategies and attention-driven feature focusing demonstrates AER-Net's ability to perform fair recognition of both majority and minority emotion classes with balanced performance results. Traditional approaches tend to fit dominant emotional categories through expensive learning, which causes damage to infrequent emotional expressions. The database design for SER enters a new era through AER-Net by adding residual connections and dynamic attention features that were not sufficiently addressed in similar speech research. The designed architectural choices allow higher efficiency in feature learning processes, which avoids typical failures like gradient vanishing and information disappearance.

AER-Net's shallow acoustic input processing enables the detection of crucial emotional information patterns at different octave and time frequencies. AER-Net exploits multiple levels of representation to enhance its capability for sound generalization when analyzing different speech scenarios and emotional intensity ranges. AER-Net outperforms traditional machine learning approaches on experimental tests that improve accuracy, precision, recall and F1 scores. The confusion matrix analysis generates qualitative findings showing that the model avoids mistake classifications between similar emotions. AER-Net provides a scalable speech-emotion recognition system that defines core design elements for advanced emotional AI architecture development. The presented work tackles essential data imbalance flaws, acoustic learning limitations, and robustness issues to advance emotional detection methods that align with human nature.

5.1. Future Work

AER-Net showcases impressive results and reliable performance, but researchers have multiple possible development paths for future work. The model would serve broader populations better when behavioral data detection functions are expanded to process speech input across different languages and cultures. Research about emotional expression recognition examines facial expressions and physiological signals combined as multi-modal features for enhancing the discovery of emotional contexts. AER-Net researchers should analyze lightweight versions of the model to establish real-time functionality for mobile devices and embedded platforms. Finally, advancing interpretability methods to visualize better and explain the model's decision-making process would contribute to building more transparent and trustworthy emotion-aware AI systems.

References

[1] Md Saroar Jahan et al., "A Comprehensive Study on NLP Data Augmentation for Hate Speech Detection: Legacy Methods, BERT, and LLMs," *arxiv*, pp. 1-31, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Sadil Chamishka et al., "A Voice-Based Real-Time Emotion Detection Technique using Recurrent Neural Network Empowered Feature Modelling," *Multimedia Tools and Applications*, vol. 81, pp. 35173-35194, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Haibin Wu et al., "EMO-SUPERB: An In-depth Look at Speech Emotion Recognition," *arxiv*, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Papers with Code, Speech Emotion Recognition. [Online]. Available: https://paperswithcode.com/task/speech-emotion-recognition?page=9&q=

[5] Sushadevi Shamrao Adagale, and Praveen Gupta, "Speech-based Sentiment Recognition System using PDCNN and LSTM Algorithms," *Research Square*, pp. 1-24, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Minoo Shayaninasab, and Bagher Babaali, "Multi-Modal Emotion Recognition by Text, Speech and Video Using Pretrained Transformers," *arxiv*, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Jiang Li et al., "CFN-ESA: A Cross-Modal Fusion Network with Emotion-Shift Awareness for Dialogue Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 15, no. 4, pp. 1919-1933, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Ravi Raj Choudhary, Gaurav Meena, and Krishna Kumar Mohbey, "Speech Emotion Based Sentiment Recognition using Deep Neural Networks," *Journal of Physics: Conference Series: 2nd International Conference on Computational Intelligence*, vol. 2236, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Rosa A. García-Hernández et al., "A Systematic Literature Review of Modalities, Trends, and Limitations in Emotion Recognition, Affective Computing, and Sentiment Analysis," *Applied Sciences*, vol. 14, no. 16, pp. 1-25, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Hussein Farooq Tayeb Al-Saadawi, Bihter Das, and Resul Das, "A Systematic Review of Trimodal Affective Computing Approaches: Text, Audio, and Visual Integration in Emotion Recognition and Sentiment Analysis," *Expert Systems with Applications*, vol. 225, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Rosa A. Garcia-Hernandez et al., "A Systematic Literature Review of Modalities, Trends, and Limitations in Emotion Recognition, Affective Computing, and Sentiment Analysis," *Applied Sciences*, vol. 14, no. 16, pp. 1-25, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Samson Akinpelu, Serestina Viriri, and Adekanmi Adegun, "An Enhanced Speech Emotion Recognition using Vision Transformer," *Scientific Reports*, vol. 14, no. 1, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Yaoxun Xu et al., "SECap: Speech Emotion Captioning with Large Language Model," *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, Vancouver, Canada, vol. 38, no. 17, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Ziyang Ma et al., "Leveraging Speech PTM, Text LLM, and Emotional TTS for Speech Emotion Recognition," *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, pp. 11146-11150, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Fatima Ahmed, "*Speech Emotion Recognition*," Thesis, Effat University, 2024. [Publisher Link]

[16] Weidong Chen et al., "Vesper: A Compact and Effective Pretrained Model for Speech Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 1711-1724, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Ahdi Hassan et al., *Federated Learning and AI for Healthcare 5.0*, IGI Global Scientific Publishing, pp. 1-391, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Sahil Kashyap et al., "Voice Sentiments Analysis," *International Journal of Progressive Research in Engineering Management and Science*, vol. 4, pp. 1130-1132, 2024. [Publisher Link]

[19] Bahman Mirheidari et al., "Automatic Detection of Expressed Emotion from Five-Minute Speech Samples: Challenges and Opportunities," *PLoS One*, vol. 19, no. 3, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Tyler C. Dalal et al., "Speech based Natural Language Profile before, during and After the Onset of Psychosis: A Cluster Analysis," *Acta Psychiatrica Scandinavica*, vol. 151, no. 3, pp. 332-347, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[21] Tushar Anand et al., *Voice and Speech Recognition Application in Emotion Detection*, IGI Global Scientific Publishing, pp. 242-268, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Santosh Singh et al., "AI Model for Sentiment Analysis System Using Python," *Iconic Research and Engineering Journals*, vol. 7, no. 8, pp. 41-45, 2024. [Publisher Link]

[23] Serena Taylor, and Fariza Fauzi, "Multimodal Sentiment Analysis for the Malay Language: New Corpus using CNN-based Framework," *ACM Transactions on Asian and Low-Resource Language Information Processing*, pp. 1-29, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Yee Sen Tan et al., "Video Sentiment Analysis for Child Safety," *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, Shanghai, China, pp. 783-790, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[25] Nitin Liladhar Rane et al., *Using Artificial Intelligence, Machine Learning, and Deep Learning for Sentiment Analysis in Customer Relationship Management to Improve Customer Experience, Loyalty, and Satisfaction*, Trustworthy Artificial Intelligence in Industry and Society, pp. 233-261, 2024. [CrossRef] [Publisher Link]

[26] Moshiur Rahman Faisal et al., "Bengali & Banglish: A Monolingual Dataset for Emotion Detection in Linguistically Diverse Contexts," *Data in Brief*, vol. 55, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[27] Jinchuan He, "*Voice Emotion Recognition Using Natural Language Processing Deep Learning*," Master Thesis, Northeastern University, pp. 1-18, 2022. [Google Scholar] [Publisher Link]

[28] Sophina Luitel, Yang Liu, and Mohd Anwar, "Investigating Fairness in Machine Learning-Based Audio Sentiment Analysis," *AI and Ethics*, vol. 5, pp. 1099-1108, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[29] Jianbo Zhao et al., "Sentiment Analysis of Video Danmakus based on MIBE-RoBERTa-FF-BiLSTM," *Scientific Reports*, vol. 14, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[30] Yicheng Zhong et al., "ExpCLIP: Bridging Text and Facial Expressions via Semantic Alignment," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, Canada, vol. 38, no. 7, pp. 7614-7622, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[31] A. Tepecik and E. Demir, "Emotion Detection with Pre-Trained Language Models BERT and ELECTRA Analysis of Turkish Data," *Intelligent Methods in Engineering Sciences*, vol. 3, no. 1, pp. 7-12, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[32] Xinfa Zhu et al., "METTS: Multilingual Emotional Text-to-Speech by Cross-Speaker and Cross-Lingual Emotion Transfer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1506-1518, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[33] Suja Panickar et al., "Sentiment Analysis of Custom Speech Corpus: A proof of concept for NLP," *Procedia Computer Science*, pp. 220-228, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[34] Jie Shan, "Application of Natural Language Processing-based Emotional Semantic Analysis in the 'One Core, Three Integrations' Vocal Music Teaching Model," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[35] Pamela Lopes da Cunha et al., "Automated Free Speech Analysis Reveals Distinct Markers of Alzheimer's and Frontotemporal Dementia," *PLoS One*, vol. 19, no. 6, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[36] Galym Kapyshev, Marat Nurtas, and Aizhan Altaibek, "Speech Recognition for Kazakh Language: A Research Paper," *Procedia Computer Science*, vol. 231, pp. 369-372, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[37] I. Venkata Dwaraka Srihith, and S. Ashok, "The Voice of Feeling: Exploring Emotions via Machine Learning," *Research and Reviews: Advancement in Cyber Security*, vol. 1, no. 3, pp. 1-16, 2024. [CrossRef] [Publisher Link]

[38] Mingyu Ji et al., "SS-Trans (Single-Stream Transformer for Multimodal Sentiment Analysis and Emotion Recognition): The Emotion Whisperer—A Single-Stream Transformer for Multimodal Sentiment Analysis," *Electronics*, vol. 13, no. 21, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[39] Zhicheng Liu et al., "Ensemble Pretrained Models for Multimodal Sentiment Analysis using Textual and Video Data Fusion," *WWW '24: Companion Proceedings of the ACM Web Conference 2024*, Singapore, pp. 1841-1848, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[40] Yuki Konishi, and Yoshihiro Tanaka, "An Emotional Expression System with Vibrotactile Feedback during the Robot's Speech," *arxiv*, pp. 1-4, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[41] H. Kuchuk, and A. Kuliahin, "Hybrid Recommender for Virtual Art Compositions with Video Sentiments Analysis," *Advanced Information Systems*, vol. 8, no. 1, pp. 70-79, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[42] Xingqun Qi et al., "EmotionGesture: Audio-Driven Diverse Emotional Co-Speech 3D Gesture Generation," *IEEE Transactions on Multimedia*, vol. 26, pp. 10420-10430, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[43] Jonathan Lo, and Elsie Wang, Flynn O'sullivan, Auditing Sentiment Analysis Algorithms for Bias. [Online]. Available: https://dsc180b.lojot.com/

[44] Viraj Nishchal Shah et al., "Investigation of Imbalanced Sentiment Analysis in Voice Data: A Comparative Study of Machine Learning Algorithms," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 11, no. 6, pp. 1-12, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[45] Nirmalya Thakur et al., "A Labelled Dataset for Sentiment Analysis of Videos on YouTube, TikTok, and Other Sources about the 2024 Outbreak of Measles," *26th International Conference on Human-Computer Interaction, HCII 2024*, Washington, DC, USA, pp. 220-239, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[46] Ghulam Ali Amiri et al., "Decoding Gender Representation and Bias in Voice User Interfaces (VUIs)," *International Journal of Computer Science and Mobile Computing*, vol. 13, no. 5, pp. 76-88, 2024. [CrossRef] [Publisher Link]

[47] Huan Zhao, Nianxin Huang, and Haijiao Chen, "Knowledge Enhancement for Speech Emotion Recognition via Multi-Level Acoustic Feature," *Connection Science*, vol. 36, no. 1, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[48] Chao He et al., "Mixture of Attention Variants for Modal Fusion in Multi-Modal Sentiment Analysis," *Big Data and Cognitive Computing*, vol. 8, no. 2, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[49] Houwei Cao et al., "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377-390, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[50] Kate Dupuis, and M. Kathleen Pichora-Fuller, "*Toronto Emotional Speech Set (TESS)*," University of Toronto, 2010. [Publisher Link]