Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Leveraging Transfer Learning for Speech Emotion Recognition Using ResNet50

Veeranjaneyulu Janagam

jangam.vfstr@gmail.com

Vignan's Foundation for Science, Technology and Research

P Siva Prasad

Vignan's Foundation for Science, Technology and Research

Article

Keywords: Speech Emotion Recognition, Deep learning, Transfer Learning Resnet50

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11178272/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Leveraging Transfer Learning for Speech Emotion Recognition Using ResNet50

J. Veeranjaneyulu[1], P.Siva Prasad[2]
[1,2]Department of Computer Science and Engineering,
Vignan's Foundation for Science, Technology and Research, Vadlamudi, AP, India
Email: jangam.vfstr@gmail.com, pusapatisivaprasad@gmail.com
Corresponding author: J. Veeranjaneyulu (jangam.vfstr@gmail.com)

***Abstract*—This paper introduces a speech emotion recognition (SER) system based on transfer learning with ResNet50. It uses the deep hierarchical features of Mel spectrograms of speech signals to precisely classify emotions. We improve feature learning while ensuring computational efficiency by freezing the first 143 layers of ResNet50 and fine-tuning the last layers. The model is trained on the TESS dataset, which contains seven emotions: anger, disgust, fear, happiness, neutral, surprise, and sadness. With data augmentation and optimized hyperparameters, the suggested model obtains a classification accuracy of 98.60%, surpassing traditional SER methods. The experimental results prove that transfer learning based on ResNet50 can greatly enhance the generalization and robustness of SER models, which makes it an ideal choice for real-world human-computer interaction applications.**

Keywords – Speech Emotion Recognition, Deep learning, Transfer Learning Resnet50

I. Introduction

Speech signals are the quickest and most natural way for people to communicate. As a result, speech serves as a primary means of establishing a quick connection between humans and computers. This makes speech a rich medium embedded with valuable information, making it highly significant in modern human-computer interaction [1]. Speech is a remarkably simple and fluid form of communication that has been demonstrated to efficiently and quickly convey information in the modern era. Nowadays, considerable time and effort are dedicated to teaching people how to communicate with various smart devices using speech signals [2]. Along with many significant attitudes, emotions are largely unique and powerful mental processes that can be understood through simple external observations. All physical actions, including speech, facial expressions, and body language, are fundamental components of an individual's emotional state [3]. Recognizing emotions through speech, known as speech emotion recognition (SER), has become a critical area of research. Numerous fields, including medicine, education, criminal investigation, and elder care, have found use for speech emotion recognition [4]. The ability to analyze emotional states from speech signals enables advancements in mental health monitoring, personalized education, and even criminal investigations. Many distinguishing characteristics can be extracted from speech to identify the emotions embedded in it. These characteristics include continuous features, qualitative features, and spectral features, which are analyzed to develop effective SER systems [5]. MFCC, Log Mel-Spectrogram, Chroma, Spectral Centroid, and Spectral Rolloff have all been extracted using this method. Audio signals were directly subjected to feature extraction using the Librosa library [6]. The two phases of feature extraction and emotion classification determine the effectiveness and resilience of SER models. Robust feature extraction enables the model to gain detailed knowledge about acoustic characteristics such as pitch, formant frequencies, and vocal tract features. These characteristics are crucial for identifying various emotions in speech. Generally, spectral analysis and temporal analysis are the two primary methods for extracting speech features. Conventional approaches rely on manually developed techniques to extract local features, from which statistical calculations are made to determine global features. SER deep learning models enhance generalization, robustness, and accuracy by dynamically learning both local and global features essential to the task [7]. A hybrid model of CNN-LSTM is designed for speech emotion recognition to capture both the local and global context information in terms of log-Mel spectrogram features. The structure involves four Local Feature-Learning Blocks (LFLBs) which include convolutional layers, batch normalization, exponential linear unit (ELU) activation functions, and max-pooling layers that extract the local correlations of the input features. In order to discover long-range dependencies in the sequential data, the blocks create features that are passed into an LSTM layer. This is a combination strategy that is able to process speech signals effectively and has high classification accuracy for emotions. The objective of this study is to develop a robust speech emotion recognition system using a CNN-LSTM hybrid model and hybrid feature fusion techniques. By leveraging diverse datasets and advanced feature extraction methods, this work aims to improve the generalization and accuracy of SER systems, contributing to their broader applicability in real-world scenarios.

II. Related Work

Koppula et al. [8] proposed a Firefly-based Recurrent Neural Speech Recognition (FbRNSR) model for Speech Emotion Recognition (SER). Their approach integrates a firefly optimization algorithm with a Recurrent Neural Network (RNN)

to enhance emotion classification. The steps involve noise removal, feature extraction (amplitude, frequency, and pitch), and classification based on a firefly-based fitness function. This optimization improves accuracy, precision, and recall while decreasing computational complexity, making it more efficient compared to conventional SER approaches. Cimtay et al. [9] suggested a multimodal emotion recognition system by combining facial expressions, galvanic skin response (GSR), and electroencephalogram (EEG) signals. Their hybrid fusion system uses feature-level and decision-level fusion to enhance recognition accuracy. The technique utilizes the strengths of various modalities and is more accurate in subject-independent emotion recognition tasks on databases such as DEAP and LUMED-2. The technique is especially helpful in identifying true emotions in situations where facial expressions can be deceptive. Gerczuk et al. [10] presented EmoNet, a transfer learning-based approach to Speech Emotion Recognition (SER). EmoNet applies EMOSET, a unified dataset merging 26 SER datasets with more than 84,000 audio files. By using ResNet architecture with residual adapters, EmoNet generalizes across datasets, enhancing performance in 21 out of 26 datasets. The paper emphasizes the significance of transfer learning in processing various SER datasets. S. Ramesh et al. [11] proposed a hybrid SER model integrating Naïve Bayes classifier and Gray Wolf Optimizer (GWO). The model learns speech features in the form of Mel-Frequency Cepstral Coefficients (MFCC) and performs optimization of classification through GWO. Experimental evaluation on a combined dataset of SAVEE and TESS, the method enhances accuracy, precision, and efficiency in comparison to conventional machine learning techniques. Falahzadeh et al. [12] presented a hybrid SER approach based on GWO and a Deep Convolutional Neural Network (DCNN). They presented "Chaogram," which transforms 1D speech signals into 2D images that are processed through a VGG16-based DCNN. The model is optimized with GWO to improve accuracy at the expense of less computational cost. EMO-DB and eNTERFACE05 experiments verify its better performance compared to traditional SER approaches. Zhu et al. [13] developed a three-layer brain-inspired multimodal emotion recognition model, comprising feature extraction, fusion, and classification. It handles text, audio, and visual information using cross-attention for combining the information and softmax function for classification. When tested on the IEMOCAP dataset, it recorded an 82.7% recognition rate with excellent inter-modal learning. Pan et al. [14] created a multimodal model fusing facial, speech, and EEG information. Facial is identified using GhostNet, speech by LFCNN, and EEG by tLSTM with decision-level fusion making the ultimate prediction. Improved accuracy over previous models was established during experiments conducted using CK+, EMO-DB, and MAHNOB-HCI datasets. Alsabhan et al. [15] proposed a hybrid real-time SER model based on 2D CNN, 1D CNN, LSTM, and attention mechanisms. The model employs the feature extraction using ZCR, RMSE, and MFCC, and the 1D CNN along with LSTM and attention layers to improve classification while maintaining temporal relationships. It is tested on EMO-DB, SAVEE, ANAD, and BAVED databases and is shown to have high accuracy and robust generalization for multilingual SER tasks. Singh et al. [16] suggested a gender-dependent SER model with a separately trained 1D CNN, extracting features via MFCC, delta-MFCC, and delta-delta-MFCC to form 16 separate models for various emotions. This is simpler compared to a deeper CNN baseline and performs 6.9% accuracy improvement on the RAVDESS dataset for six emotions. The research underlines the advantage of gender-specific training in enhanced SER accuracy at reduced computational complexity. Bautista et al. [17] suggested a parallel CNN-Transformer model for speech emotion recognition (SER) with mel-spectrograms as inputs. Data augmentation methods such as noise addition and spectral distortions enhanced the performance of the model on the RAVDESS dataset. The method performed better than conventional CNN-based models at low computational expense. Alluhaidan et al. [18] presented a SER model that integrated MFCC and time-domain features into a master feature vector. For classification, they employed a lightweight 1D CNN with excellent accuracy on EMO-DB (97%), SAVEE (93%), and RAVDESS (92%), outclassing conventional models. Sahoo et al. [19] proposed TLEFuzzyNet, an ensemble of pre-trained CNNs (ResNet18, InceptionV3, GoogleNet) for SER based on fuzzy rank-based fusion. The proposed model obtained the best accuracy rates: 98.57% (SAVEE), 99.38% (EmoDB), and 99.66% (RAVDESS), proving the efficiency of transfer learning and ensemble strategies. Min et al. [20] solved the problem of dataset imbalance in SER using a Sparse Learning-Based Fusion Model (SLBF). It fuses CNN-SEN and MLSTM-FCN, merging spectral and temporal features. Through retraining weak neurons and freezing well-trained ones, it attained 1.13%–9.73% accuracy gains on EMOVO, RAVDESS, SAVEE, and EMO-DB datasets. Dal R et al. [21] used a CNN with an attention mechanism for SER, with MFCCs as inputs. The model's accuracy was 83% (RAVDESS), 100% (TESS), 68% (CREMA-D), and 63% (IEMOCAP), showing good generalizability across datasets. Oh et al. [22] emphasized noise-resilient recognition of emotions through facial expressions utilizing a CNN-driven model. Having been trained in salt-and-pepper noise levels up to 20%, the system was keeping classification accuracy of 68– 99%, exhibiting its robustness in real life noisy conditions.

III. Methodology

This section describes the procedure for implementing a transfer learning method in the deployment of a Speech Emotion Recognition (SER) model. The proposed model enhances feature extraction and classification accuracy using pre-trained deep learning models. Data preprocessing, feature extraction, training, evaluation, and transfer learning-based model development are all part of the methodology.

A. Preprocessing

The preprocessing pipeline encompasses a number of steps to map raw audio data into structured inputs for deep learning-

based emotion classification. To do this, initially, every audio file is read and resampled to a set duration of 3 seconds and a sampling frequency of 22,050 Hz. The audio is then converted into a mel spectrogram with 128 mel filter banks and a maximum frequency of 8,000 Hz and then converted to the decibel (dB) scale for more effective feature representation. To make the input size consistent, spectrograms are padded or truncated to a fixed size of (128, 130). As ResNet50 demands three-channel input, the one-channel spectrograms are replicated over three channels. Moreover, data augmentation methods including rotation, zoom, shifts, and flipping are used to enhance model generalization. These processes guarantee well-structured data after processing, making the accuracy and robustness of the emotion recognition model improved.

B. Feature Extraction

The process of feature extraction in the code given here is the transformation of raw speech signals into Mel Spectrograms that serve as inputs for the deep learning model. The primary feature extraction processes involve:

1) Short-Time Fourier Transform (STFT): The first step is converting the time-domain audio signal into a time-frequency representation using STFT. The STFT of a signal $x(t)$ is computed as:

$$X(m,k) = \sum_{n=0}^{N-1} x(n)\, w(n - mR)\, e^{-j2\pi kn/N} \tag{1}$$

where:

- $X(m,k)$ represents the STFT coefficient at frame m and frequency bin k,
- $x(n)$ denotes the discrete audio signal,
- $w(n)$ is the windowing function,
- R represents the hop length,
- N denotes the FFT size.

This transforms the signal into a spectrogram that represents amplitude across time and frequency.

2) Mel Spectrogram Calculation: The spectrogram obtained from STFT is mapped onto the Mel scale, which mimics human auditory perception. This is done by applying a Mel filter bank to the power spectrogram:

$$S_{\text{mel}}(m,j) = \sum_{k=0}^{N/2} |X(m,k)|^2 H_j(k) \tag{2}$$

where:

- $S_{\text{mel}}(m,j)$ represents the Mel spectrogram,
- $|X(m,k)|^2$ denotes the power spectrogram obtained from STFT,
- $H_j(k)$ is the j^{th} Mel filter.

The Mel-scale frequency f_{mel} is computed as:

$$f_{\text{mel}} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{3}$$

where f represents the original frequency measured in Hertz (Hz).

3) Logarithmic Power Scaling (Decibel Conversion): Since human auditory perception follows a logarithmic response to sound intensity, the Mel spectrogram is transformed into the decibel (dB) scale as follows:

$$S_{\text{dB}} = 10 \log_{10}(S_{\text{mel}} + \epsilon) \tag{4}$$

where ϵ is a small constant added to prevent numerical instability caused by $\log(0)$ issues.

C. Proposed Model

The proposed framework employs transfer learning using the ResNet50 architecture for speech emotion recognition. Mel spectrograms generated from raw speech signals are utilized as input features. Since ResNet50 requires three-channel input images, the single-channel spectrograms are replicated across three channels before being fed into the network. The pretrained ResNet50 model, originally trained on the ImageNet dataset, is adopted as a feature extractor. To preserve the generic visual features learned during pretraining, the initial 143 layers are frozen, while the remaining layers are fine-tuned for the speech emotion recognition task. The extracted deep feature representations are subsequently processed using a Global Average Pooling (GAP) layer to reduce feature dimensionality and computational complexity. A fully connected dense layer with L2 regularization is employed to minimize overfitting and improve generalization performance. In addition, a dropout layer with a dropout rate of 0.3 is incorporated to further enhance model robustness. Finally, a softmax classification layer consisting of seven output neurons is used to classify speech samples into their corresponding emotional categories. The model training process utilizes the Adam optimizer with a learning rate of 0.0001. Sparse categorical cross-entropy is adopted as the loss function to efficiently handle encoded class labels. Furthermore, data augmentation techniques such as rotation, zooming, and shifting are applied during training to improve model robustness and classification performance. By leveraging pretrained hierarchical representations and adapting them to speech-based emotional patterns, the proposed approach achieves improved recognition accuracy and computational efficiency.

1) Feature Extraction using ResNet50: We employ a pretrained ResNet50 model to extract deep hierarchical features from the spectrograms. To retain general feature representations, we freeze the first 143 layers of ResNet50 while fine-tuning the remaining layers for domain adaptation. The feature extraction process can be defined as:

$$F = \text{ResNet50}(M(t,f); \theta_f) \tag{5}$$

where F represents the extracted feature vector, and θ_f are the trainable parameters of the fine-tuned ResNet50 layers.

2) Classification Layers: The extracted features are passed through a classification head consisting of a Global Average Pooling (GAP) layer, a fully connected (FC) layer, and a softmax output layer. The classification output is computed as:

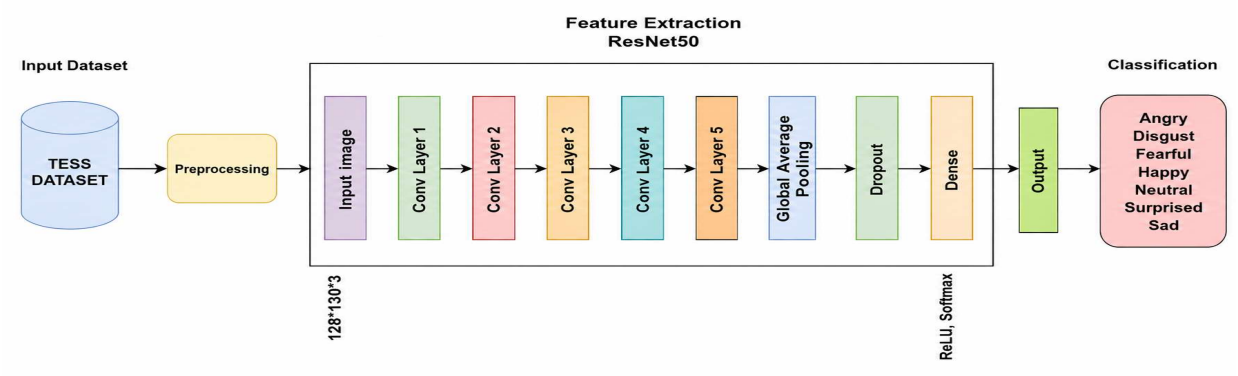


Fig. 1. Proposed Model Architecture for Speech Emotion Recognition using ResNet50

$$P(y|X) = \text{softmax}(W_c F + b_c) \tag{6}$$

where W_c and b_c are the learnable weights and biases of the final FC layer, and $P(y|X)$ represents the predicted probability distribution over the eight emotion classes.

The proposed architecture Fig 1 efficiently utilizes transfer learning, improving performance while maintaining low computational complexity. To enhance the interpretability of spectral information, the power spectrogram is converted into a logarithmic scale (Log-Mel Spectrogram) using a decibel (dB) transformation:

D. Hyperparameters Optimization

The suggested model uses a set of specifically chosen hyperparameters to balance performance and generalization. The spectrograms used as input are resized to $(128, 130, 3)$ to accommodate the expected size of ResNet50. The learning rate is set to 0.0001 using the Adam optimizer to ensure stable convergence. A batch size of 32 is chosen for efficient training, and the model is trained for 50 epochs to balance performance and training time. L2 regularization $(\lambda = 0.01)$ is applied to the dense layer to prevent overfitting. A dropout rate of 0.3 is introduced in the classification layers to enhance generalization. The first 143 layers of ResNet50 are frozen, allowing only the top layers to be fine-tuned for emotion classification. The softmax activation function is used in the final output layer to classify emotions into eight categories. These hyperparameters are optimized to ensure efficient learning and robust emotion recognition performance.

$$S_{\text{dB}}(m,t) = 10 \times \log_{10}(S(m,t)) \tag{7}$$

This transformation helps make subtle variations in frequency and amplitude more distinguishable, improving feature representation.

Since audio recordings vary in duration, the extracted spectrograms are resized to a fixed shape of $(128, 130)$ time frames. If a spectrogram is shorter than the target length, zero-padding is applied as follows:

$$S_{\text{padded}}(m,t) = \begin{cases} S_{\text{dB}}(m,t), & \text{if } t < T_{\text{target}} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $T_{\text{target}} = 130$ represents the target number of time frames.

If the spectrogram is longer than the target, truncation is applied:

$$S_{\text{truncated}}(m,t) = S_{\text{dB}}(m,t), \quad \text{for } t \leq T_{\text{target}} \tag{9}$$

This ensures that all input spectrograms have a consistent shape, which is necessary for deep learning models to process the data effectively.

IV. EXPERIMENTAL RESULTS

This section presents the performance evaluation of the proposed Speech Emotion Recognition (SER) model. The experiments were conducted to assess the effectiveness of transfer learning using ResNet50 for feature extraction and classification.

A. Dataset and Experimental Setup

The TESS (Toronto Emotional Speech Set) dataset is employed for speech emotion recognition in this study. It consists of 200 target words from two female speakers, representing seven emotions: anger, disgust, fear, happiness, neutral, surprise, and sadness. The dataset contains 2800 audio recordings specifically designed to mimic natural emotional speech. The audio recordings are captured in a controlled setting with quality audio to guarantee clarity and consistency. TESS is a commonly used database for emotion recognition tasks because of its well-balanced distribution of emotions and capacity to record expressive speech variations, which makes it a rich dataset for deep learning-based emotion classification.

B. Model Training and Testing

The model is trained under transfer learning with ResNet50 as the base. The TESS dataset is preprocessed to transform audio signals to Mel spectrograms and used as input images with three channels.Data augmentation is used for training

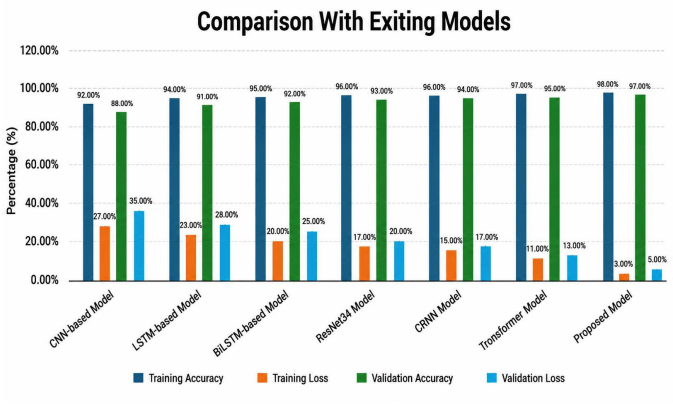


Fig. 2. Performance Analysis Of Proposed Model With Other Models

and testing sets to improve generalization. ResNet50 is fine-tuned by freezing the first 143 layers and then training the other layers, along with custom dense layers for emotion classification. The model is trained on 50 epochs with the Adam optimizer and sparse categorical cross-entropy loss. Performance is measured after training through accuracy metrics, a confusion matrix, and a classification report by comparing predictions against true labels in the test set. Fig. 2 shows the comparative performance analysis of the proposed model with existing models.

C. Performance Evaluation

From the Table I, it can be observed that the proposed speech emotion recognition model achieved strong performance in terms of precision, recall, and F1-score across different emotional classes. The obtained results indicate that the model can effectively distinguish between various emotional categories with high classification accuracy. Most emotional classes achieved precision and recall values greater than 95%, demonstrating the robustness and reliability of the proposed approach. In particular, emotions such as Sad, Happy, and Neutral exhibited higher recognition performance with nearperfect F1-score values, indicating efficient feature extraction and classification capabilities. The overall accuracy of 98.60% further confirms the effectiveness of the transfer learningbased ResNet50 framework in capturing discriminative emotional patterns from speech signals.

V. DISCUSSION

This work investigates the efficacy of transfer learning employing ResNet50 for speech emotion recognition (SER). Utilizing pre-trained hierarchical features, the model attains high accuracy (98.60%) in emotion classification from Mel spectrograms. Freezing the first 143 layers and fine-tuning the rest enhances generalization and computation savings. Robustness is further enhanced using data augmentation techniques. Although the model works well, there are issues with separating similar emotions and fitting different datasets. Future research must emphasize dataset extension, multimodal feature integration, and optimizing the model for real-time

TABLE I
PERFORMANCE ANALYSIS OF THE EMOTION RECOGNITION MODEL BASED ON PRECISION, RECALL, AND F1-SCORE. THE MODEL OBTAINED A TOTAL ACCURACY OF 98.60%, DEMONSTRATING HIGH EFFICIENCY IN CORRECTLY IDENTIFYING EMOTIONS

Emotion	Precision	Recall	F1-Score
Angry	0.97	0.95	0.97
Disgust	0.95	0.97	0.99
Fearful	0.95	0.98	0.97
Happy	0.98	0.97	0.99
Neutral	0.99	0.96	0.98
Sad	0.99	0.99	0.99
Surprised	0.99	0.89	0.99
Accuracy = 98.60%			

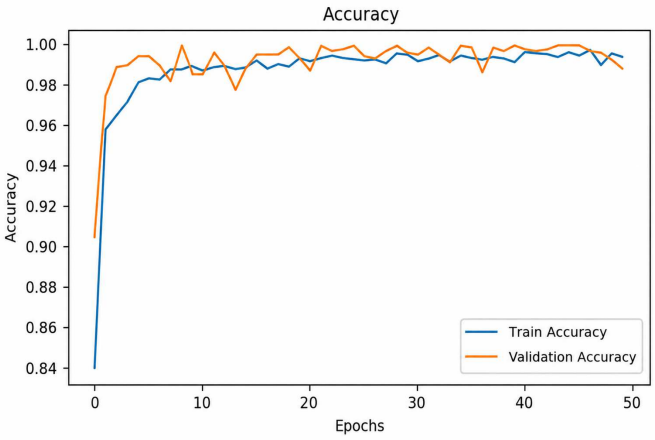


Fig. 3. Training and Validation Accuracy over Epochs

human-computer interaction. Fig. 3 presents the training and validation accuracy curves of the proposed model. Fig. 4 illustrates the training and validation loss curves during model training. Fig. 5 shows the confusion matrix of the proposed speech emotion recognition model.

VI. CONCLUSION

The present research proves the efficiency of transfer learning along with ResNet50 for the task of speech emotion recognition. With the fine-tuning of deep hierarchical features from Mel spectrograms, the model obtains an accuracy of 98.60% on the TESS dataset. The experimental outcome explains the strengths of ResNet50 in the detection of complex speech features, resulting in higher performance than conventional techniques. The work also verifies that transfer learning promotes model generalization with lower computational expense. It is possible for future research to expand the dataset with more heterogeneous emotional expressions and further

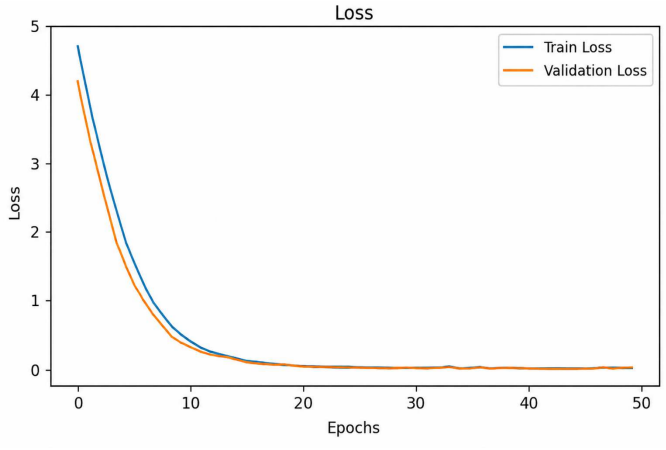


Fig. 4. Training and Validation Loss over Epochs

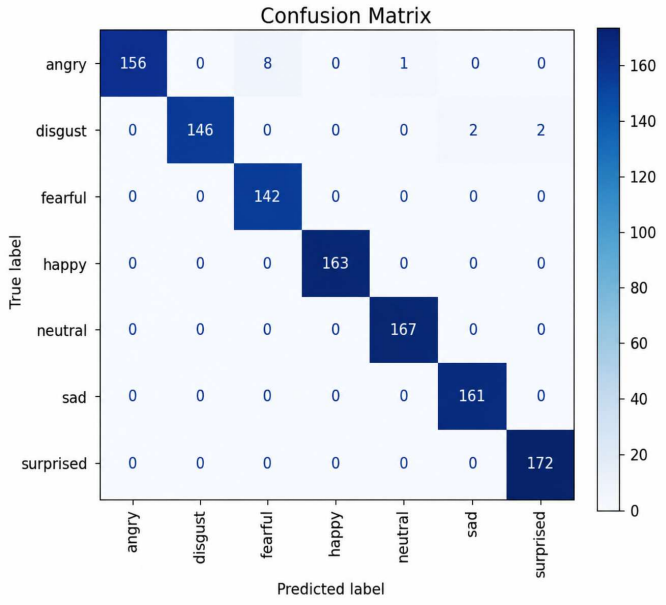


Fig. 5. Confusion Matrix of the Model Predictions

enhance the model's adaptability in noisy conditions. The methodology has deep-rooted implications in applications in mental health tracking, smart assistants, and human-computer interaction systems.

Ethics Declaration

The authors declare no competing financial interests

Funding

The authors received no specific funding for this work.

References

[1] M. J. Al-Dujaili and A. Ebrahimi-Moghadam, "Speech emotion recognition: a comprehensive survey," *Wireless Personal Communications*, vol. 129, no. 4, pp. 2525–2561, 2023.

[2] S. Akinpelu, S. Viriri, and A. Adegun, "Lightweight deep learning framework for speech emotion recognition," *IEEE Access*, 2023.

[3] M. B. Er, "A novel approach for classification of speech emotions based on deep and acoustic features," *IEEE Access*, vol. 8, pp. 221640–221653, 2020.

[4] L. M. Zhang, et al., "A parallel-model speech emotion recognition network based on feature clustering," *IEEE Access*, 2023.

[5] A. A. Abdelhamid, et al., "Robust speech emotion recognition using CNN+LSTM based on stochastic fractal search optimization algorithm," *IEEE Access*, vol. 10, pp. 49265–49284, 2022.

[6] A. Aggarwal, et al., "Two-way feature extraction for speech emotion recognition using deep learning," *Sensors*, vol. 22, no. 6, p. 2378, 2022.

[7] S. Kakuba, A. Poulose, and D. S. Han, "Deep learning approaches for bimodal speech emotion recognition: Advancements, challenges, and a multi-learning model," *IEEE Access*, 2023.

[8] N. Koppula, et al., "A novel optimized recurrent network-based automatic system for speech emotion identification," *Wireless Personal Communications*, vol. 128, no. 3, pp. 2217–2243, 2023.

[9] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020.

[10] M. Gerczuk, et al., "EmoNet: A transfer learning framework for multicorpus speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1472–1487, 2021.

[11] S. Ramesh, et al., "Automatic speech emotion detection using hybrid of gray wolf optimizer and naïve Bayes," *International Journal of Speech Technology*, vol. 26, no. 3, pp. 571–578, 2023.

[12] M. R. Falahzadeh, et al., "Deep convolutional neural network and gray wolf optimization algorithm for speech emotion recognition," *Circuits, Systems, and Signal Processing*, vol. 42, no. 1, pp. 449–492, 2023.

[13] X. Zhu, et al., "Emotion recognition based on brain-like multimodal hierarchical perception," *Multimedia Tools and Applications*, vol. 83, no. 18, pp. 56039–56057, 2024.

[14] J. Pan, et al., "Multimodal emotion recognition based on facial expressions, speech, and EEG," *IEEE Open Journal of Engineering in Medicine and Biology*, 2023.

[15] W. Alsabhan, "Human–computer interaction with a real-time speech emotion recognition with ensembling techniques 1D convolution neural network and attention," *Sensors*, vol. 23, no. 3, p. 1386, 2023.

[16] V. Singh and S. Prasad, "Speech emotion recognition system using gender dependent convolution neural network," *Procedia Computer Science*, vol. 218, pp. 2533–2540, 2023.

[17] J. L. Bautista, Y. K. Lee, and H. S. Shin, "Speech emotion recognition based on parallel CNN-attention networks with multi-fold data augmentation," *Electronics*, vol. 11, no. 23, p. 3935, 2022.

[18] A. S. Alluhaidan, et al., "Speech emotion recognition through hybrid features and convolutional neural network," *Applied Sciences*, vol. 13, no. 8, p. 4750, 2023.

[19] K. K. Sahoo, et al., "TLEFuzzyNet: fuzzy rank-based ensemble of transfer learning models for emotion recognition from human speeches," *IEEE Access*, vol. 9, pp. 166518–166530, 2021.

[20] D. J. Min and D. H. Kim, "Speech Emotion Recognition via Sparse Learning-based Fusion Model," *IEEE Access*, 2024.

[21] F. A. Dal Rì, F. C. Ciardi, and N. Conci, "Speech Emotion Recognition and Deep Learning: an Extensive Validation using Convolutional Neural Networks," *IEEE Access*, 2023.

[22] S. Oh and D. K. Kim, "Noise-Robust Deep Learning Model for Emotion Classification using Facial Expressions," *IEEE Access*, 2024.

Contributions

Both authors contributed equally to the conceptualization, methodology, implementation, experimentation, analysis, and interpretation of the results. Both authors substantially contributed to writing, reviewing, and revising the manuscript and approved the final version.