Citation: Ayush Porwal, Praveen Kumar Tyagi, Ajay Sharma, Dheeraj Kumar Agarwal. Deep learning-based speech emotion recognition: Leveraging diverse datasets and augmentation techniques for robust modeling. *Journal of Harbin Institute of Technology (New Series)*. DOI: 10.11916/j.issn.1005-9113.2024005.

Deep Learning-Based Speech Emotion Recognition: Leveraging Diverse Datasets and Augmentation Techniques for Robust Modeling

Ayush Porwal[1], Praveen Kumar Tyagi[2], Ajay Sharma[3] and Dheeraj Kumar Agarwal[2]*

(*1. Department of Electronics and Instrumentation Engineering, Shri G.S.Institute of Technology and Science, Indore 452001, Madhya Pradesh, India;*
2. Department of Electronics and Communication Engineering, Maulana Azad National Institute of Technology, Bhopal 462003, Madhya Pradesh, India;
3. School of Computing Science and Engineering, VIT Bhopal University, Kothrikalan, Sehore, 466114, Madhya Pradesh, India)

Abstract: In recent years, Speech Emotion Recognition (SER) has developed into an essential instrument for interpreting human emotions from auditory data. The proposed research focuses on the development of a SER system employing deep learning and multiple datasets containing samples of emotive speech. The primary objective of this research endeavor is to investigate the utilization of Convolutional Neural Networks (CNNs) in the process of sound feature extraction. Stretching, pitch manipulation, and noise injection are a few of the techniques utilized in this study to improve the data quality. The investigation includes coverage of these methods. Feature extraction methods including Zero Crossing Rate, Chroma_stft, MFCC, RMS, and MelSpectogram are used to train a model. By using these techniques, audio signals can be transformed into recognized features that can be utilized to train the model. Ultimately, the study will produce a thorough evaluation of the model's performance. When this method was applied, the model achieved an impressive accuracy of 94.57% on the test dataset. Proposed work also validated on EMO-BD and IEMOCAP dataset. These consist of further data augmentation, feature engineering, and hyperparameter optimization. By following these development paths, SER systems will be able to be implemented in real-world scenarios with greater accuracy and resilience.

Keywords: voice signal; emotion recognition; deep learning; CNN
CLC number: TN18, TN912.3 **Document code**: A **Article ID**: 1005-9113(2024)00-0000-12

0 Introduction

Acoustic signals are created by the human vocal tract during the process of speech generation. Speech signals are acoustic signals[1]. In the form of sound waves, they carry information that is used to express spoken language. Examples of words, moods, intonations, and other linguistic elements are provided in this information. The waveform of these signals is what sets them apart, moreover, it shows how changes in air pressure are caused by the movement of the vocal cords, the articulation of the tongue, lips, and palate, and the modulation of airflow via the vocal tract. These signals' distinctive features are brought about by these variations in air pressure. Since their inception, voice signals have been primarily analyzed through the lens of signal processing. This discipline employs a diverse range of methodologies to extract relevant data from the signals that are being analyzed. To comprehend and discern spoken language, one must possess knowledge of the attributes associated with pitch, intensity, frequency, and temporal patterns. Possessing these attributes is crucial for understanding communication. The aforementioned qualities are inherent in this substance. Speech signals serve as the auditory manifestation of spoken words, providing insights into the intricate physiological processes that contribute to their generation. The respiratory system, in conjunction with the articulators (mouth, lips, and palate) and vocal cords, generates these transmissions. Speech signals are generated

Received 2024-01-15.
* Corresponding author: Ayush Porwal, Bachelor. Email: porwal.ayush2002@ gmail.com.

through the modulation of ventilation and air pressure. These signals utilize sound vibrations to transmit linguistic information. Waveform structures, frequencies, amplitudes, and temporal configurations contribute to the intricacy and profundity of spoken communication. To comprehend speech, signal processing analyses and extracts voice characteristics. The processing of these signals yields the spectral characteristics, including intensity, intonation, and content. These attributes exhibit the frequency, intensity, and frequency components of the speaker. The interpretation of speech signals is complicated by the modifications that cadence, pauses, and temporal variations introduce. Emotion information is conveyed through voice signals in voice emotion recognition[2]. Through variations in pitch, intonation, tempo, and other acoustic components, speech conveys emotion. Therefore, the analysis of speech signals is critical to comprehend these emotional indicators and to identify and categorize emotional states expressed through spoken language. By extracting features that capture these emotional shifts, models capable of identifying and classifying speech emotions can be developed. Within the domain of SER, these signals convey emotional indicators that may be identified and classified utilizing machine learning and signal processing methodologies. This allows for the identification of underlying emotions that are communicated through spoken words. Since emotions are frequently reflected in variations in pitch, tone, strength, and other acoustic aspects of speech signals, these characteristics are essential for the study of emotions[3].

Convolutional Neural Network (CNN): The CNN is a type of deep neural network that is utilized for the deep learning methodology[4]. Fig. 1 shows the working of a CNN model for classification tasks.

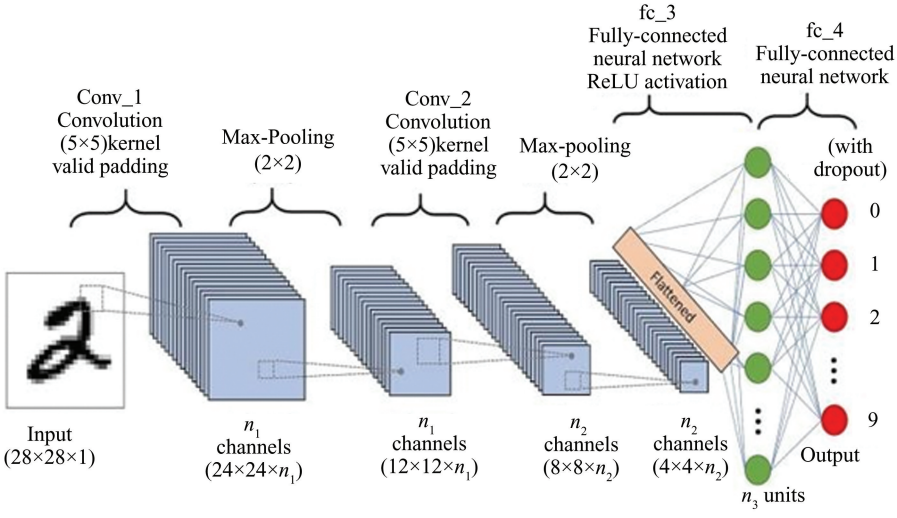

Fig.1 Convolutional neural network

The evaluation of speech signals, graphics, and signals are the primary applications of this tool. To identify patterns in the information that they were receiving, whether it was photos, sounds, sights, or position data, CNN attempted to function in a manner that was analogous to the way the human nervous system operates. In most cases, the structure of a CNN is composed of multiple layers that are stacked one on top of the other. Every layer has its unique approach to managing the data. When there are numerous layers in a system, the information gained from the layer that comes before it is utilized to process the information gained from the layer that comes before it. Naturally, each CNN can have more than one layer, and each layer can have its own set of parameters. Both of these possibilities are possible. To improve CNN's functionality, it is necessary to arrange all of these layers in the appropriate sequence and to configure their features appropriately[5]. The remaining parts of the paper are put together in the following manner: In Section 1, the related work is presented, and in Section 2, the methodology for the datasets, features extracted, and model employed are discussed. The comparative results and general comments are presented in Section 3. The paper concludes with Section 4.

1 Related Works

Krishna et al.[6] employed Support Vector Machine (SVM) and Multi-Layer Perception (MLP) classifiers along with MFCC, MEL, chroma, and Tonnetz audio features for emotion recognition, achieving an accuracy of 86.5%. Khalil[7], provides an overview of deep learning techniques applied in SER literature without focusing on a specific model. Anusha et al.[8] utilized classifiers trained on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset with features such as MFCC, Mel Spectrogram, and chroma for SER. While Wani[9] conducted a comprehensive review of SER systems and methodologies without a specific focus on a singular model. Arun[10], explored diverse machine learning models for emotion recognition in speech, utilizing various feature sets across Indian languages, aiming to identify optimal model-feature combinations for detecting emotions, including sarcasm. In the presented work[11], various machine learning models including SVM, Long Short-Term Memory network (LSTM), random forests, and CNNs were employed for emotion classification in speech signals, with the 2D CNN model achieving the highest accuracy of around 70% on the testing dataset. In Ref. [12], Mel-Frequency Cepstral Coefficients (MFCC) alongside pitch and Short-Term Energy (STE) features, utilizing SVM for emotion classification in North American English speech datasets. Kumar et al.[13] employed deep learning techniques for speech emotion recognition based on feature extraction and model creation. Ref. [14] implemented a deep learning-based system for emotion detection in speech signals, achieving an efficiency rate of 81.82%. Ref.[15] presented an emotion detection system for speech signals, validated with a dataset of 250 utterances from two Chinese female speakers. Ref.[16] introduced a Deep Neural Network (DNN) architecture for SER achieving a 96.97% accuracy on the Berlin Database of Emotional Speech (3 class subset). Cherif et al.[17] employed machine learning-based models, CNNs, LSTM, and BLSTM for Speech Emotion Recognition in the Algerian dialect, achieving a top accuracy of 93.34% with the LSTM-CNN model on their collected dataset. Yoon et al.[18] introduced a novel deep dual recurrent encoder model that combines text and audio signals for SER, outperforming prior methods with accuracies between 68.8% to 71.8% on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. Ramdinmawii[19], explored emotion recognition in speech signals using signal processing methods, analyzing four basic emotions (anger, happy, fear, neutral) by extracting features like F0, formants, dominant frequencies, ZCR, and signal energy. The study cross-validates findings across German and Telugu Emotion databases, revealing distinct differences between emotions, particularly in high-arousal states, providing insights for diverse applications. Ref. [20] introduced an end-to-end SER system employing multi-level acoustic data and a unique co-attention module, achieving competitive results on the IEMOCAP dataset. The model leveraged MFCC, spectrogram, and high-level acoustic data extracted through CNN, BiL-STM, and wav2vec2, respectively, fused using a proposed co-attention mechanism. In Ref. [21], Mountzouris used six deep learning networks: Deep Belief Network (DBN), DNN, LSTM, LSTM with Attention Mechanism (LSTM-ATN), CNN, and CNN with Attention Mechanism (CNN-ATN). Mountzouris trained and evaluated on the Surrey Audio-Visual Expressed Emotion (SAVEE) and RAVDESS databases, the models incorporated techniques like dropout and batch normalization for improved generalization and faster training. Results indicated that models with attention mechanisms outperformed others, with CNN-ATN achieving the highest accuracy of 74% for SAVEE and 77% for RAVDESS, surpassing existing state-of-the-art systems for these datasets.

2 Methodology

2.1 Dataset

Fig. 2 shows count versus emotions for the used dataset. For training and evaluating the SER model, the project makes use of several diverse datasets. These datasets provide a large variety of emotional speech examples, which provides a bank of audio samples that is both rich and varied and can be used for analysis and categorization.

The RAVDESS is the first dataset that is utilized. The audio recordings in this compilation feature a variety of performers articulating premeditated words that symbolize an extensive spectrum of emotions.

Each audio sample in RAVDESS is labeled with an emotion representing one of eight distinct sentiments. The vocal samples are annotated with labels that function as symbolic representations of these emotions. This class of sentiments comprises, among others, feelings of tranquility, joy, sorrow, anger, fear, revulsion, and awe. An exhaustive dataset about emotion recognition is presented, encompassing variables such as utterances and emotional intensity that are encoded within the filenames.

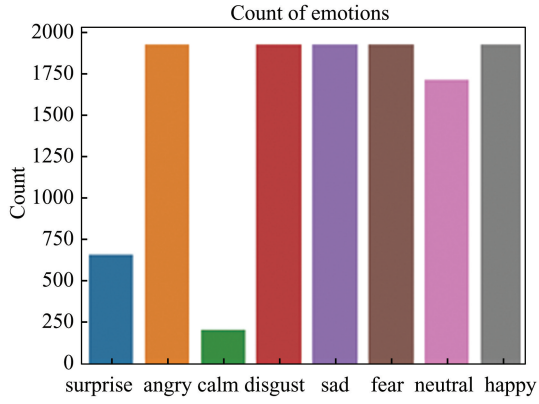

Fig. 2 Count of emotions

The study also makes use of the Crowd-sourced Emotional Multimodal Actors Dataset, also known as Crema-D. The audio recordings in this dataset depict a variety of emotional states through the use of expressive facial features; these states include melancholy, rage, contempt, dread, happiness, and neutrality. These expressions embody particular emotional states. Several instances of emotive speech elicited by the actors' performances are included in the dataset. Furthermore, this contributes an additional level of genuineness and inherent diversity to the emotional expressions. Additionally, a variety of emotional states are represented in the audio samples accessible via the SAVEE collection. These emotions include surprise, anger, disgust, fear, neutrality, and sorrow, among others. SAVEE is composed of spoken word recordings from extracted audio samples, where the listener can be access to an extensive array of emotions expressed through vocalizations. Audio recordings of emotional states constitute the final component of the Toronto Emotional Speech Set (Tess) collection. The majority of these expressions are comprised of surprise and other emotional states. Tess makes multiple contributions to the overall dataset pool, including the inclusion of unexpected instances and a variety of emotional expressions. This has led to a greater extent of the datasets. The amalgamation of these datasets yields an extensive and varied compilation of instances of impassioned speech. This compilation encompasses an extensive spectrum of affective states that are capable of being articulated orally. The versatility and reliability of the SER model, which underwent training using this compilation of datasets, are due to the genuineness of the recordings and the extensive array of emotional expressions employed. The model's capacity to extrapolate its findings is impacted by both factors.

To evaluate the reliability, the proposed work validates the on-Berlin Database of Emotional Speech (EMO-BD)[22] and IEMOCAP[23] datasets. The EMO－BD dataset is consisted of 535 emotional speech files. The dataset, which includes anxiety, happiness, neutrality, disgust, sadness, boredom, and anger, was recorded by five male and five female professional speakers, and is widely used for SER purposes by researchers. The IEMOCAP dataset, consisted of five sessions with male and female speakers, is used to analyze emotions. The dataset aggregates excited utterances into happy categories, considering four distinct emotion classes: angry, happy, neutral, and sad. The results show a distribution of 1103 angry utterances, 1636 happy utterances, 1708 neutral utterances, and 1084 sad utterances.

2.2 Features Extracted

To obtain crucial information for the emotion detection process, a substantial collection of features is extracted from the audio signals. To analyze feature importance in emotion recognition, employ techniques like permutation importance, SHAP values, and partial dependence plots. Permutation importance assesses each feature's impact by shuffling values and observing performance changes. SHAP (SHapley Additive exPlanations) values offer insights into individual feature contributions to predictions. Partial dependence plots visualize feature-emotion relationships, additionally, conducts statistical tests for feature significance. Rank features based on these analyses to identify the most influential ones, and uses findings to guide future feature engineering efforts, focusing on enhancing crucial features or exploring new representations. Validation through cross-validation ensures robustness. Document and report results clearly to facilitate understanding and guide further research. This systematic approach provides

insights into feature importance, aiding in optimizing emotion recognition models. The purpose of these attributes is to encapsulate the distinctive qualities of speech signals that communicate nuanced emotional information. They comprise a variety of the constituent elements of the audio data.

The following are the most significant characteristics that were extracted.

2.2.1 Zero Crossing Rate (ZCR)[24]

The rate at which the audio signal's sign changes is determined by this function, and this information can be used to understand the waveform's temporal fluctuations and transitions.

Sign changes indicate rapid shifts in the waveform, which can correlate with abrupt changes in emotional expression such as sudden outbursts, transitions between emotions, or changes in vocal intensity. For example, a higher ZCR may indicate a more dynamic and expressive vocal delivery, which could be associated with emotions like excitement or agitation.

2.2.2 Chroma_stft

Extracted from the Short-Time Fourier Transform (STFT), this feature captures the spectral content of the audio signal in different musical pitch classes, providing information about tonal qualities and musical content.

Emotions often manifest through changes in tonal qualities, such as pitch variations or musical content. Chroma_stft helps in capturing these musical features, providing insight into the melodic aspects of vocal expression. For instance, shifts in pitch or melody can convey emotions like joy, sadness, or tension.

2.2.3 Mel Frequency Cepstral Coefficients (MFCCs)[25]

MFCCs are a depiction of the audio signal's spectral characteristics that focus primarily on the frequency ranges that are perceptible to humans. Like how the human hearing system responds to outside noises, it can detect important frequency components.

MFCCs are particularly effective in capturing the nuanced variations in vocal timbre and texture that are indicative of different emotional states. They help in capturing the subtle differences in vocal tone, resonance, and articulation that accompany various emotions. For example, changes in the distribution of MFCCs may reflect variations in vocal tension, which could correspond to emotions like anger or fear.

2.2.4 Root Mean Square (RMS) value[26]

The audio signal's total amplitude or energy is represented by the Root-Mean-Square (RMS) value, which also provides information about the signal's loudness or intensity.

Emotions often manifest through variations in vocal intensity, ranging from whispers to loud exclamations. RMS helps in quantifying these variations in loudness, which can be indicative of emotional arousal or expressiveness. For instance, higher RMS values may indicate heightened emotional intensity, potentially corresponding to emotions like anger or excitement.

2.2.5 MelSpectogram[27]

This feature highlights the distribution of spectrum energy by providing a visual representation of the frequencies across time. It is derived from the Mel-frequency spectrogram. Figs. 3 and 4 focus on the spectrogram for audio with sad and happy emotions respectively, while Figs. 5 and 6 show the spectrogram for audio with fear and angry emotions respectively. In emotion recognition tasks using voice signals, each of the mentioned features plays a crucial role in capturing different aspects of the audio signal related to emotional expression.

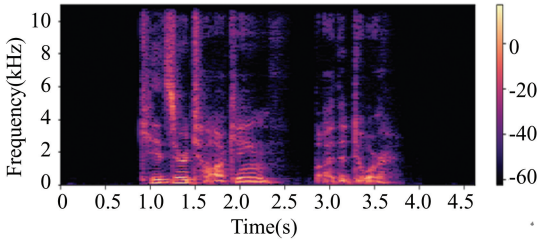

Fig. 3 Spectrogram for audio with sad emotion

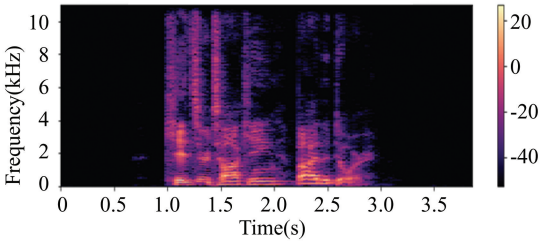

Fig. 4 Spectrogram for audio with happy emotion

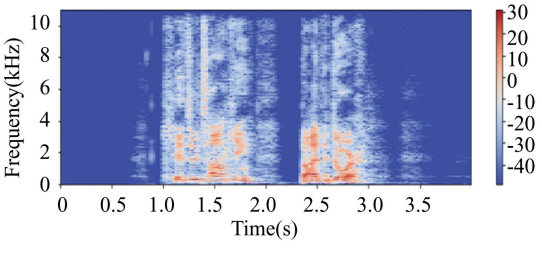

Fig. 5 Spectrogram for audio with fear emotion

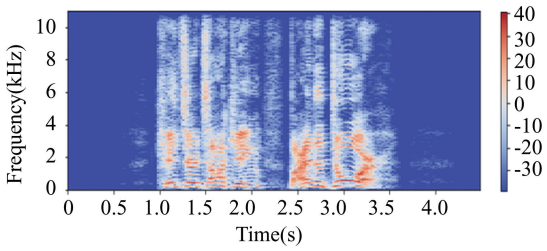

Fig. 6 Spectrogram for audio with angry emotion

MelSpectrogram offers a comprehensive overview of the spectral characteristics of the audio signal, capturing both temporal and frequency dynamics. It helps in identifying patterns and structures in the audio signal that correspond to different emotional expressions. For example, distinct patterns in the MelSpectrogram may correspond to specific emotional states, such as the presence of high-frequency energy bursts associated with fear or the presence of lower-frequency components associated with sadness. When these features are merged, the audio signals are represented in a diverse manner. This representation encompasses temporal and spectral elements, in addition to aspects related to intensity. They enable the model to identify and classify emotions more effectively since they have essential qualities that represent emotions that are expressed through speech. Furthermore, the dataset is improved through the application of augmentation techniques like stretching, pitch modulation, and noise injection. These methods strengthen the model's ability to handle variations and enhance its generalization capabilities.

2.3 Model Employed

The main model for emotion categorization based on auditory inputs is a CNN architecture. Throughout the process, this model serves as the main model. The goal of the CNN model is to process and automatically learn features from the input audio. This is achieved by the application of its hierarchical architecture, which makes it easier to find patterns and connections between the collected characteristics. From the audio spectrograms, the CNN architecture is composed of numerous layers that are designed to extract and abstract information that is pertinent to the problem at hand. All of these levels typically consist of the following layers.

1) Convolutional layers: By applying filters to the input spectrogram, these layers can identify a wide variety of characteristics and patterns included within the audio signals.

2) Max-pooling layers: After the convolutional layers, the max-pooling layers down sample the learned features, thereby lowering the dimensionality and concentrating on the most significant information.

3) Dropout layers: There are dropout layers that are incorporated to prevent overfitting. These layers randomly deactivate a portion of the neurons during training, which encourages the model to generalize more effectively.

4) Fully connected layers: The retrieved features and patterns are included in these layers, which are then used for the final classification into several categories of emotions. The CNN model architecture that was utilized for the project most likely consists of numerous convolutional blocks, each of which includes convolutional, pooling, and potentially dropout layers, followed by dense layers for classification. When doing multi-class classification, it is a usual practice to use activation functions such as ReLU (Rectified Linear Unit) in the convolutional layers and softmax in the final output layer.

The proposed CNN model consists of a sequential stack of several layers with used parameters as follows.

1) Input layer: The input shape is determined by the 'input_shape' parameter, which is '(x_train.shape[1], 1)'. It implies that the input data consists of sequences with a single feature.

2) Convolutional layers: There are four convolutional layers added sequentially. Each convolutional layer has a different number of filters and kernel sizes. The first convolutional layer has 256 filters with a kernel size of 5, followed by the subsequent layers with 256, 128, and 64 filters respectively. The activation function used in each convolutional layer is ReLU, which introduces non-linearity into the model and helps in capturing complex patterns in the data. Padding is set to 'same', which ensures that the output size remains the same as the input size.

3) Max pooling layers: After each convolutional layer, there is a max-pooling layer. Max-pooling is used for downsampling the feature maps, reducing computational complexity, and helping the model to focus on the most important features. Each max-pooling layer has a pool size of 5 and a stride of

2, which means it takes the maximum value within a window of size 5 and moves by 2 steps at a time. Padding is set to 'same' to ensure that the output size remains consistent.

4) Dropout layers: Two dropout layers are added to prevent overfitting. Dropout randomly sets a fraction of input units to zero during training, which helps in reducing overfitting by preventing the network from relying too much on specific activations.

5) Flatten layer: This layer flattens the output of the previous layer into a one-dimensional array, which is required before passing it to the fully connected layers.

6) Dense layers: There are two fully connected dense layers. The first dense layer has 32 units and utilizes the ReLU activation function. The second dense layer has 8 units with a softmax activation function, which is used for multi-class classification tasks. Softmax normalizes the output into a probability distribution over the 8 classes.

7) Compilation: The model is compiled with the Adam optimizer, categorical cross-entropy loss function (suitable for multi-class classification), and accuracy as the evaluation metric.

8) Callbacks: A ReduceLROnPlateau callback is used to reduce the learning rate when the training loss plateaus. It monitors the loss, and if the loss does not decrease for a certain number of epochs (patience), it reduces the learning rate by a factor specified (0.4 in this case) until it reaches the minimum specified learning rate (0.0000001).

Table 1 shows the different layers used in the model. Overall, this model architecture leverages convolutional layers for feature extraction from sequential data, max-pooling layers for down-sampling, dropout layers for regularization, and fully connected layers for classification. The ReduceLROnPlateau callback helps in optimizing the learning process by adjusting the learning rate dynamically during training. Additionally, optimization strategies such as ReduceLROnPlateau, which allows for the adjustment of learning rates and categorical cross-entropy loss functions, in conjunction with the Adam optimizer, are frequently utilized to effectively train the model and optimize its performance in recognizing emotions based on the audio features.

Table 1 Architecture of used CNN model

Layer (type)	Output shape	Parameter
conv1d_28 (Conv1D)	(None, 162, 256)	1536
max_pooling1d_28 (MaxPooling)	(None, 81, 256)	0
conv1d_29 (Conv1D)	(None, 81, 256)	327936
max_pooling1d_29 (MaxPooling)	(None, 41, 256)	0
conv1d_30 (Conv1D)	(None, 41, 128)	163968
max_pooling1d_30 (MaxPooling)	(None, 21, 128)	0
dropout_13 (Dropout)	(None, 21, 128)	0
conv1d_31 (Conv1D)	(None, 21, 64)	41024
max_pooling1d_31 (MaxPooling)	(None, 11, 64)	0
flatten_7 (Flatten)	(None, 704)	0
dense_13 (Dense)	(None, 32)	22560
dropout_14 (Dropout)	(None, 32)	0
dense_14 (Dense)	(None, 8)	264

3 Results and Discussion

Fig. 7 shows a graph for training and testing loss. Several encouraging findings were obtained by the SER model in terms of identifying and categorizing emotions based on speech signals. After being evaluated, the model demonstrated an accuracy of roughly 94.57% across the board when applied to the demonstration dataset. There were notable variations in the categorization performance amongst the different emotional categories. This difference was noteworthy enough to mention. The results showed that some emotions were more accurately classified than others, with rage and surprise being two examples of these emotions. Performance is constantly fluctuating, which is in line with the intrinsic complexity of the task of differentiating between distinct emotional expressions seen in speech signals.

The model showed a higher degree of accuracy in differentiating between surprise and anger. This could be explained by the distinct aural qualities associated with different emotional states. Pitch, tone, and intensity changes are a few characteristics that

characterize this occurrence. Nevertheless, the task of differentiating discrete emotions was complicated by factors such as neutrality and slight variations in facial expressions of emotion. Because some emotions may share more auditory characteristics than others, it may be more difficult to distinguish between them solely by listening to their sounds.

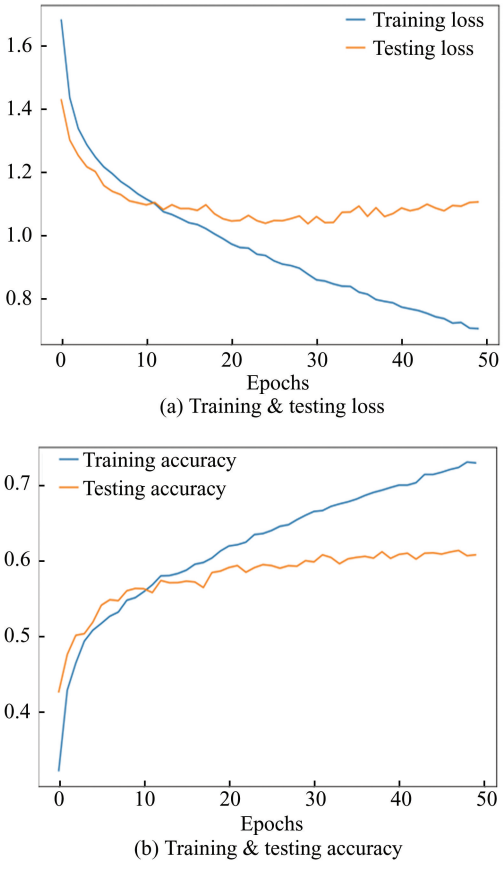

Fig. 7 Training and testing results for number of epochs with the loss and accuracy representations

The confusion matrix between predicted and actual labels is illustrated in Fig. 8. The model's ability to accurately discern emotions from a variety of speech instances demonstrated its generalizability. This outcome was attained through the attainment of extreme precision. The fact that the accuracy remains only moderately accurate suggests that there are possible avenues for enhancement. This implies that to enhance the model's ability to differentiate between different emotional states via application, it might be imperative to refine feature extraction methods, investigate more complex network architectures, or consider novel augmentation techniques.

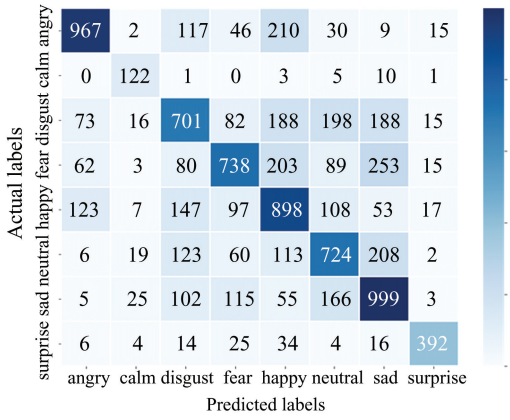

Fig. 8 Confusion matrix of the proposed work

In addition, the incorporation of additional modalities or contextual information could potentially improve the model's capability of discerning intricate emotional signals concealed within speech signals. This is due to the model's failure to consider the context of the speech signals. Further research initiatives are warranted in the domain of speech emotion recognition, which has the potential to yield substantial advancements and enhancements. In addition to the CNNs utilized in this research, an opportunity arises to investigate more complex neural network topologies such as Transformer-based models or Recurrent Neural Networks, which can improve the model's comprehension of the affective indicators inherent in speech signals. By investigating more sophisticated techniques for feature extraction and process selection, it is possible to further improve the model's ability to identify subtle fluctuations in emotional state. By integrating various modalities, such as facial expressions or physiological data, the model's comprehension of emotions could potentially be enhanced through the implementation of multimodal fusion techniques. This would be a positive step to take. Combining datasets with a wide range of ethnicities, languages, and cultural backgrounds could help achieve this. This would increase the model's inclusiveness and applicability. To enhance reproducibility and applicability provide specific details on hyperparameter optimization in the study, and outline the hyperparameters tuned, such as learning rate, batch size, and regularization strength, along with their ranges or distributions. The optimization method used, like grid search, random search, or Bayesian optimization. The impact of hyperparameters on model performance metrics, such

as accuracy or F1 score is shown in Table 2, include any trade - offs or interactions observed between hyperparameters. By the hyperparameter tuning process, it can replicate and adapt the methodology effectively, fostering transparency and facilitating further advancements in the field of emotion recognition. Notwithstanding the advances that have been achieved, some challenges remain to be solved. A few of these challenges are the dataset's constraints, the subjectivity associated with emotion classification, and the challenge of differentiating between emotions that are closely related to one another. Consequently, this will aid in the advancement of SER, thereby unlocking its potential in numerous field applications. This objective can be addressed by tackling these problems by applying improved techniques and gaining a more profound understanding of the emotional cues contained in speech signals.

The study compares to the suggested approach, which employs deep learning with CNN, as shown in Table 2. This analysis evaluates the proposed technique in several previous research attempts. The evaluation metrics employed are accuracy, F1 score, and processing time. The proposed method, utilizing CNNs, obtains an accuracy of 94.57% and an impressive F1 score of 0.947, with a processing time of 12 ms. By contrast, Krishna et al.[6] employed SVM and MLP with diverse features. Their accuracy reached 86.50% and their F1 score was 0.850, all accomplished within a processing time of 20 ms. Mittal et al.[11] combined SVM, LSTM, Random Forests, and CNNs, resulting in a processing time of 25 ms, an accuracy of 70.00 %, and an F1 test score of 0.680. Babu et al.[14] employed Librosa for deep learning and achieved an accuracy of 81.82%, an F1 score of 0.80, and a processing time of 15 ms. Cherif et al.[17] employed machine learning models, CNNs, and LSTM and achieved a high accuracy of 93.34%, an F1 score of 0.920, and a processing time of 22 ms. Yoon et al.[18] utilized a Deep dual recurrent encoder model, achieving a processing time of 30 ms, an accuracy of 68.80%, and an F1 score of 0.67. Mountzouris et al.[21] employed DBN, DNN, LSTM, LSTM-ATN, CNN, CNN-ATN and CNN-ATN achieved a high accuracy of 74.00% for SAVEE and 77.00% for RAVDESS, an F1 score of 0.726, and a processing time of 28 ms. The table offers a comprehensive summary, depicting the trade-offs between processing time, F1 score, and accuracy for each methodology. The suggested method demonstrates remarkable performance in terms of F1 score and processing time, rendering it a feasible choice in some situations.

Table 2 Comparative results of proposed methodology with different state-of-the-art networks

Reference	Methodology	Accuracy(%)	F1 score	Processing time(ms)
Proposed method	Deep learning with CNN	94.57	0.947	12
Krishna et al.[6]	SVM and MLP with various features	86.50	0.850	20
Mittal et al.[11]	SVM, LSTM, Random Forests, CNNs	70.00	0.680	25
Babu et al.[14]	Deep learning with Librosa	81.82	0.800	15
Cherif et al.[17]	Machine learning models, CNNs, LSTM	93.34	0.920	22
Yoon et al.[18]	Deep dual recurrent encoder model	68.80	0.670	30
Mountzouris et al.[21]	CNN with the addition of an attention mechanism (CNN-ATN)	74.00% for SAVEE 77.00% for RAVDESS	0.726	28

The proposed research presents a comprehensive investigation into SER utilizing deep learning techniques across multiple datasets. Our study primarily focuses on the utilization of CNNs for sound feature extraction, augmented by various techniques such as stretching, pitch manipulation, and noise injection to enhance data quality. We aim to provide an in-depth analysis of these methods and their impact on SER performance which is shown in Table 3. In the proposed research, a CNN-based feature extraction method with data augmentation techniques achieved impressive performance with 95.67% accuracy on the EMO-BD dataset and 84.21% on the IEMOCAP dataset. The method proposed by Ref. [28] combined SSL models and spectral features through MoE, achieving a weighted accuracy of 73.91% and an unweighted accuracy of 72.29%, addressing the domain shift problem in SER. Similarly, Ref. [29] utilized multi-resolution variational mode decomposition, achieving 90.51% accuracy, while

facing challenges like dataset dependency and potential overfitting. Additionally, Ref. [30] introduced a fusion of spectral and temporal features using CNNs and a convolution layer-based transformer, obtaining 94.20% accuracy on EMO-BD and 81.10% on IEMOCAP datasets, highlighting computational complexity as a challenge. However, the proposed research stands out due to its superior performance, achieving notably higher accuracy rates on both datasets, surpassing the limitations encountered in the other studies, and demonstrating the effectiveness of the CNN-based feature extraction method with data augmentation techniques in enhancing SER.

Table 3 Comparison on different methods with proposed method on EMO-BD and IEMOCAP dataset

Aspect	Proposed research	Hyeon et al.[28]	Mishra et al.[29]	Saleem et al.[30]
Methodology	CNN-based feature extractionwith data augmentation techniques	Combination of SSL models and spectral features using MoE	Multi-resolution variational mode decomposition method	Fusion of spectral and temporal features using CNNs and a convolution layer-based transformer
Dataset	EMO-BD, IEMOCAP dataset	IEMOCAP dataset	EMO-DB datasets	EMO-BD, IEMOCAP dataset
Achieved performance	95.67%, 84.21 % on EMO-BD and IEMOCAP datasets respectively.	Weighted Accuracy (WA) of 73. 91% and an Unweighted Accuracy (UA) of 72.29%	90.51% accuracy	94. 20% accuracy on EMO-BD, 81. 10% accuracy on IEMOCAP
Features	CNN-based feature extraction, data quality enhancement techniques	Self-supervised learning models, spectral features, MoE	Multi-resolution variational mode decomposition, MRVMMFCC, MRVMAE, MRVMPE	Parallel CNNs, convolution layer-based transformer, attention gated recurrent unit
Limitations/ Challenges		Domain shift problem of SSL models, potential computational complexity	Dependency on dataset characteristics, potential overfitting	Complexity and computational cost

4 Conclusions

Convolutional neural networks have been applied to voice emotion recognition with great success, allowing for the identification and categorization of emotions from voice input. The emotion recognition technology has enabled these advancements. It was easier to develop a strong model that could identify a range of emotional states based on aural inputs by using many datasets that covered a wide range of emotional expressions. The fact that the datasets included a variety of emotional expressions allowed for this. Among the features that were taken out of the audio signals were MFCCs, RMS values, Zero Crossing Rate, Chromist, and Mel Spectrograms. These attributes provided a comprehensive elucidation of the auditory signals. These attributes resulted in the accumulation of noteworthy characteristics that are linked to an assortment of emotions. Pitch modulation, stretching, and noise injection were among the augmentation techniques implemented to increase the model's capacity to generalize across a broad spectrum of affective expressions. The diversity of the dataset was expanded as a result of the implementation of these strategies. Despite attaining a respectable accuracy of around 94. 57% on the test dataset, the model demonstrated substantial variability in its performance across various emotional categories. The proposed work attained an accuracy of 95. 67% and 84.21% on the validation datasets of the EMO-BD and IEMOCAP datasets, respectively. Even though it achieved such a high degree of accuracy. It is particularly effective at reliably classifying certain emotions to a greater extent than others. This exemplifies the necessity for further refinement and expansion of the model's discriminatory capabilities, particularly about discerning dim or closely associated emotional states. To augment the precision of emotion identification, prospective domains of research might encompass the examination of improved techniques for feature extraction, alternative architectures for networks, or multimodal methodologies that integrate

contextual comprehension. Furthermore, the integration of more extensive and varied datasets, along with the enhancement of augmentation techniques, may potentially augment the model's capacity to understand and classify intricate emotional expressions conveyed in speech signals. Overall, the SER model represents a positive advancement in the direction of automated emotion recognition from speech. To more accurately capture the intricacies of human emotional indicators conveyed via audio signals, ongoing advancements and improvements are required.

References

[1] Buza O, Toderean G, Nica A, et al. Voice signal processing for speech synthesis. IEEE International Conference on Automation, Quality and Testing, Robotics. Piscataway: IEEE. 2006. 360-364. DOI: 10.1109/AQTR.2006.254660.

[2] Islam A R, Tarique M, Abdel-Raheem E. A survey on signal processing based pathological voice detection techniques. IEEE Access, 2020, 8: 66749-66776. DOI: 10.1109/ACCESS.2020.2985280.

[3] Dhananjaya N, Yegnanarayana B. Voiced/Nonvoiced detection based on robustness of voiced epochs. IEEE Signal Processing Letters, 2010, 17(3): 273-276. DOI: 10.1109/LSP.2009.2038507.

[4] Grossi E, Buscema M. Introduction to artificial neural networks. European Journal of Gastroenterol Hepatology, 2007, 19(12): 1046-1054. DOI: 10.1097/meg.0b013e3282f198a0.

[5] O'Shea K, Nash R. An introduction to convolutional neural networks. 2015, arXiv:1511.08458. DOI: 10.48550/arXiv.1511.08458.

[6] Krishna K V, Sainath N, Posonia A M. Speech emotion recognition using machine learning. 6th International Conference on Computing Methodologies and Communication (ICCMC). Piscataway: IEEE, 2022. 1014-1018. DOI: 10.1109/ICCMC53470.2022.9753976.

[7] Khalil R A, Jones E, Babar M I, et al. Speech emotion recognition using deep learning techniques: A review. IEEE Access, 2019, 7: 117327-117345. DOI: 10.1109/ACCESS.2019.2936124.

[8] Anusha R, Subhashini P, Jyothi D, et al. Speech emotion recognition using machine learning. 5th International Conference on Trends in Electronics and Informatics (ICOEI). Piscataway: IEEE. 2021. 1608-1612. DOI: 10.1109/ICOEI51242.2021.9453028.

[9] Wani T M, Gunawan T S, Qadri S A A, et al. A comprehensive review of speech emotion recognition systems. IEEE Access, 2021, 9: 47795-47814. DOI: 10.1109/ACCESS.2021.3068045.

[10] Arun A, Rallabhandi I, Hebbar S, et al. Emotion recognition in speech using machine learning techniques. 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). Piscataway: IEEE, 2021. 1-7. DOI: 10.1109/ICCCNT51525.2021.9580028.

[11] Mittal R, Vart S, Shokeen P. Speech emotion recognition. 2nd International Conference on Intelligent Technologies (CONIT). Piscataway: IEEE, 2022. 1-6. DOI: 10.1109/CONIT55038.2022.9848265.

[12] Deshmukh G, Gaonkar A, Golwalkar G, et al. Speech based emotion recognition using machine learning. 3rd International Conference on Computing Methodologies and Communication (ICCMC). Piscataway: IEEE, 2019.812-817. DOI: 10.1109/ICCMC.2019.8819858.

[13] Kumar A, Kumar V, Rajakumar P. Speech emotion recognition using machine learning. 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM). Piscataway: IEEE, 2023. 1-6. DOI: 10.1109/ICIPTM57143.2023.10118251.

[14] Babu P A, Siva Nagaraju V, Vallabhuni R R. Speech emotion recognition system with Librosa. 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT). Piscataway: IEEE, 2021. 421-424. DOI: 10.1109/CSNT51715.2021.9509714.

[15] Luoh L, Su L Z, Hsu C F. Speech signal processing-based emotion recognition. 2010 International Conference on System Science and Engineering. Piscataway: IEEE, 2010. 487-490. DOI: 10.1109/ICSSE.2010.5551812.

[16] Harár P, Burget R, Dutta M K. Speech emotion recognition with deep learning. 4th International Conference on Signal Processing and Integrated Networks (SPIN). Piscataway: IEEE, 2017. 137-140. DOI: 10.1109/SPIN.2017.8049931.

[17] Cherif R Y, Moussaoui A, Frahta N, et al. Effective speech emotion recognition using deep learning approaches for Algerian dialect. 2021 International Conference of Women in Data Science at Taif University (WiDSTaif). Piscataway: IEEE, 2021. 1-6. DOI: 10.1109/WiDSTaif52235.2021.9430224.

[18] Yoon S, Byun S, Jung K. Multimodal speech emotion recognition using audio and text. IEEE Spoken Language Technology Workshop (SLT). Piscataway: IEEE, 2018. 112-118. DOI: 10.1109/SLT.2018.8639583.

[19] Ramdinmawii E, Mohanta A, Mittal V K. Emotion recognition from speech signal. TENCON 2017-2017 IEEE Region 10 Conference. Piscataway: IEEE, 2017. 1562-1567. DOI: 10.1109/TENCON.2017.8228105.

[20] Zou H, Si Y, Chen C, et al. Speech emotion recognition with co-attention based multi-level acoustic information. ICASSP 2022-2022 IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2022. 7367-7371. DOI: 10.1109/ICASSP43922.2022.9747095.

[21] Mountzouris K, Perikos I, Hatzilygeroudis I. Speech emotion recognition using convolutional neural networks with attention mechanism. Electronics, 2023, 12(20): 4376. DOI: 10.3390/electronics12204376.

[22] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech. Interspeech, 2005, 5: 1517-1520. DOI: 10.21437/Interspeech.2005-446.

[23] Busso C, Bulut M, Lee C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008, 42: 335-359. DOI: 10.1007/s10579-008-9076-6.

[24] Kathirvel P, Sabarimalai Manikandan M, Senthilkumar S, et al. Noise robust zerocrossing rate computation for audio signal classification. 3rd International Conference on Trendz in Information Sciences & Computing (TISC2011). Piscataway: IEEE, 2011. 65-69. DOI: 10.1109/TISC.2011.6169086.

[25] Abdul Z K, Al-Talabani A K. Mel frequency cepstral coefficient and its applications: A review. IEEE Access, 2022, 10: 122136-122158. 2022, DOI: 10.1109/ACCESS.2022.3223444.

[26] Serov A N, Shatokhin A A, Serov N A. Application of the signal samples approximation for accurate RMS measurement. 44th International Convention on Information, Communication and Electronic Technology (MIPRO). Piscataway: IEEE, 2021. 147-153. DOI: 10.23919/MIPRO52101.2021.9597075.

[27] Meghanani A, Anoop C S, Ramakrishnan A G. An exploration of log-mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech. 2021 IEEE Spoken Language Technology Workshop (SLT). Piscataway: IEEE, 2021. 670-677. DOI: 10.1109/SLT48900.2021.9383491.

[28] Hyeon J, Oh Y H, Lee Y J, et al. Improving speech emotion recognition by fusing self-supervised learning and spectral features via mixture of experts. Data and Knowledge Engineering, 2023, 150: 102262. DOI: 10.1016/j.datak.2023.102262.

[29] Mishra S P, Warule P, Deb S. Improvement of emotion classification performance using multi-resolution variational mode decomposition method. Biomedical Signal Processing and Control, 2024, 89: 105708. DOI: 10.1016/j.bspc.2023.105708.

[30] Saleem N, Gao J, Irfan R, et al. DeepCNN: Spectro-temporal feature representation for speech emotion recognition. CAAI Transactions on Intelligence Technology, 2023, 8(2): 401-417. DOI: 10.1049/cit2.12233.