SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V13I5P114

Volume 13 Issue 5, 172-182, May 2026

Original Article

Advanced Signal Processing and Deep Learning-Based Speech Emotion Recognition in the Bodo Language

Rupali Khaklary[1], Nabankur Pathak[2]

[1,2]*Krishna Kanta Handiqui State Open University, Guwahati, Assam, India.*

[1]*Corresponding Author : khaklaryrupali9@gmail.com*

Received: 22 February 2026 Revised: 21 March 2026 Accepted: 20 April 2026 Published: 30 May 2026

Abstract *- In the current generation of communication systems, it is essential for accurate emotion recognition across linguistically diverse settings. Extensive research has addressed emotion recognition from speech (SER) in various languages, whereas investigations involving the Bodo language are still scarce. This work addresses a Bodo-specific SER framework, combining signal processing techniques with deep learning models. The study introduces an innovative audio data collection method, specifically tailored for Bodo emotional speech, which has not been previously explored. To represent the acoustic characteristics of the Bodo speech signals, MFCC, Mel-spectrogram, Chroma, Zero Crossing Rate, and Root Mean Square Energy are extracted and organized as input features for training the proposed model. Extraction is performed under two conditions using original data and augmented samples for comparative evaluation. The resulting feature sets train the proposed Convolutional Neural Network model (CNN), optimized through hyperparameter tuning. Performance is compared between augmented and non-augmented datasets. The proposed CNN-based model, combined with augmented data, demonstrates higher accuracy (81.71%) and robustness in emotion recognition. This work also provides a novel analysis of the unique spectral and prosodic characteristics of Bodo speech, offering fresh insights into its acoustic properties. The proposed approach achieves higher accuracy in Bodo speech emotion recognition and contributes to further research in this area.*

Keywords *- Bodo Language, Data augmentation, Deep Learning, Signal Processing, Speech Emotion Recognition.*

1. Introduction

The analysis of emotional information in speech signals, referred to as Speech Emotion Recognition (SER), has gained increasing attention due to its importance in developing more effective human–machine interaction systems. SER supports the systems that require effective interaction between users and machines, such as speech synthesis, customer service systems, e-learning platforms, call centres, forensic analysis, and medical diagnostics [1-4]. To communicate with a better understanding of each other with accurate emotional cues, it is essential to enable more natural interaction between users and intelligent systems. However, in many digital communication environments, the emotional intent of the speaker cannot be easily perceived, which motivates the development of automatic SER technologies.

Significant research has been carried out on several languages that enabled the integration of various real-world applications. Recent advancements in SER have been driven by deep learning, Transfer learning, and ensemble learning. Likewise, the Speech Former++ framework has been proposed for paralinguistic speech processing [5], incorporating temporally aware architectures such as Bi-directional Multi-scale Networks (TIM-Nets) to capture emotional variations across time. Additionally, the approach integrates spatial and temporal representations, combining parallel Convolutional architectures with a transformer encoder [6]. This method is inspired by emotional perception theories derived from brain science for implicit emotional attribute classification [7] and utilizes excitation features and Zero-Shot Learning (ZSL) for emotion detection [8]. Recent investigations suggest that CNN architectures are well-suited for modeling speech signals and improving emotion classification performance. Hybrid feature-based CNN models have been suggested to achieve more accurate classification of emotional state [9]. Furthermore, deep learning architectures such as CNN and CNN–LSTM have been employed for speech emotion recognition by incorporating diverse acoustic features, for example, MFCC, Zero Crossing Rate, Root Mean Square Energy, and pitch [10]. In Consistent with earlier studies, CNN-based models trained on spectrogram representations achieved strong emotion recognition results [11]. Irrespective of these developments, SER research for the Bodo language is limited due to its low-resource nature, with few studies addressing emotional speech analysis. The study in [12] is one of the earliest studies on Bodo SER, where a Gaussian Mixture Model (GMM)-based approach was proposed for voice

emotion recognition across multiple languages, including Bodo. Emotional characteristics of the speech signals are modeled using a combination of cepstral features, namely WPCC2, MFCC, tfWPCC2 derived through Teager Energy operation, and tfMFCC. As an early study in Bodo SER, these findings provide some initial insight into emotional characteristics present in Bodo speech. Sharma [13] applied HMM (Hidden Markov Model) for classifying emotions in Bodo speech using pitch-related features and MFCCs. Subsequently, research [14] suggested that a hybrid framework for generating emotional prosody in Bodo speech synthesis by combining rule-based and template-based methods to model pitch, duration, and intensity. Barnali [15] further developed an emotion speech synthesis for the Bodo language at the syllable, word, and sentence levels using HMM. Research on Bodo emotional speech remains at an early stage, where most approaches are centered on traditional statistical methods and constrained feature sets.

The overall workflow of the proposed system includes preprocessing of the recorded Bodo speech signals, computation of spectral and prosodic descriptors, and subsequent classification using a deep learning model. Among these stages, the extraction of acoustic features is critical for encoding information relevant to emotion recognition [16]. Nevertheless, it is not an easy task to determine a set of features that can be effectively used to differentiate between different emotional states [17]. Improved SER performance has been reported when spectral and prosodic representations are utilized together [18-23]. Although Sharma [24] suggested that tonal languages may have limited influence on prosodic features, the Bodo language is significantly influenced by the prosodic features of its regional linguistic structure. Therefore, analysing both spectral and prosodic features may provide valuable insights into emotional characteristics in Bodo speech. Nevertheless, deep learning methods of speech emotion recognition in the Bodo language are poorly studied. In the context of Bodo emotional speech analysis, earlier approaches have predominantly relied on GMM- and HMM-based frameworks, with comparatively fewer studies exploring deep neural models for automated feature extraction and emotion recognition. Thus, this paper will explore the issue of spectral and prosodic features together with a CNN-based model as an effective SER method in Bodo to enhance a higher classification accuracy as compared to existing methods.

This study introduces several novel contributions to SER in Bodo speech as outlined below:

1. The introduction of a novel emotion-labelled Bodo speech data collection method that allows the formation of more diverse and representative data to train models is presented.
2. The research investigates key spectral and prosodic actualizations of Bodo speech, analyzing their contributions to the performance of emotion classification.
3. An introduction of a deep learning–based CNN model is applied on speech emotion recognition to improve performance.
4. Application of data augmentation to achieve an improved and more robust emotion recognition performance with the integration of data augmentation strategies with CNN.

1.1. The Bodo Language

From a linguistic perspective, Bodo is recognized as a member of the Tibeto-Burman subgroup within the wider Sino-Tibetan family of languages [25]. It is among the 8th-scheduled Indian languages of the government of India. In northeastern India, especially in Assam, this language is predominantly spoken by the Bodo community. In addition to its core region, the language is also found in surrounding northeastern states, namely Tripura, Meghalaya, Nagaland, Arunachal Pradesh, and in some districts of West Bengal [26, 27]. Apart from this, Bodo-speaking populations can also be found in nearby countries such as Bangladesh, Nepal, and Bhutan, indicating their broader regional distribution.

2. Materials and Methods

2.1. Dataset Description

Being a low-resource language, Bodo does not have a publicly available and well-structured emotional speech corpus, which motivates the development of a dedicated dataset in this study. The proposed study used a self-built Bodo emotional speech corpus. The corpus was designed from an interview-based speech dataset. A questionnaire consisting of 15 emotionally rich Bodo sentences in text-independent format and of varying lengths was documented for data collection. For the purpose of modeling, six discrete emotion states are defined, namely neutral, happiness, sadness, anger, fear, and surprise, which form the target classes of the system. In this context, emotional scenarios were created to bring such emotions into play. Each speaker was asked to read out the specific sentence and imagine the emotional situation or scenario mentioned in the set of scripts of each of the six emotions. They were required to repeat each sentence twice for every emotion. The process of audio collection spanned six sessions per speaker, dedicating one session to each specific emotion. After each session, a 10-15-minute break was given. The age of the speaker is limited to 15-45 years, possessing a good speaking voice quality with at least a matriculation qualification. In this way, for each speaker, there are 180 audio speeches in total, i.e., 30 speeches for each emotion, repeating each sentence two times. So, the dataset contains a total of 3960 audio files in .wav format, with a total of 660 files for each emotion. The speech dataset was developed using recordings from 22 native speakers, including 10 males and 12 females, to ensure diversity in speaker representation. Speech samples were acquired using a condenser microphone (SF-666), keeping approximately 24 inches from each participant. Recording was carried out in stereo mode using

Audacity, an open-source audio acquisition tool, with a fixed sampling frequency of 44.1 kHz to maintain consistent signal quality. Recorded audio files are then pre-processed for the implementation process.

2.2. Features Selection for the Study

As discussed in [28-31], the effectiveness of any SER system is strongly influenced by the selection of appropriate acoustic representations. SER commonly relies on acoustic descriptors similar to those used in speech processing, such as fundamental frequency (pitch), signal energy, speaking rate, and spectral representations including Mel-frequency cepstral coefficients (MFCCs) [3, 32]. In this study, 3 spectral features: MFCC, chroma, and mel-spectrogram, and 2 prosodic features: ZCR and RMSE were considered.

2.3. Features Extraction Approach

Realizing a strong SER of an environment with a scarcity of data is not easy, especially when it comes to a language with limited data. To achieve significant performance, training a deep learning model typically necessitates a huge amount of data [33, 34]. Training a deep learning model, challenges with data scarcity and lack of generalization can hinder the training process [35]. Therefore, the training dataset was expanded using additive white noise and temporal stretching to improve model robustness under noisy conditions and enhance generalization to unseen samples. Feature extraction was performed in two approaches: first, without adding data augmentation, and second, with data augmentation on the training dataset. This work examines whether augmentation strategies enhance the classifier's ability to generalize, particularly by improving accuracy and robustness under varied conditions. The extracted feature sets are subsequently used as input into the proposed DL SER model. Figure 1 represents the block diagram of the model to be used for implementation.

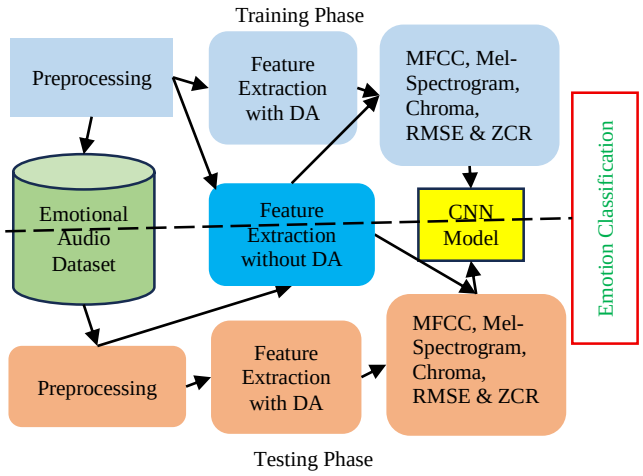


Fig. 1 Block Diagram of Speech Emotion Recognition

2.4. Features Extraction Methods

2.4.1. Spectral Features

MFCC

The flowchart in Figure 2 describes the MFCC extraction procedure, where the time-domain speech signal is first divided into overlapping windows and then converted into its corresponding spectral form through FFT-based analysis of each segment. Since human auditory perception is nonlinear at frequencies higher than 1000Hz, the mel-scale theory, which describes the sound characteristics, is considered. A triangular filter, the Mel filter bank, is applied to the spectrum using convolution between the spectrum and filter bank coefficients to get the Mel-spectrum. A Mel frequency, F, is obtained from a normal frequency f by the formula as in (1).

$$Mel(F) = 2595 * log_{10}(1 + \frac{f}{700}) \quad (1)$$

Then the logarithm of the mel-spectrums at each Mel frequency is calculated. MFCCs are derived through DCT as in (2)

$$Mn = \sum_{k=1}^{L}(logS_k)Cos[n(k - \frac{1}{2})\frac{\pi}{L}] \quad (2)$$

Where S_k denotes the energy corresponding to the output of the k^{th} triangular filter, with k ranging from 1 to L.

Chroma

Chroma features represent how spectral energy is distributed across 12 pitch classes. The filter bank stage aggregates spectral information by computing the mean log-magnitude of DFT coefficients within a predefined frequency group. As defined in (3), V_k is calculated by averaging values

$$V_k = \sum_{n \in S_k}\left(\frac{x_i(n)}{N_k}\right), k = 0,1,2, \dots 11 \quad (3)$$

Over the subset S_k, corresponding to the k^{th} filter bank region and N_k is the size of the subset, the number of frequency components involved in the averaging process [34]. Chroma is obtained by applying STFT to the signal, mapping the log frequency spectrogram onto the closest pitch class. Then, aggregating the log frequency magnitude spectrum within each pitch class across time, a single coefficient is determined.

Mel-Spectrogram

A mel-spectrogram represents the time–frequency structure of an audio signal using a perceptual frequency mapping, where frequency components are transformed into a scale that reflects human auditory sensitivity [36]. Time-frequency representation of the signal was obtained using STFT, which allows spectral variations to be analyzed over short temporal segments. Following this,

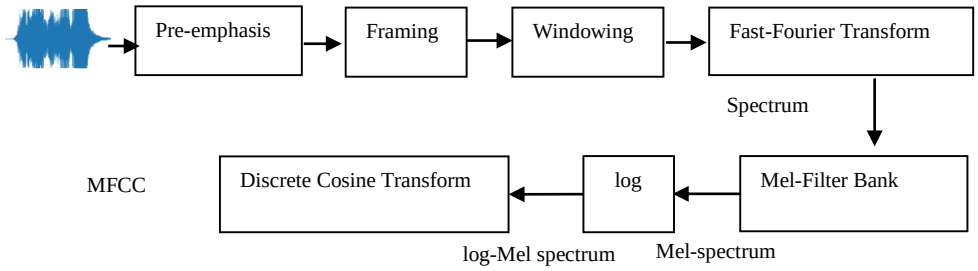


Fig. 2 Flowchart of MFCC Feature Extraction

The resulting frequency components were subjected to a nonlinear transformation, and their amplitudes were converted into a decibel scale to improve perceptual interpretability. This representation aligns with the human auditory perception system, which has nonlinear responses to frequency, particularly with reduced sensitivity at higher frequencies. Figure 3 is the flowchart of the mel-spectrogram feature extraction.

2.4.2. Prosodic Features

Zero Crossing Rate (ZCR)

ZCR reflects the temporal density of polarity reversals in an audio signal, serving as an indicator of its oscillatory behavior. Its corresponding mathematical formulation is expressed as equation (4).

$$Z(i) = \frac{1}{zW_L}\sum_{n=1}^{W_L} |\ sign[x_i(n)] - sign[x_i(n-1)]\ | \qquad (4)$$

Where sign () is the sign function by (5)

$$sign[x_i(n)] = \begin{cases} -1, x_i(n) < 0 \\ 1, x_i(n) \geq 0 \end{cases} \qquad (5)$$

And x_i(n), corresponds to the nth discrete sample of the ith segmented frame of the speech signal. Each frame consists of W_L samples, which determines the temporal resolution of the analysis.

Root Mean Square Energy (RMSE)

RMSE is used as a measure of the energy content of a speech frame. It is computed as the square root of the average of squared amplitude values within each frame, as given in equation (6).

$$RMSE = \sum_{n=1}^{W_L} |\ x_i(n)|^{\ 2} \qquad (6)$$

Where $x_i(n)$ represents the signal amplitude at the ith frame and W_L denotes the frame length [37]. In speech emotion analysis, RMSE reflects the intensity or loudness of speech segments and is used as a basic energy-based feature [38]. Similar methods were used in the second stage of feature extraction separately. The only difference is that in this case, two data augmentations, namely noise injection and stretching, were added.

2.4.3. Data Augmentation (DA)

Using a Gaussian-based stochastic process, as expressed in (7):

$$y(t) = x(t) + \alpha.n(t) \qquad (7)$$

Here, n(t) represents a randomly generated noise component characterized by a zero-centered distribution with variance σ2, which is applied to the input signal. α is a scaling factor that controls the noise level and depends on the SNR value, which defines how much noise will be injected into the signal, as in (8)

$$SNR = 10.log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) \qquad (8)$$

Time Stretching

The original speech signal is time-stretched with a constant stretch factor of 0.91 and used to simulate changes in speaking rate. The transformation is implemented using a phase vocoder-based approach. Analysis is performed using frames of 512-1024 samples with a hop size of N/4 (i.e., 75% overlap), and the reconstruction of the signal is done using a Hann window.

2.5. Classification Model: Design and Setup

The proposed model is built with a deep learning CNN using Python 2.14.1. The proposed architecture consists of four convolutional layers activated using ReLU (Rectified Linear Unit), each followed by max-pooling operations for dimensionality reduction. The learned feature representations are further processed through two dense layers with ReLU activation, and finally classified using a SoftMax output layer. Figure 4 is its corresponding detailed neural architecture. Dropout is incorporated between fully connected layers to prevent overfitting and enhance generalization. A Rectified Linear Unit (ReLU), formulated as $ReLU\ (x) = max\ (0, x)$ is incorporated between the fully connected layers to enhance non-linear feature learning. The output layer is configured with a neuron count equal to the number of target classes, enabling direct multi-class prediction [39]. Output activations are normalized via SoftMax to represent relative probabilities for each target class. Model parameters are optimized using the Adam algorithm to enable efficient convergence.

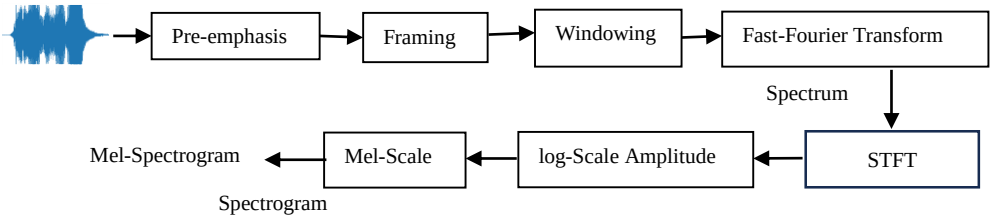


Fig. 3 Flowchart of Mel-Spectrogram Feature Extraction

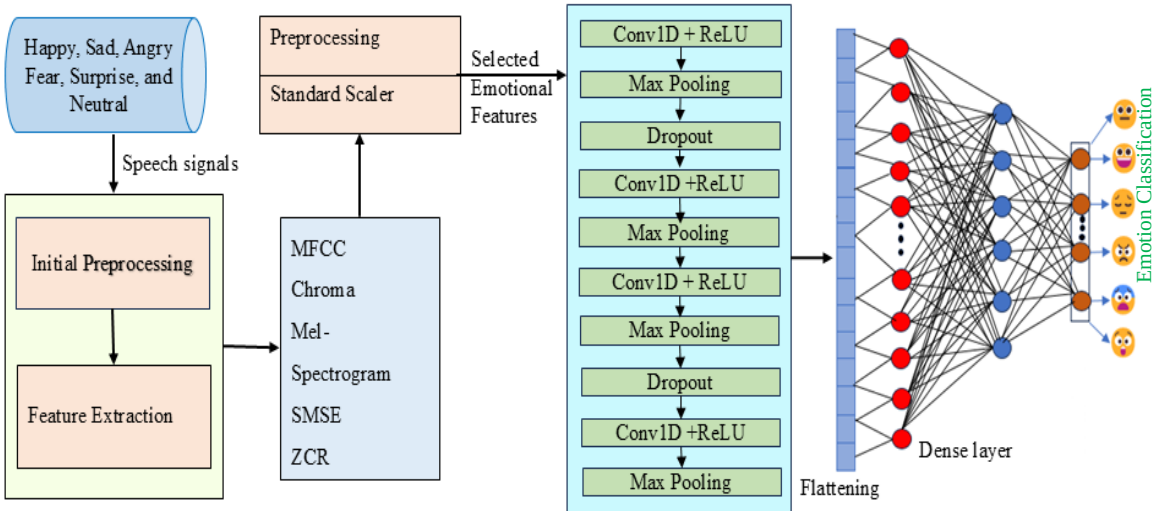


Fig. 4 Architecture for the proposed SER model

Additive White Gaussian Noise (AWGN) injection: Random noise is superimposed onto each audio waveform

To ensure effective model training and evaluation, the dataset was randomly split into 80% training data and 20% testing data. Model configurations corresponding to the baseline (original data) and augmentation-based feature extraction settings are detailed in Tables 1 and 2, respectively, with training conducted over 100 epochs using a batch size of 64.

3. Experimental Results and Analysis

This section presents a detailed analysis of the experimental results to assess the performance of the proposed SER model. ZCR-based feature analysis reveals distinct fluctuations for happy and angry emotions compared to the other emotional categories. For the remaining emotional categories, only minimal variation was observed. Table 3 presents the zero-crossing counts obtained from a sample speech "देरहागिरिया बान्था मोनगोन" (The winner will get an award) under both non-augmented and augmented conditions. Figure 5 illustrates the corresponding waveform of this sample in six emotions (a, b, c, d, e, and f) under normal recording conditions. Figure 6 depicts an example of the same, but adding data augmentations, namely, noise injection (all figures are not included). Figures 7(a) and 7(b) highlight differences in RMSE behavior between non-augmented and augmented data conditions. The results show remarkable differences from each other. From the visual graphs of MFCC, it is seen that MFCCs with non-augmented and augmented are unable to be distinguished clearly from one another. Figure 8(a) and 8(b) present this result for the same speech (देरहागिरिया बान्था मोनगोन) in Happy emotion. The extracted chroma feature in both conditions also shows a clear, remarkable difference.

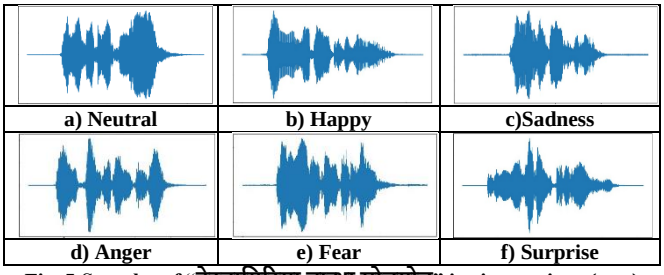


Fig. 5 Samples of "देरहागिरिया बान्था मोनगोन" in six emotions (raw)

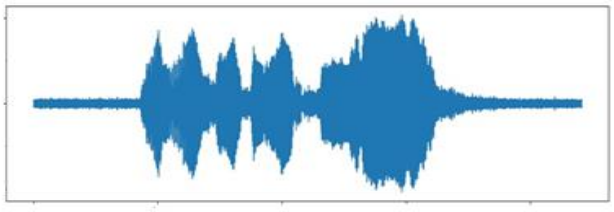


Fig. 6 Speech sample of "देरहागिरिया बान्था मोनगोन" in Happy emotion after adding noise injection

Table 1. CNN architecture under no augmentation: convolutional layers configured with 96, 64, and 34 filters

Layer (Operation)	Feature Dimension	Trainable Parameters
Block 1 – Convolution	(None,2,96)	576
Downsampling (Max Pooling)	(None,1,96)	0
Regularization (Dropout)	(None,1,96)	0
Block 2 – Convolution	(None,1,96)	46176
Downsampling (Max Pooling)	(None,1,96)	0
Block 3 – Convolution	(None,1,64)	30784
Downsampling (Max Pooling)	(None,1,64)	0
Regularization (Dropout)	(None,1,64)	0
Block 4- Convolution	(None,1,34)	10914
Downsampling (Max Pooling)	(None,1,34)	0
Flattening Stage	(None,34)	0
Dense Layer	(None,16)	560
Dropout Stage	(None,16)	0
Output Layer (Softmax)	(None,6)	102

Table 2. CNN architecture with augmented data: convolutional layers with 256, 128, and 64 filters

Layer (Operation)	Feature Dimension	Trainable Parameters
Block 1 - Convolution	(None,162,256)	1024
Downsampling (Max Poolinng	(None,81,256)	0
Block 2 - Convolution	(None,81,256)	196864
Downsampling (Max Pooling)	(None,41,256)	0
Block 3 - Convolution	(None,41,128)	98432
Downsampling (Max Pooling)	(None,21,128)	0
Regularization (Dropout)	(None,21,128)	0
Block 4 - Convolution	(None,21,64)	24640
Downsampling (Max Pooling)	(None,11,64)	0
Flattening Stage	(None,704)	0
Dense Layer	(None,32)	22560
Dropout Stage	(None,32)	0
Output Layer (Softmax)	(None,6)	198

Table 3. Number of ZCR of one of the speeches in different emotions with and without data augmentation

Speech (देरहागिरिया बान्था मोनगोन) in Emotion	Number of ZCR (Feature extracted without data augmentation)	Number of ZCR (Feature when extracted with data augmentation
Neutral	**4**	**4**
Happy	**11**	**11**
Sadness	**5**	**12**
Angry	**1**	**11**
Fear	**4**	**4**
Surprise	**21**	**21**

Also shows a clear, remarkable difference. Figure 9 is an example of a visual representation of the speech "देरहागिरिया बान्था मोनगोन" in Happy emotion. In the case of the mel-Spectrogram, when data is augmented, the color of the waveform becomes black in some periods. It represents that the amplitude is reduced, which means energy is reduced at times. But the brightness of the color remained the same. It shows that irrelevant energy(loudness) of audio is discarded. The brighter the color, the higher the energy. Since energy signifies the loudness of a sound, it is useful for classifying anger, surprise, and fear from sadness and neutral emotions. Based on the arousal dimension of emotion representation, high-energy states include emotions such as joy, anger, and fear, while low-energy conditions are typically represented by sadness, boredom, and neutrality [40].

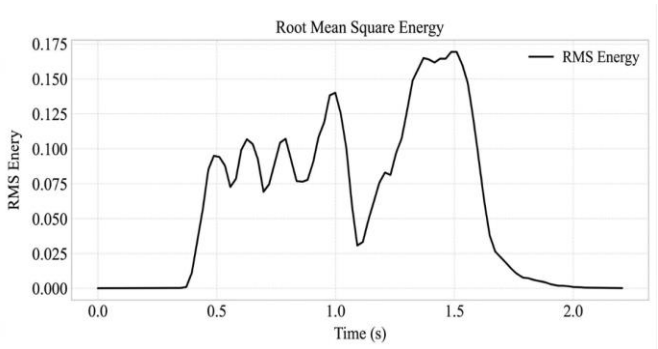


Fig.7 (a) RMSE before adding data augmentation

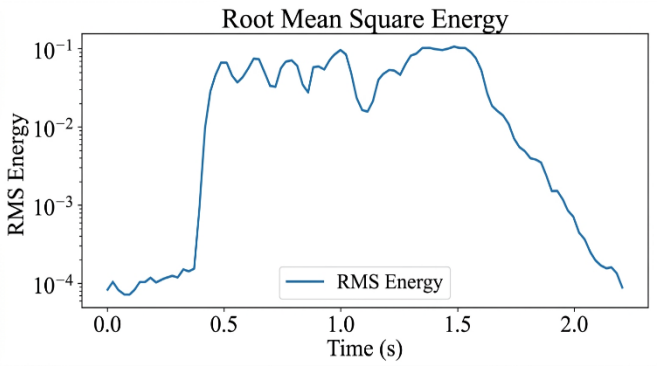
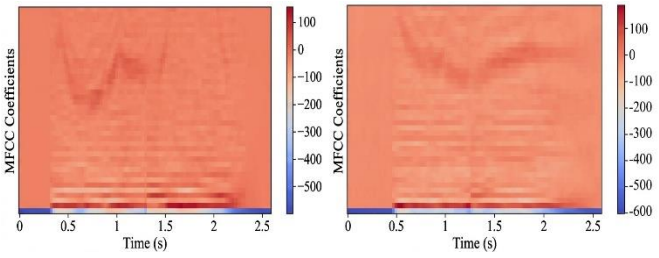
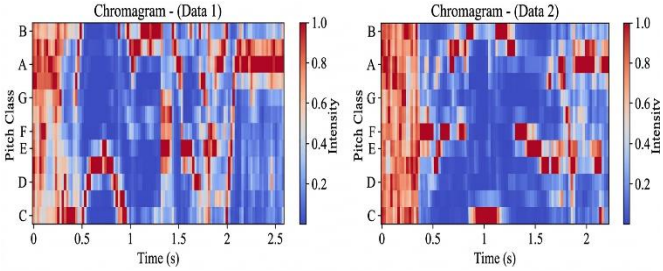


Fig. 7 (b) RMSE after adding data augmentation



(a) **(b)**

Fig. 8 MFCC, (a) normal (b) after adding data augmentation.



(a) **(b)**

Fig. 9 Chroma (a) Normal, (b) After adding data augmentation.

Other emotional states, such as anger, happiness, and fear, have relatively high amplitude [41, 42]. The total parameters, including trainable weights derived by the classification model extracted features with data augmentation, are 343,718, and those derived by the classification model using extracted features without data augmentation are 98,112. This indicates an enhancement of 245606 parameters of data.

The model achieved training and testing accuracies of 85.67% and 66.66% on non-augmented data, and 94.57% and 81.71% on augmented data, respectively. This indicates that applying augmentation to the features improved the recognition rate of the model by 15.05%. Figures 10 and 11 depict the model's loss and accuracy curves over the non-augmented vs augmented datasets, respectively.

Figures 12 and 13 illustrate the classification performance through confusion matrices for models trained on non-augmented and augmented datasets, respectively. Table 4 presents the detailed results of the proposed SER model in two conditions.

4. Comparative Analysis with Existing Studies

The comparison analysis showing the development of emotional speech processing methods on the Bodo language is presented in Table 5.

It is important to note that the datasets, strategies of extracting features, and modeling frameworks vary across the studies; thus, the comparison is qualitative and demonstrates the methodological development in the same low-resource Bodo language context instead of being strictly quantitative. Early Bodo emotion recognition relied on GMM and HMM with input features selected from spectral and prosodic acoustic characteristics [12, 13].

Although these approaches demonstrate that emotion recognition is feasible, they struggle to represent intricate speech features, making it difficult to capture nuanced emotional variations and high-dimensional patterns.

Moreover, research [14, 15] did not concentrate on emotion recognition but rather on the synthesis of emotional speech, focusing on prosodic speech-generation models, but not classification. Even though these studies are very insightful into the emotional characteristics of Bodo speech, they are not directly comparable to recognition-based frameworks.

These limitations make it clear that more flexible modelling approaches are important, which could improve recognition performance, especially with the low-resource Bodo language. In this context, the proposed CNN-based model utilizes multiple spectral features and automatic feature learning, which allows better discrimination of emotional states and proves a remarkable improvement as compared to previous studies.

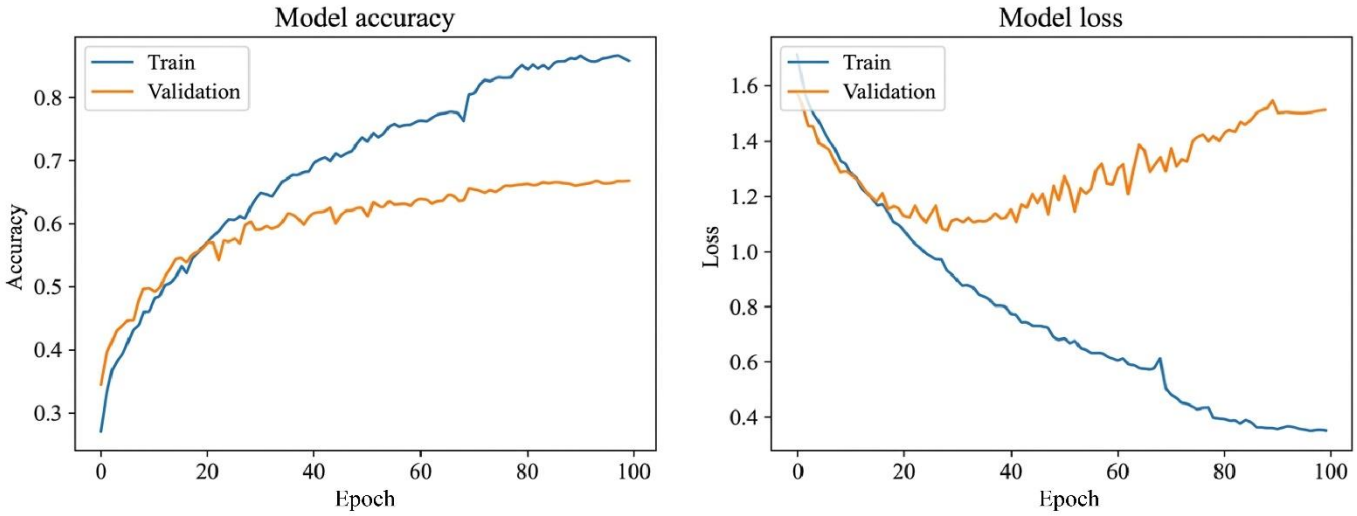
The findings of the proposed approach show that the model is better at capturing more complex emotional patterns in comparison to previous approaches using GMM and HMM (75.12% and 55.12% accuracy) based on the hierarchical feature representation. Further, unlike synthesis-oriented literature [14, 15], the current study, which focused on creating emotional speech, prioritizes correct categorization of emotional states and thus fills a research gap in Bodo speech processing with improved performance.

5. Discussion and Conclusion

Feature extraction and classification of emotion in the Bodo language have been experimented with in two approaches: one without augmenting data, and another that augments the data.

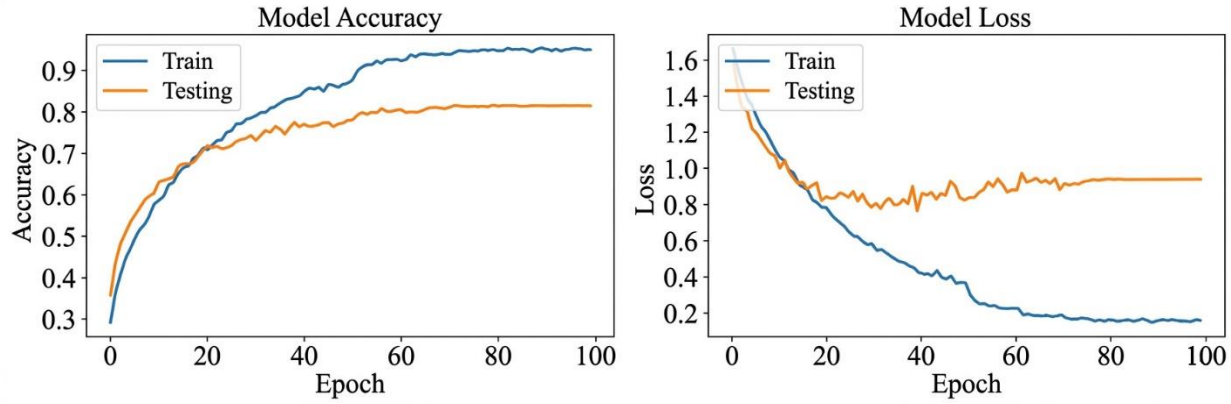
The results found that extracted features and parameters are highly different in the two cases. Training a model and achieving a high performance depends on preparing the dataset, feature extraction, and the classification method used. Data augmentation generates additional training samples by slightly altering the original data, which helps improve model robustness.

When data augmentation was added to feature extraction, the performance of the previous model (without augmentation) achieved a huge change, with an improved rate of 15.05% error-free. This is possible because of the greater variety of training data, which strengthens the model's generalizability and mitigates overfitting.



(a) Training & Testing (Validation) Accuracy **(b) Training & Testing (Validation) loss**

Fig.10. Model Accuracy graph without data augmentation, the blue curve represents training, and the yellow represents testing.



(a) Training & Testing (Validation) Accuracy **(b) Training & Testing (Validation) loss**

Fig. 11 Model Accuracy graph with data augmentation, the blue curve represents training, and the yellow represents testing.

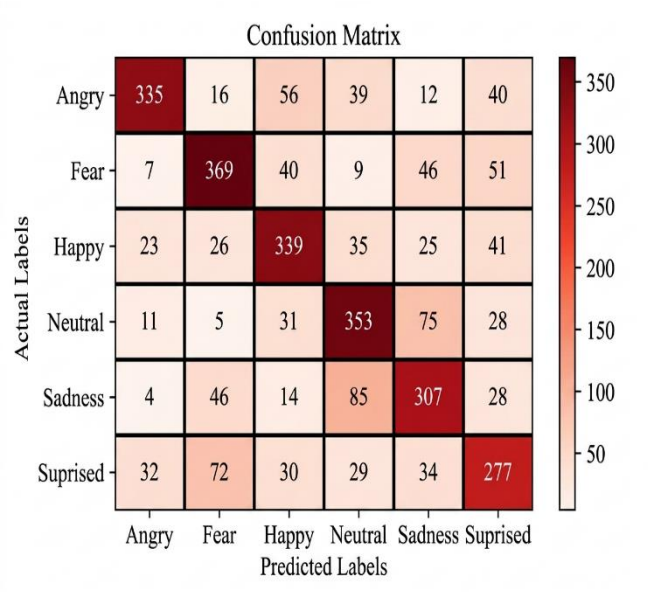


Fig. 12 Confusion Matrix of SER model without data augmentation

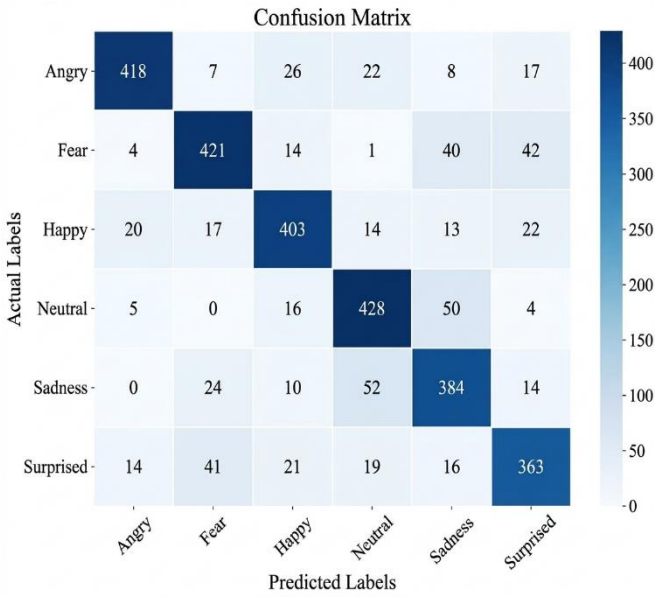


Fig. 13 Confusion Matrix of SER model with data augmentation

Moreover, the successful combination of both MFCC and Mel-spectrogram characteristics is another factor that promotes improvement in performance, as it captures complementary spectral features. CNN-based architecture effectively acquires spatial patterns using these rich feature representations and achieves improved discrimination of the emotional states. This study describes an effective method of speech emotion recognition concerning the Bodo language, which will be very helpful for further investigation. The proposed model can be used in smart rooms, business (or commercial) fields, and psychological diagnosis.

This study provides a robust and novel resource for future research endeavors for upgrade. This can also be added to the multilingual and cross-cultural speech emotion recognition system to advance SER technology.

Table 4. Performance comparison of the proposed model under augmented and non-augmented speech data conditions

Emotion	Precision (%)		Recall (%)		F1-Score (%)		Overall accuracy (%)	
	With DA	Without DA	With DA	Without DA	With DA	Without DA	With DA	Without DA
Angry	91	81	82	66	87	74	81.71	66.66
Fear	80	69	80	71	80	70		
Happy	78	66	83	69	81	68		
Neutral	82	64	86	70	84	67		
Sadness	77	62	78	63	78	62		
Surprise	78	60	76	58	77	59		

Table 5. Comparative Analysis of Existing Studies and Proposed Approach in Bodo Emotional Speech Processing

Study	Task	Features	Model	Focus of the study	Key contribution	Limitations
12	Emotion Recognition	MFCC, WPCC, tfMFCC, tfWPCC	GMM	Cross-lingual emotion recognition using advanced spectral features	Demonstrated effectiveness of wavelet-based features	Not specific to Bodo; depends on handcrafted features
13	Emotion Recognition	Pitch, MFCC	HMM	Analysis of emotional characteristics in Bodo speech	Highlighted the role of features in tonal speech	Limited performance and generalization
14	Emotional Speech Synthesis	Prosodic features	Rule-based + Template-based with GMM	Prosody modeling for emotional speech generation	Model of emotional variations	Not applicable to recognition task
15	Emotional speech synthesis and analysis	Acoustic + prosodic features	HMM	Emotional speech synthesis and analysis	emotional speech processing	Focus on synthesis; limited classification capability
Current study	Emotion Recognition	MFCC, Chroma, Mel-spectrogram, RMSE & ZCR	Deep CNN	Robust emotion recognition in low-resource Bodo speech	Automatic feature learning and improved classification	Limited to the low-resource Bodo language rather than other low-resource languages.

Note: The comparison is limited to studies related to the Bodo language only and is qualitative in nature as the datasets, feature extraction methods, and experimental conditions vary across studies. In addition, certain previous studies were concerned with the synthesis of emotional speech, but not recognition, and they incorporated elements to reflect the whole research scenario in Bodo emotional speech processing.

6. Discussion and Conclusion

Since pitch(F0) is the fundamental feature to distinguish the tone of a language, and the Bodo language is a tonal language, adding pitch with prosodic features and pitch-shifting data augmentation may extend emotion recognition using a tonal-based emotional speech corpus.

Conflict of interest

The authors confirm that there are no financial or personal interests that might have affected the conduct or reporting of this research.

Funding Statement

This research received no specific grant or support from any funding agency in the public, commercial, or non-profit organisations.

References

[1] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby, "Deep Learning Techniques for Speech Emotion Recognition: From Databases to Model," *Sensors*, vol. 21, no. 4, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Mai El Seknedy, and Sahar Fawzi, "Speech Emotion Recognition System for Human Interaction Applications," *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, Cairo, Egypt, pp. 361-368, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[4] D.J. France et al., "Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829-837, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[5] S. Maheshwari, R. Bhuvana, and S. Sasikala, "Emotion Recognition Using Deep Learning," *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, vol. 3, no. 1, pp. 16-22, 2023. [Publisher Link]

[6] Rizwan Ullah et al., "Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer," *Sensors*, vol. 23, no. 13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Gang Liu, Shifang Cai, and Ce Wang, "Speech Emotion Recognition Based on Emotion Perception," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, pp. 1-7, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Palak Kaushik, and Ashish Sharma, "Analysing Paralinguistic Information from Human Speech and its Applications in Medicine," *2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS)*, Bangalore, India, pp. 55-59, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Rashmi Rani, and Manoj Kumar Ramaiya, "Enhancing Speech Emotion Recognition with Multi-Modal Hybrid Features and CNN," *SSRG International Journal of Electronics and Communication Engineering*, vol. 12, no. 7, pp. 35-46, 2025. [CrossRef] [Publisher Link]

[10] Izza Nur Afifah, Tri Budi Santoso, and Titon Dutono, "Indonesian Speech Emotion Recognition: Feature Extraction and Neural Network Approaches," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 15, no. 4, pp. 3769-3778, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[11] Vidhi Sareen, and Seeja K.R., "Speech Emotion Recognition Using Mel Spectrogram and Convolutional Neural Networks," *Procedia Computer Science*, vol. 258, pp. 3693-3702, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[12] Aditya Bihar Kandali, Aurobinda Routray, and Tapan Kumar Basu, "Vocal Emotion Recognition in Five Native Languages of Assam Using New Wavelet Features," *International Journal of Speech Technology*, vol. 12, pp. 1-13, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[13] Uzzal Sharma, "Identification of Emotion from Speech Signal," *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, pp. 2805-2807, 2016. [Google Scholar] [Publisher Link]

[14] Laba Kr. Thakuria et al., "Integrating Rule and Template-Based Approaches to Prosody Generation for Emotional BODO Speech Synthesis," *2014 Fourth International Conference on Communication Systems and Network Technologies*, Bhopal, India, pp. 939-943, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[15] Kalita Barnali, "*Bodo Emotional Speech Synthesis and Recognition Using HMM*," Ph.D. Thesis, Gauhati University, 2018. [Publisher Link]

[16] Shashidhar G. Koolagudi, and K. Sreenivasa Rao, "Emotion Recognition from Speech: A Review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99-117, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[17] Masaki Kurematsu, Jun Hakura, and Hamido Fujita, "An Extraction of Emotion in Human Speech Using Speech Synthesis and Classifiers for Each Emotion," *WSEAS Transactions on Information Science and Applications*, vol. 5, no. 3, pp. 246-251, 2008. [Google Scholar] [Publisher Link]

[18] Akalpita Das, Purnendu Acharjee, and Pranhari Talukdar, "An Improved Approach of Emotion Recognition Combining Spectral and Prosodic Features with Reference to Assamese Language," *International Journal of Innovative Research and Advanced Studies*, vol. 4, no, 4, pp. 111-114, 2017. [Publisher Link]

[19] Kishor Bhangale, and Mohanaprasad Kothandaraman, "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network," *Electronics*, vol. 12, no. 4, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Kudakwashe Zvarevashe, and Oludayo Olugbara, "Ensemble Learning of Hybrid Acoustic Features for Speech Emotion Recognition," *Algorithms*, vol. 13, no. 3, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[21] Lamiaa Abdel-Hamid, Nabil H. Shaker, and Ingy Emara, "Analysis of Linguistic and Prosodic Features of Bilingual Arabic-English Speakers for Speech Emotion Recognition," *IEEE Access*, vol. 8, pp. 72957-72970, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[22] Yu Zhou et al., "Speech Emotion Recognition Using Both Spectral and Prosodic Features," *2009 International Conference on Information Engineering and Computer Science*, Wuhan, China, pp. 1-4, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[23] I. Manolekshmi, and M. A. Mukunthan, "Speech Emotion Recognition Using Hybrid Deep Learning and Ensemble Approaches," *SSRG International Journal of Electronics and Communication Engineering*, vol. 12, no. 1, pp. 216-235, 2025. [CrossRef] [Publisher Link]

[24] Uzzal Sharma, "*A Study on Intonation and Prosody of Bodo Language*," Ph.D. Thesis, Department of Instrumentation & USIC, Gauhati University, Assam, India, 2012. [Publisher Link]

[25] Satyendranarayan N. Goswami, Studies in Sino-Tibetan Language, Assam, India: Mandira Goswami, 1988. [Online]. Available: https://search.worldcat.org/title/Studies-in-Sino-Tibetan-languages/oclc/246649500

[26] A. Brahma, *Modern Bodo Grammar*, 1st ed., vol. 1, no.1, Guwahati, India: N. L. Publications, 2012. [Google Scholar]

[27] Sanjib Narzary et al., "Generating Monolingual Dataset for Low Resource Language Bodo from Old Books Using Google Keep," *Proceedings of the 13th Conference on Language Resources and Evaluation*, Marseille, France, pp. 6563-6570, 2022. [Google Scholar] [Publisher Link]

[28] Anusha Koduru, Hima Bindu Valiveti, and Anil Kumar Budati, "Feature Extraction Algorithm to Improve the Speech Emotion Recognition Rate," *International Journal of Speech Technology*, vol. 23, no. 1, pp. 45-55, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[29] Babak Basharirad, and Mohammadreza Moradhaseli, "Speech Emotion Recognition Methods: A Literature Review," *AIP Conference Proceedings*, vol. 1891, no. 1, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[30] Samson Akinpelu, and Serestina Viriri, "Robust Feature Selection-Based Speech Emotion Classification Using Deep Transfer Learning," *Applied Sciences*, vol. 12, no. 16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[31] Shadi Langari, Hossein Marvi, and Morteza Zahedi, "Efficient Speech Emotion Recognition Using Modified Feature Extraction," *Informatics in Medicine Unlocked*, vol. 20, pp. 1-11, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[32] Dimitrios Ververidis, and Constantine Kotropoulos, "Emotional Speech Recognition: Resources, Features, and Methods," *Speech Communication*, vol. 48, no. 9, pp. 1162-1181, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[33] LeiLei Xu et al., "A Large-Scale Remote Sensing Scene Dataset Construction for Semantic Segmentation," *International Journal of Image and Data Fusion*, vol. 14, no. 4, pp. 299-323, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[34] Iqbal H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Computer Science*, vol. 2, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[35] Laith Alzubaidi et al., "A Survey on Deep Learning Tools Dealing with Data Scarcity: Definitions, Challenges, Solutions, Tips, and Applications," *Journal of Big Data*, vol. 10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[36] Leland Roberts, Understanding the Mel Spectrogram, Medium, 2020. [Online]. Available: https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

[37] Sarfaraz Masood, Jeevan Singh Nayal, and Ravi Kumar Jain, "Singer Identification in Indian Hindi Songs Using MFCC and Spectral Features," *2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, Delhi, India, pp. 1-5, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[38] G. Tzanetakis, and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[39] Somenath Bera, Vimal K. Shrivastava, and Suresh Chandra Satapathy, "Advances in Hyperspectral Image Classification Based on Convolutional Neural Networks: A Review," *CMES - Computer Modeling in Engineering and Sciences*, vol. 133, no. 2, pp. 219-250, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[40] Juraj Kacur et al., "On the Speech Properties and Feature Extraction Methods in Speech Emotion Recognition," *Sensors*, vol. 21, no. 5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[41] Petri Laukka et al., "The Expression and Recognition of Emotions in the Voice across Five Nations: A Lens Model Analysis Based on Acoustic Features," *Journal of Personality and Social Psychology*, vol. 111, no. 5, pp. 686-705, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[42] Roza G. Kamiloğlu, Agneta H. Fischer, and Disa A. Sauter, "Good Vibrations: A Review of Vocal Expressions of Positive Emotions," *Psychonomic Bulletin & Review*, vol. 27, no. 2, pp. 237-265, 2020. [CrossRef] [Google Scholar] [Publisher Link]