Exploring Speech Emotion Recognition in Tribal Language with Deep Learning Techniques

Original Scientific Paper

Subrat Kumar Nayak*
Department of Computer Science and Engineering,
Siksha 'O' Anusandhan Deemed to be University,
Bhubaneswar, Odisha, India
subratsilicon28@gmail.com

Ajit Kumar Nayak
Department of Computer Science and Information Technology, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha, India
ajitnayak@soa.ac.in

Smitaprava Mishra
Department of Computer Science and Information Technology, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha, India
smitamishra@soa.ac.in

*Corresponding author

Prithviraj Mohanty
Department of Computer Science and Information Technology, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha, India
prithvirajmohanty@soa.ac.in

Nrusingha Tripathy
Department of Computer Science and Engineering,
Siksha 'O' Anusandhan Deemed to be University,
Bhubaneswar, Odisha, India
nrusinghatripathy654@gmail.com

Kumar Surjeet Chaudhury
Department of Computer Engineering,
KIIT Deemed to be University,
Bhubaneswar, Odisha, India
surjeet.chaudhuryfcs@kiit.ac.in

Abstract *– Emotion is fundamental to interpersonal interactions since it assists mutual understanding. Developing human-computer interactions and a related digital product depends heavily on emotion recognition. Due to the need for human-computer interaction applications, deep learning models for the voice recognition of emotions are an essential area of research. Most speech emotion recognition algorithms are only deployed in European and a few Asian languages. However, for a low-resource tribal language like KUI, the dataset is not available. So, we created the dataset and applied some augmentation techniques to increase the dataset size. Therefore, this study is based on speech emotion recognition using a low-resourced KUI speech dataset, and the results with and without augmentation of the dataset are compared. The dataset is created using a studio platform for better-quality speech data. They are labeled using six perceived emotions:* ସଡାଙ୍ଗି *(angry),* ରେହା *(happy),* ଆଜି *(fear),* ବିକାଲି *(sad),* ବିଜାରି *(disgust), and* ଡ଼େକ୍ *(surprise). Mel-frequency cepstral coefficient (MFCC) is used for feature extraction. The deep learning technique is an alternative to the traditional methods to recognize speech emotion. This study uses a hybrid architecture of Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs) as classification techniques for recognition. The results have been compared with existing benchmark models, with the experiments demonstrating that the proposed hybrid model achieved an accuracy of 96% without augmentation and 97% with augmentation.*

Keywords*:* *KUI Dataset, Speech Emotion Recognition, Deep Learning, Long Short-Term Memory, Data Augmentation*

Received: May 3, 2024; Received in revised form: September 23, 2024; Accepted: September 23, 2024

1. INTRODUCTION

Speech is the simplest, quickest, and most natural way to establish interaction between humans among the numerous forms of communication. A crucial component of regular human activity is emotion. Emotions support decision-making and help people understand one another. It facilitates communication in terms of safety and security. Human emotions can be recognized using a variety of modalities, including speech, writing, and facial expressions. Speech Emotion Recognition (SER) aims to identify emotions as they are communicated in spoken language. While speech communication between humans and machines is improving, it is still not interactive, natural, or organic communication as all the machines are not fully equipped to understand human emotions, specifically in low-resource scenarios. This issue has given rise to a new area of study for researchers. Speech Emotion Recognition is the term for techniques that can successfully assist us in comprehending human emotions by identifying the speaker's emotional state from their speech. Speech can clearly express emotions, which may be utilized later to extract essential semantic information from the uttered words and enhance the effectiveness of speech recognition [1].

SER systems can be used in several applications that need human interaction, such as the caller's emotions in a call centre that tracks the problem. A device that could behave like a human is also thought to require adding emotion recognition features. Doctors can learn about the patient's psychological and physical condition, which is an excellent achievement in the case of speech-emotion recognition. As a result, many researchers are getting more interested in SER research to create a recognition model that performs better.

Most speech emotion datasets are available in German, English, and Spanish. Several SER studies have also been used in languages, including Odia, Tamil, Mandarin, and other European and Asian languages. For KUI, a Low Resourced Tribal language, there is a lack of speech emotion dataset even though it is one of the tribal languages of Odisha, spoken by over 10 million people in the Kandhamal District of Odisha and other states of India [2]. This work addresses this critical gap by developing a novel dataset specifically for the KUI language, thereby facilitating new research and development in underrepresented languages and advancing the field of Natural Language Processing (NLP). It introduces a hybrid model combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for emotion classification. This innovative approach leverages the strengths of both architectures: CNNs are effective for spatial feature extraction, while LSTMs excel at capturing temporal dependencies. By employing this hybrid model, we explore its performance using both original and augmented datasets, enhancing the diversity of training data through techniques such as noise injection, time shifting, random gain, and polarity inversion. Additionally, we detail the design and development process of our CNN-LSTM model, including parameter selection and feature extraction using Mel-Frequency Cepstral Coefficients (MFCCs) [3]. This comprehensive approach contributes to understanding emotion recognition in the KUI language and is a foundation for future research in low-resource language processing [4].

The structure of this document is as follows. In Section II, relevant research is provided, and the findings are discussed. The procedure for creating the KUI dataset and the feature extraction technique is explained in Section III. The suggested model categorization is shown in Section IV, and the performance metrics are shown in Section V. Section VI provides the outcomes. The paper's conclusion, which includes a scope and future work, is included in Section VII.

2. LITERATURE SURVEY

Researchers now have more efficient opportunities to develop SER models using deep learning. Language, dataset properties, feature extractions, and various classifiers are all important factors in speech-emotion recognition systems [5]. Several researchers have applied machine learning and deep learning to detect emotions in English speech.

More research needs to be done on the KUI language. Some KUI commands are trained and tested using deep learning models, which yields significant results [6]. Several studies have been conducted to identify emotions in speech for a wide range of languages, but they have yet to be done for KUI speech, as shown in Table 1.

Jo et al. [7] presented a CNN-based transfer learning model for recognising emotions in Korean. Using Chosun University's Korean speech emotion database, they achieved an accuracy of 94.91%. This is a good performance, but accuracy is still required to improve in a tribal language.

Taj et al. [8] presented a 1D-CNN model for voice emotion identification in the URDU language. Although Urdu has fewer resources, the model still gives an accuracy of 97%. They used the URDU dataset for the experiments. In this research, the authors compare the results with existing work based on CNN and SVM.

Amjad et al. [9] proposed deep convolutional neural networks (DCCNs) with two layers of long-short-term memory (LSTM). They identify the emotions from the spontaneous speech. They used SAVEE, IEMOCAP, and BAUM-1 datasets for emotion recognition. They got an accuracy of 94.78% for speaker-independent with the raw SAVEE dataset.

Atila et al. [10] proposed an attention-guided 3D CNN-long short-term memory (LSTM) system for speech-based emotion identification. They used three datasets: RAVDESS, RML, and SAVEE, and a mixture of them. They compared the outcomes using the F1-score, sensitivity, specificity, and classification accuracy. The RML dataset, which consists of different languages, gave an accuracy of 93%.

Table 1. Emotion Recognition in different languages

Ref	Year of publication	Dataset	Language	Model Used
[11]	2023	RAVDESS	English	CNN-LSTM
[12]	2023	CREMA-D	English	LSTM
[13]	2023	EMO-DB SAVEE RAVDESS	German English English	CNN
[14]	2023	EMO-DB ENTERFACE05	German English	DCNN-GWO
[15]	2023	RAVDESS	English	Deep LSTM
[16]	2023	EMO-DB	German	DNN-SVM

[17]	2023	RAVDESS EMO-DB	English German	1-D DCNN
[18]	2023	RAVDESS TESS CREMA-D IEMOCAP	English English English English	CNN
[19]	2023	AVEC, AFEW	English	ASP-MTL
[20]	2022	EMO-DB URDU	German Urdu	CADCN
[21]	2021	NSSED	Sindhi	1D- CNN
[22]	2020	EMO-DB	German	CNN
[23]	2020	EMO-DB IEMOCAP RAVDESS	German English English	BiLSTM
[24]	2019	ARABIC	Arabic	CNN-LSTM
[25]	2019	EMIRATI	Arabic	GMM-DNN
[26]	2018	CUSTOM	Malayalam	DNN
[27]	2018	NCKU-ES	Chinese	CNN LSTM
[28]	2017	TELGU EMOTION	Telgu	ANN KNN
		TAMIL EMOTION	Tamil	ANN KNN
[29]	2013	MANDARIN	Chinese	HMM
[30]	2010	ODIA	Odia	K-Means

3. MATERIAL AND METHODS

3.1. DATASET

A KUI emotion speech dataset was generated from several speakers in the Kandhamal district of Odisha for this study. A platform has been developed for data preparation in the KUI language [31]. The platform snapshot is shown in Fig. 1. The manual dataset preparation takes more time. We have taken six common emotions in the dataset: ସଡାଙ୍ଗି, ରେହା, ଆଜି, ବିକାଲି, ବିଜାରି and ଡେ଼କ୍. The meaning of the KUI emotions is shown in Table 2.

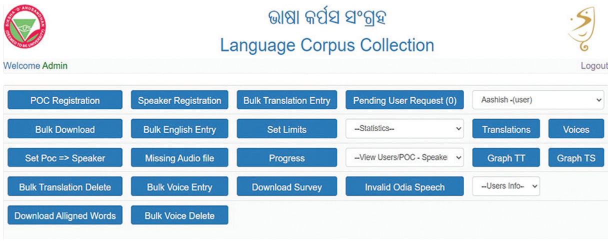

Fig. 1. Speech emotion data collection platform

There are 2,383 expressions of 6 different emotions in our dataset. At a rate of 16 kHz, each file is saved in wave format as linear 16-bit single-channel Pulse Code Modulation (PCM) values. We used a Zoom audio recorder, a laptop, and a mobile device to capture the speech expressions. To reduce noise, it is recorded in a studio. Our collection's male-to-female voices are equal to avoid bias. After the completion of data collection, the sampling rate of all the recordings is checked. The pre-processing stage involved the removal of unwanted background noise. The total dataset includes 2383 different utterances.

Table 2. KUI Speech emotion and its meaning

Emotion (Kui)	Emotion (English)	Emotion (IPA)	#files		
			Original	Augmented	Total
ସଡାଙ୍ଗି	Angry	ɾagɔ	397	1588	1985
ରେହା	Happy	k^{h}usi	400	1600	2000
ଆଜି	Fear	b^{h}ɔjɔ	398	1592	1990
ବିକାଲି	Sad	d̪uhk^{h}ɔ	397	1588	1985
ବିଜାରି	Disgust	biɾɔkt̪i	396	1584	1980
ଡେ଼କ୍	Surprise	astʃɔɾdʒjɔ	395	1580	1975

3.2. METHODOLOGY

The flow diagram of speech emotion recognition is shown in Fig. 2. KUI speech emotion recognition mainly depends on feature extraction methods. It mostly takes several features out of the audio streams. After feature extraction, the features are sent into the classifier, frequently referred to as input. The various emotions are identified using the inputs. The first and foremost stage of speech conversion is feature extraction. The vital goal of this process is to find the details of a speech signal. Over time, feature extraction needs to be consistent. In speaking, it must happen regularly and dynamically. There are several kinds of feature extraction techniques available [32]. The Mel-Frequency Cepstral Coefficient (MFCC) feature extraction approach is used in this paper.

3.2.1. DATA AUGMENTATION

The volume of data directly impacts the performance of deep learning. Deep learning is subject to overfitting when used with limited datasets. Typically, the first thing that is thought about is approaching this challenge from the data level [33].

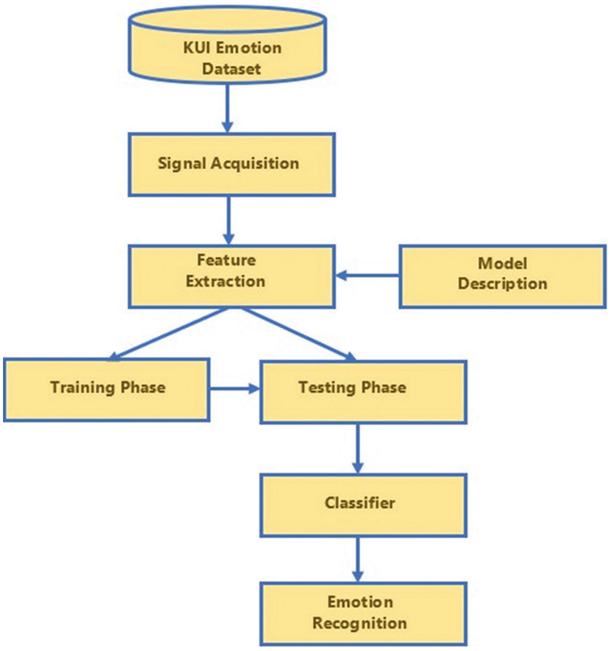

Fig. 2. Flow diagram of KUI emotion recognition

For low-resourced languages, gathering extra information can be challenging. Data augmentation aims to enhance the volume of data required for speech recognition system training. Data augmentation effectively increases current data availability and allows model training without requiring new data. We can present the audio data in two ways, i.e., raw audio and spectrogram [34]. Data augmentation, a regularisation technique, generates new, slightly modified samples from the original data to increase the training set. In this paper, four data augmentation strategies are considered as follows:

i. Noise injection: We can use different types of noises for data augmentations. Noise may be environmental, background, white, or thick. In our case, we inject white noise, adding a random value to the original data.

ii. Shifting time: It just moves the audio left or right. If there is not enough trailing silence, the audio will wrap around.

iii. Random gain: Random gain can change the amplitude. This method measures the loudness of the audio. The original audio signal is multiplied by a random factor, which converts into an augmented signal.

iv. Polarity: An audio time-frequency representation is independent of polarity. An audio data augmentation that reverses the phase of an audio stream might be beneficial for raw waveforms. All phases will cancel out when the phase-inverted signal is added to the original signal. In this case, the signal is multiplied by -1 for the phase insertion.

3.2.2. FEATURE EXTRACTION

Mel-Frequency Cepstral Coefficient (MFCC):

The European Telecommunications Standards Institute defines the MFCC algorithm. MFCC is an efficient method for feature extraction. It considers the frequencies of human perception sensitivity, which can be treated as one of the best tools for speech recognition. The block diagram of MFCC is described in Fig. 3. Various types of steps for finding the Mel-Frequency Cepstral Coefficient [35] are described in eqn 1 through 5.

Pre-emphasis: This step increases energy to higher frequencies, possibly related to vowels with more energy at low frequencies than at high frequencies. It improves the detection accuracy and uses a filter to increase higher frequencies.

$$y[n]=x[n]-\alpha\cdot x[n-1] \qquad (1)$$

Where $x[n]$ is the input signal, and $y[n]$ is the pre-emphasized signal

Framing: In this step, the signal is split into small time frames where each frame can be independently analyzed and represented as a single feature vector. The frame time length of the extracted speech is 25-30 ms. The overlapping of the frames is very useful to reduce the loss of information. The advantage is not to do the Fourier transform across the entire signal.

$$x_f[i]=x[i.R:(i+1).R-1], i= 0,1,\ldots,L-1 \qquad (2)$$

Here, $x_f[i]$ represents the i-th frame. R is the frame length and L is the number of frames.

Windowing: It involves slicing the sound's waveform into different frames. But it cannot split at the boundary of the frame. For slicing, the audio signal amplitude should drop near the edge of a frame. Therefore, it is better to use Hamming windows to chop the signal.

$$x_w[i]=x_f[i]\cdot w[i], i=0,1,\ldots,R-1 \qquad (3)$$

Here, $x_w[i]$ is the windowed frame.

Fast Fourier Transform (FFT): The frequency domain is created from the time domain using the Fast Fourier Transform (FFT) technique, as the time domain calculation is more complicated than the prevalence domain.

$$X[k]=\text{FFT}(x_w) \qquad (4)$$

Where $X[k]$ is the spectrum of the frame.

Mel Filter Bank: The way of receiving the sound of our ears and the machine is different. We can differentiate easily if we hear sound at 10HZ and 30HZ, but it is not easy to distinguish if it becomes 1000HZ and 1020HZ. But the machine resolution is the same at all frequencies. Thus, the human hearing property will improve performance. Therefore, we use the mel scale to map the actual frequency. The power spectrum uses the filter bank to sum up the energies. This energy is applied with the algorithm to find the mel frequency co-efficient.

$$H_m[k]=\sum_{i=0}^{N-1} X[i]\cdot H_m[i,k], m=0,1,,,,,M-1 \qquad (5)$$

Here, $H_m[k]$ is the output of the m-th filter bank. N is the number of FFT points and M is the number of mel filters.

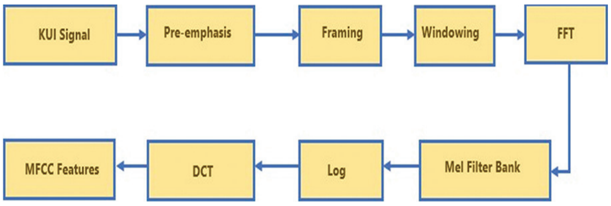

Fig. 3. Block diagram of Mel-Frequency Cepstral Coefficient

4. MODEL BUILDING

Many neural network approaches are utilized for voice emotion recognition. Speech emotion recognition in a low-resource KUI language has not yet been researched or developed. This work used the Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) models. Taking the above two models, we propose a hybrid model and compare the model accuracy with performance metrics. We also compare the accuracy before data augmentation and after data augmentation. We used Python to implement our research work. The data were split into three different training and testing sets for classification, i.e., 90%-10%, 80%-20%, and 70%-30%. We also compare the results by taking the different epoch sizes.

4.1. CONVOLUTIONAL NEURAL NETWORKS (CNN) MODEL

As Convolutional Neural Networks (CNNs) are good at identifying local patterns in spectrogram representations of audio data, they have become essential tools in speech emotion recognition [13]. For this purpose, CNNs usually have convolutional layers and pooling layers, allowing them to extract discriminative features from raw audio input automatically. Through multiple layers of convolution and pooling, CNNs hierarchically extract features at different levels of abstraction, facilitating the identification of emotional cues such as pitch variations, spectral changes, and temporal dynamics in speech signals. After learning these representations, fully connected layers are provided for classification, where the model can predict the emotional state associated with the input speech segment. With appropriate training data and optimization strategies, CNNs have demonstrated promising performance in various emotion recognition tasks, offering robustness to noise and variability in speech signals while requiring minimal preprocessing of the input data [36]. Using convolutional layers, the Convolutional Neural Network (CNN) in KUI speech emotion recognition extracts hierarchical characteristics from the speech signal representations, as stated in equation 6.

$$CNN(x)=Conv1D(ReLU(BatchNorm(x))) \tag{6}$$

Where x represents the input feature of the speech signal, $Conv1D$ denotes the $1D$ convolutional operation, $ReLU$ is the rectified linear activation function and $BatchNorm$ represents batch normalization and accelerates training.

4.2. LONG SHORT-TERM MEMORY (LSTM)

Emotion recognition has demonstrated the significant efficacy of Long Short-Term Memory (LSTM) models, especially when evaluating sequential input like speech. Recurrent Neural Networks, or RNNs, are particularly good at recognizing the temporal dynamics and long-range dependencies in voice signals. By processing speech input over time through recurrent connections with gated memory cells, LSTMs can learn intricate patterns and contextual information crucial for discerning emotional states [27]. This capability allows them to capture nuanced features like prosody, intonation, and subtle variations in speech, which indicate different emotions. Furthermore, LSTMs can handle variable-length input sequences, making them suitable for analyzing speech segments of varying durations by integrating LSTMs with additional layers, such as attention mechanisms or combining them with other architectures like CNNs. Researchers have demonstrated the accuracy of LSTM models in capturing the temporal dynamics and complex nuances inherent in emotional speech by achieving state-of-the-art performance in speech emotion recognition [37]. It is possible to build an LSTM model for emotion recognition, as stated in eqn 7.

$$LSTM(x)=LSTM(x_m, h_{m-1}, \mathrm{c}_{m-1}) \tag{7}$$

Where x_m symbolizes the input at that moment m, h_{m-1} denotes the previous hidden state, a time step m-1. c_{m-1} represents the previous cell state at the moment m-1 and LSTM denotes the cell operation.

4.3. HYBRID MODEL

We present a hybrid model that includes CNN and LSTM. The convolution layer in our HYBRID model shows the feature map sequence and recognizes significant areas and varied length utterances. In the activation layer, a non-linear activation function is used. The Rectified Linear Unit (ReLU) has been utilized. When a layer is dense, the Softmax activation function is applied. The design, development, and assessment of the model, parameter selection, and feature selection for the MFCC in the HYBRID model for emotion classification are the primary contributions of this study.

The model uses multiple Conv1D layers with various filter and kernel sizes (3,5), extracting significant features from the input data. The hierarchical extraction captures local and global patterns in the data, which is crucial for tasks such as emotion classification, where subtle differences in the features matter. The LSTMs employed capture temporal dependencies within the sequential data, while the CNN is excellent for spatial feature extraction, making the proposed hybrid architecture leverage the strengths of both architectures. After the feature extraction and sequence learning phases, the model flattens the output from the LSTM layer and feeds it into dense layers. The 512-unit dense layer with ReLU activation and batch normalization processes the features before the final classification layer, passing it to the output layer. It uses a softmax activation function for multi-class classification, making the model suitable for categorizing emotions into distinct classes. The KUI dataset with limited resources is used in this research. The hybrid model can be described as a sequence of operations performed on the input data, as stated below.

1. Input:
- Input data shape:
  (batch_size, sequence_length, 1)
2. Convolutional Layers:
- $Conv1D$ with 512 filters, kernel size 5, strides 1, and ReLU activations, followed by batch normalization and max pooling:

  $Conv1D$ 512 (x)
  $\rightarrow BatchNormalization(x)$
  $\rightarrow MaxPool1D(x)$
- Output shape:
  (batch_size, sequence_length/2, 512)
3. Convolutional Layers:
- Conv1D with 512 filters, kernel size 5, strides 1, and ReLU activations, followed by batch normalization and max pooling:

$$Conv1D_{512}(x) \rightarrow BatchNormalization(x) \rightarrow MaxPool1D(x)$$

- Output shape: (batch_size, sequence_length/4, 512)

4. Convolutional Layers:

- $Conv1D$ with 256 filters, kernel size 5, strides 1, and ReLU activations, followed by batch normalization and max pooling:

$$Conv1D_{256}(x) \rightarrow BatchNormalization(x) \rightarrow MaxPool1D(x)$$

- Output shape: (batch_size, sequence_length/8, 256)

5. Convolutional Layers:

- $Conv1D$ with 256 filters, kernel size 3, strides 1, and ReLU activations, followed by batch normalization and max pooling:

$$Conv1D_{256}(x) \rightarrow BatchNormalization(x) \rightarrow MaxPool1D(x)$$

- Output shape: (batch_size, sequence_length/16, 256)

6. Convolutional Layers:

- $Conv1D$ with 128 filters, kernel size 3, strides 1, and ReLU activations, followed by batch normalization and max pooling:

$$Conv1D_{128}\ (x) \rightarrow BatchNormalization(x) \rightarrow MaxPool1D(x)$$

- Output shape: (batch_size, sequence_length/32, 128)

7. LSTM Layer:

- LSTM layer with 256 units, returning sequences.

8. Flatten Layer:

- Flatten the output of the LSTM layer to prepare it for the dense layers.

9. Dense Layers:

- ReLU activation and 512-unit dense layer are followed by batch normalization.
- Output shape: (batch_size, 512)

10. Output Layer:

- Dense layer with 6 units and softmax activation for multi-classification.
- Output shape: (batch_size, 6)

5. PERFORMANCE MEASURE OF HYBRID DEEP LEARNING MODELS

Performance metrics in speech emotion recognition are crucial for evaluating the effectiveness of models in classifying emotional states from speech signals. Commonly used metrics include accuracy, precision, recall, and F1-score [38]. Mathematically, all the metrics are represented in eqn 8 through 11.

- **Precision**: Voice emotion recognition systems must achieve high precision for practical uses in sentiment analysis, customer service, human-computer interaction, and psychology research. This guarantees that the system can accurately recognize and react to spoken language's emotional content, which is essential for offering suitable and efficient communication interfaces and services.

$$Precision = TP/(TP+FP) \tag{8}$$

Where TP denotes the number of positive instances correctly classified by the model and FP denotes the number of negative instances incorrectly classified as positive by the model.

- **Recall**: Recall in speech emotion recognition refers to a system's or model's capacity to accurately identify every occurrence of an emotion class from the dataset's total number of instances of that emotion. It measures the system's ability to capture all relevant instances of a specific emotion without missing any.

$$Recall = TP/(TP+FN) \tag{9}$$

Where TP are the instances correctly identified as positive and FN denotes instances incorrectly identified as negative.

- **Accuracy**: Accuracy in speech emotion recognition refers to the degree to which a system can correctly identify the emotional state conveyed by human speech. This is typically measured as the percentage of correctly identified emotions out of the total number of emotions analyzed.

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \tag{10}$$

Where TN denotes the number of correctly predicted negative instances.

- **F1 Score**: An often-used statistic in speech emotion identification is the F1 score, which assesses how well emotion categorization models perform. The F1 score is a model accuracy metric considering the model's recall and precision.

$$F1\ Score = 2*(Recall*Precision)/(Recall+Precision) \tag{11}$$

6. RESULTS AND ANALYSIS

The study presented in this article compares prominent deep-learning algorithms to identify the emotion from KUI speech. We have considered six emotions ସଡାଙ୍ଗି (angry), ରେହା (happy), ଆଜି (fear), ବିକାଲି (sad), ବିଜାରି (disgust), and ଡ଼େକ୍ (surprise). We have compared the results of CNN and LSTM with our proposed HYBRID model. The experiment was done for data without augmentation as well as with augmentation. As mentioned before, we test deep learning models using our KUI dataset. We use 90:10, 80:20, and 70:30 ratios for training and testing. During the training phase, the batch size is set to 32. First, we train the model without data augmentation using different parameters for our dataset. We used 0.01 as the learning rate. Initially, we used an epoch size of 50 in our experiment and trained with various split ratios. The testing accuracy of CNN is 0.91, whereas LSTM gives 0.93, and our proposed model has a maximum accuracy of 0.94. Similarly, the training and testing procedure was done for different epoch sizes. The graphical representation of the testing accuracy of all models having a split ratio of 80:20 with varying sizes of epoch is shown in Fig. 4 through 7.

Next, we train the model with the augmented dataset. After the data augmentation methods, our dataset size increased to 11915, and we input it into our models. The graphical representation of testing accuracy of all models with data augmentation having a split ratio of 80:20 with different epoch sizes is shown in Fig. 8 through 9. Table 3 displays details of all tests' accuracy of both augmentation and without augmentation

It is evident in Fig. 4 to Fig. 9 that across all the epoch sizes taken, the hybrid model consistently achieves the highest accuracy, showing fewer fluctuations in accuracy, indicating a more stable training process than the individual CNN and LSTM models. It also shows the benefits of combining CNN and LSTM architectures. This is also shown in the classification report in Tables 3 and 4. All the models taken show convergence to a high accuracy over time, but the proposed hybrid model achieves the best accuracy faster and maintains it better.

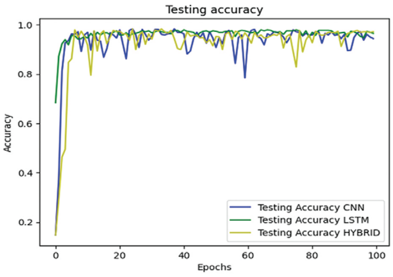

Fig. 4. Testing Accuracy without augmentation using epoch size 100

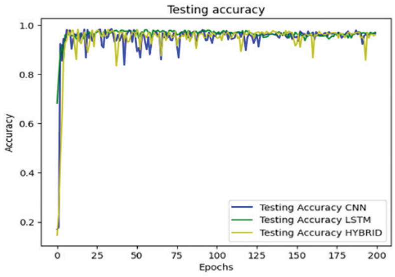

Fig. 5. Testing Accuracy without augmentation using epoch size 200

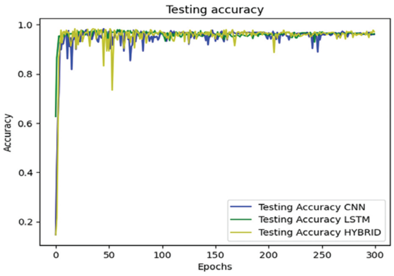

Fig. 6. Testing Accuracy without augmentation using epoch size 300

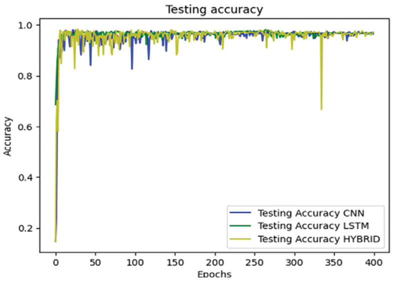

Fig. 7. Testing Accuracy without augmentation using epoch size 400

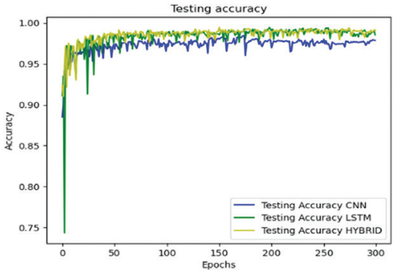

Fig. 8. Testing Accuracy with augmentation using epoch size 300

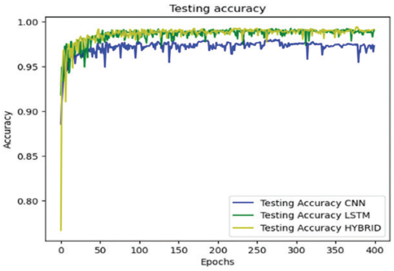

Fig. 9. Testing Accuracy with augmentation using epoch size 400

Table 3. Accuracy of different Models

		Accuracy of different Models					
No of epochs	split-ratio	without augmentation			with augmentation		
		CNN	LSTM	HYBRID Model	CNN	LSTM	HYBRID Model
50	70-30	0.81	0.92	0.93	0.91	0.93	0.95
	80-20	**0.91**	**0.93**	**0.94**	**0.96**	**0.97**	**0.97**
	90-10	0.86	0.93	0.95	0.94	0.95	0.96
100	70-30	0.89	0.93	0.94	0.93	0.94	0.96
	80-20	**0.93**	**0.94**	**0.95**	**0.96**	**0.96**	**0.97**
	90-10	0.91	0.95	0.95	0.94	0.96	0.97
200	70-30	0.91	0.94	0.95	0.95	0.95	0.96
	80-20	**0.94**	**0.95**	**0.96**	**0.96**	**0.97**	**0.97**
	90-10	0.92	0.95	0.96	0.95	0.96	0.97
300	70-30	0.93	0.95	0.96	0.94	0.96	0.96
	80-20	**0.94**	**0.95**	**0.96**	**0.94**	**0.95**	**0.97**
	90-10	0.92	0.95	0.95	0.94	0.96	0.97
400	70-30	0.91	0.93	0.94	0.94	0.95	0.96
	80-20	**0.95**	**0.95**	**0.96**	**0.95**	**0.96**	**0.97**
	90-10	0.91	0.95	0.96	0.95	0.97	0.97

The classification metrics of all the models with different epoch sizes without augmentation are shown in Table 4. Our proposed model's performance matrix gives better results than others. The classification metrics of all the models with different epoch sizes with augmentation are also shown in Table 4. Our proposed model's performance matrix gives better results than others in the case of data augmentation.

Table 4. Comparing various performance metrics

		Comparing various performance metrics																	
		without augmentation									**with augmentation**								
Epoch Size	**Performance indicators**	**CNN**			**LSTM**			**HYBRID Model**			**CNN**			**LSTM**			**HYBRID Model**		
		70-30	**80-20**	**90-10**	**70-30**	**80-20**	**90-10**	**70-30**	**80-20**	**90-10**	**70-30**	**80-20**	**90-10**	**70-30**	**80-20**	**90-10**	**70-30**	**80-20**	**90-10**
50	Precision	0.8	0.88	0.88	0.90	0.91	0.91	0.92	**0.92**	0.89	0.89	0.95	0.93	0.90	0.98	0.95	0.93	**0.91**	0.97
	Recall	0.84	0.89	0.82	0.91	0.95	0.93	0.93	**0.95**	0.94	0.91	0.97	0.96	0.95	0.97	0.94	0.95	**0.99**	0.94
	F1-Score	0.82	0.88	0.85	0.90	0.93	0.92	0.92	**0.93**	0.91	0.9	0.96	0.94	0.92	0.97	0.94	0.94	**0.95**	0.95
100	Precision	0.91	0.89	0.94	0.93	0.95	0.95	0.96	**0.95**	0.93	0.94	0.95	0.92	0.89	0.96	0.94	0.95	**0.97**	0.98
	Recall	0.9	0.94	0.91	0.94	0.93	0.96	0.93	**0.94**	0.98	0.91	0.93	0.94	0.94	0.94	0.98	0.96	**0.94**	0.95
	F1-Score	0.9	0.91	0.92	0.93	0.94	0.95	0.94	**0.94**	0.95	0.92	0.94	0.93	0.91	0.95	0.96	0.95	**0.95**	0.96
200	Precision	0.93	0.89	0.92	0.92	0.92	0.94	0.95	**0.93**	0.96	0.95	0.97	0.93	0.97	0.98	0.93	0.94	**0.98**	0.95
	Recall	0.9	0.92	0.93	0.95	0.93	0.93	0.96	**0.96**	0.97	0.93	0.94	0.96	0.98	0.95	0.96	0.98	**0.97**	0.98
	F1-Score	0.91	0.9	0.92	0.93	0.92	0.94	0.95	**0.94**	0.96	0.94	0.95	0.94	0.97	0.96	0.94	0.96	**0.97**	0.96
300	Precision	0.93	0.91	0.89	0.93	0.95	0.88	0.92	**0.93**	0.91	0.93	0.93	0.92	0.94	0.93	0.92	0.97	**0.95**	0.96
	Recall	0.92	0.92	0.90	0.96	0.96	0.93	0.96	**0.97**	0.92	0.96	0.9	0.89	0.95	0.94	0.89	0.96	**0.93**	0.98
	F1-Score	0.92	0.91	0.89	0.94	0.95	0.90	0.94	**0.95**	0.91	0.94	0.91	0.90	0.94	0.93	0.90	0.96	**0.94**	0.97
400	Precision	0.89	0.93	0.95	0.95	0.91	0.96	0.92	**0.93**	0.92	0.92	0.91	0.89	0.92	0.92	0.94	0.98	**0.95**	0.91
	Recall	0.92	0.89	0.94	0.92	0.92	0.91	0.96	**0.96**	0.89	0.96	0.93	0.93	0.93	0.94	0.97	0.94	**0.99**	0.93
	F1-Score	0.9	0.91	0.94	0.93	0.91	0.93	0.94	**0.94**	0.90	0.94	0.92	0.91	0.92	0.93	0.95	0.96	**0.97**	0.92

The confusion matrix, which shows the different performances of the classifiers for predicting the various emotions for KUI speech, is a table frequently used to characterize the performance of a classification model. The predicted labels are shown on the x-axis of the confusion matrix, while the actual labels are shown on the y-axis. Figures 10, 12, and 14 show the confusion matrix for CNN, LSTM, and HYBRID models without data augmentation. Overall, disgust was well-predicted, with a value of 90 by all classifiers.

Our proposed HYBRID model is the sole classifier with predictive ability for emotions, surprise, and anger, having 81 and 73, respectively, while another classifier failed to predict it. In contrast, fear is poorly predicted by all of our models compared to other emotions. A potential reason might be the minimal number of records in the dataset used to train classifiers to predict this emotion. CNN and LSTM are better at recognizing fear based on our proposed model. Our suggested model produced reasonable prediction rates for the emotions of surprise and anger. It can be concluded that our proposed HYBRID model scored better at predicting emotions than the other two classifiers in the case of without data augmentation.

Using the data augmentation methods, we enhance our datasets. We also generated the confusion matrix using data augmentation. As shown in Fig. 11, 13, and 15, the confusion matrix of CNN, LSTM, and HYBRID models, respectively. All classifiers correctly identified sad emotions, as indicated by their excellent prediction score. Except for the angry emotion, all predicted values are more in the case of our proposed HYBRID model. The prediction of sad is equal for both LSTM and our proposed model. From the above, it can be seen that our suggested HYBRID model outperformed the other two classifiers regarding emotion recognition. The accuracy of several deep learning models in different languages is shown in Table 5. Our proposed model outperformed better as we used the low-resourced tribal dataset.

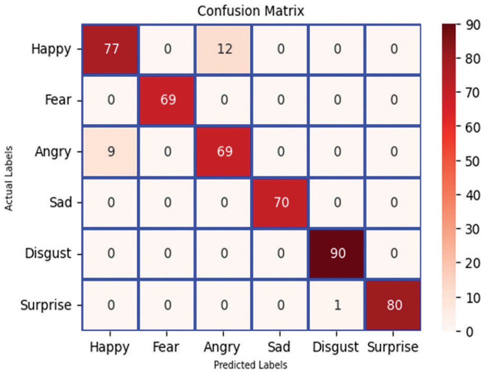

Fig. 10. Confusion matrix of CNN without augmentation

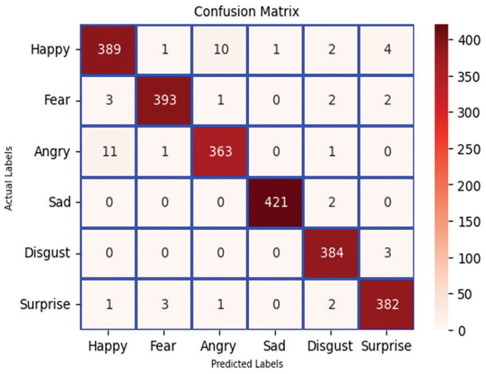

Fig. 11. Confusion matrix of CNN with augmentation

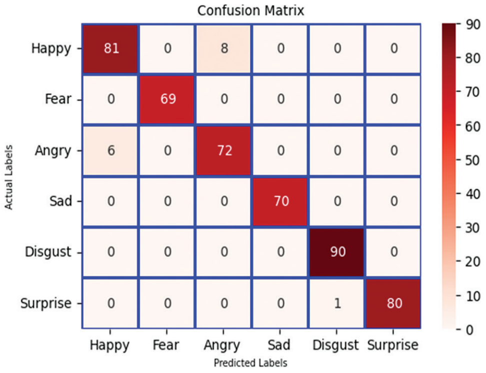

Fig. 12. Confusion matrix of LSTM without augmentation

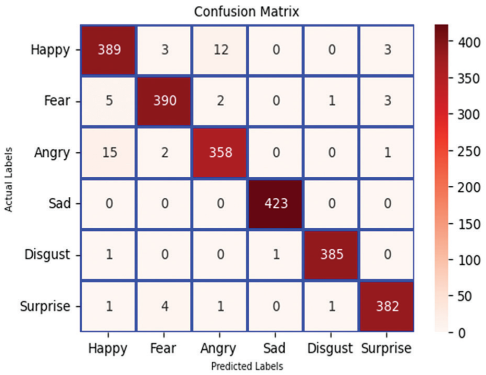

Fig. 13. Confusion matrix of LSTM with augmentation

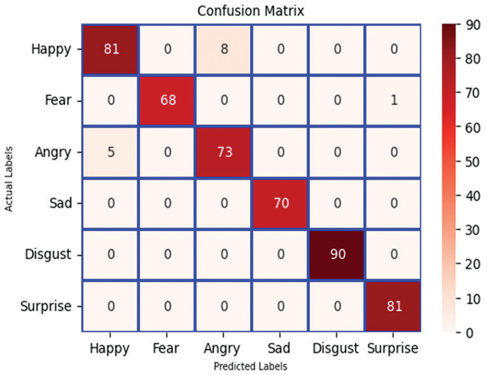

Fig. 14. Confusion matrix of HYBRID without augmentation

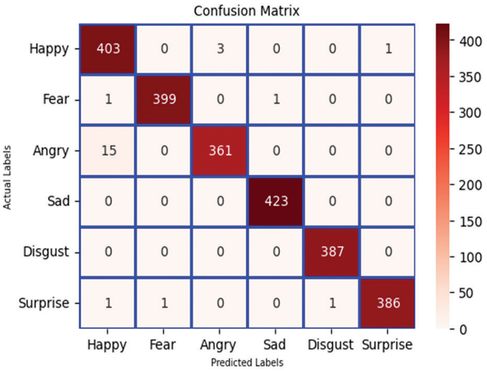

Fig. 15. Confusion matrix of HYBRID with augmentation

Table 5. Speech emotion recognition of different models for different languages

Work	Methods	Accuracy
Jo et al. [7]	Bi-LSTM	94.91
Taj et al. [8]	CNN	97.00
Atila et al. [10]	3D-CNN	96.00
Mohan et al. [11]	CNN-LSTM	70.56
Itponjaroen et al. [12]	LSTM	89.72
Alluhaidan et al. [13]	CNN	96.60
Suprava et al. [15]	D D LSTM	98.50
Dal Rì et al. [18]	CNN	100.00
Laghari et al. [21]	1D-CNN	91.00
Issa et al. [22]	CNN	86.10
Sajjad et al. [23]	BiLSTM	85.57
Hifny et al. [24]	CNN-LSTM-DNN	87.20
Huang et al. [27]	LSTM	82.00
Proposed Method	CNN+LSTM	97.00

7. CONCLUSIONS

In recent years, identifying emotions in speech has emerged as a prominent study area. The systems can improve direct communication with the machines. Several factors may affect these types of systems, such as diverged emotions, datasets, feature extraction methods, data preprocessing, and classifiers.

Developing a deep learning approach for KUI speech emotion recognition was the primary objective of this work. The CNN, LSTM, and HYBRID model evaluations are presented in experiments. The hybrid model achieved the best accuracy, with rates ranging from 93% to 97%. This study highlights insufficient research on categorizing emotions in a tribal language. The outcome also demonstrated that the best emotion to be predicted by all classifiers is sadness. Furthermore, a statistical analysis shows the reliability of the suggested approach's capability to recognize KUI emotions.

KUI is a low-resourced language, making dataset preparation a significant challenge. To enhance the language-independent KUI emotion recognition capability, more datasets in different languages may be added. We want to extend our methodology by including a broader range of emotion recognition algorithms to increase our emotion recognition system's robustness and accuracy.

8. REFERENCES:

[1] M. Hamidi, F. Barkani, O. Zealouk, H. Satori, "Assessing the Performance of a Speech Recognition System Embedded in Low-Cost Devices", International Journal of Electrical and Computer Engineering Systems, Vol. 14, No. 6, 2023, pp. 677-683.

[2] S. K. Nayak, A. K. Nayak, S. Mishra, P. Mohanty, N. Tripathy, S. Prusty, "Improving Kui digit recognition through machine learning and data augmentation techniques", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 35, No. 2, 2024, pp. 867-877.

[3] M. Hamidi, H. Boulal, J. Barkani, M. Abarkan, "Amazigh Spoken Digit Recognition using a Deep Learning Approach based on MFCC", International Journal of Electrical and Computer Engineering Systems, Vol. 14, No. 7, 2023, pp. 791-798.

[4] N. Tripathy, S. Hota, D. Mishra, P. Satapathy, S. K. Nayak, "Empirical Forecasting Analysis of Bitcoin Prices: A Comparison of Machine Learning, Deep Learning, and Ensemble Learning Models", International Journal of Electrical and Computer Engineering Systems, Vol. 15, No. 1, 2024, pp. 21-29.

[5] Y. B. Singh, S. Goel, "A systematic literature review of speech emotion recognition approaches", Neurocomputing, Vol. 492, 2022, pp. 245-263.

[6] S. K. Nayak, A. K. Nayak, S. Mishra, P. Mohanty, "Deep Learning Approaches for Speech Command Recognition in a Low Resource KUI Language", International Journal of Intelligent Systems and Applications in Engineering, Vol. 11, No. 2, 2023, pp. 377-386.

[7] A. H. Jo, K. C. Kwak, "Speech emotion recognition based on a two-stream deep learning model using Korean audio information", Applied Sciences, Vol. 13, No. 4, 2023, p. 2167.

[8] S. Taj, G. M. Shaikh, S. Hassan, "Urdu Speech Emotion Recognition using Speech Spectral Features and Deep Learning Techniques", Proceedings of the 4th International Conference on Computing, Mathematics and Engineering Technologies, Sukkur, Pakistan, 17-18 March 2023, pp. 1-6.

[9] A. Amjad, L. Khan, N. Ashraf, M. B. Mahmood, H. T. Chang, "Recognizing semi-natural and spontaneous speech emotions using deep neural networks", IEEE Access, Vol. 10, 2022, pp. 37149-37163.

[10] O. Atila, A. Şengür, "Attention guided 3D CNN-LSTM model for accurate speech-based emotion recognition", Applied Acoustics, Vol. 182, 2021, p. 108260.

[11] M. Mohan, P. Dhanalakshmi, R. S. Kumar, "Speech emotion classification using ensemble models with MFCC", Procedia Computer Science, Vol. 218, 2023, pp. 1857-1868.

[12] N. Itponjaroen, K. Apsornpasakorn, E. Pimthai, K. Kaewkaisorn, S. Panitchart, T. Siriborvornratanakul, "Speech emotion classification using semi-supervised LSTM", Advances in Computational Intelligence, Vol. 3, No. 4, 2023, p. 12.

[13] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, O. S. Neffati, "Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network", Applied Sciences, Vol. 13, No. 8, 2024, p. 4750.

[14] M. R. Falahzadeh, F. Farokhi, A. Harimi, R. Sabbaghi-Nadooshan, "Deep convolutional neural

network and gray wolf optimization algorithm for speech emotion recognition", Systems, and Signal Processing, Vol. 42, No. 1, 2023, pp. 449-492.

[15] S. Patnaik, "Speech emotion recognition by using complex MFCC and deep sequential model", Multimedia Tools and Applications, Vol. 82, No. 8, 2023, pp. 11897-11922.

[16] P. Singh, M. Sahidullah, G. Saha, "Modulation spectral features for speech emotion recognition using deep neural networks", Speech Communication, Vol. 146, 2023, pp. 53-69.

[17] K. Bhangale, M. Kothandaraman, "Speech emotion recognition based on multiple acoustic features and deep convolutional neural network", Electronics, Vol. 12, No. 4, 2023, p. 839.

[18] F. A. Dal Rì, F. C. Ciardi, N. Conci, "Speech Emotion Recognition and Deep Learning: an Extensive Validation using Convolutional Neural Networks", IEEE Access, Vol. 11, 2023, pp. 116638-116649.

[19] L. Yunxiang, Z. Kexin, "Design of Efficient Speech Emotion Recognition Based on Multi-Task Learning", IEEE Access, Vol. 11, 2023, pp. 5528-5537.

[20] S. Kakuba, D. S. Han, "Speech Emotion Recognition using Context-Aware Dilated Convolution Network", Proceeding of the 27th Asia Pacific Conference on Communications, Jevu Island, Korea, 19-20 October 2022, pp. 601-604.

[21] M. Laghari, M. J. Tahir, A. Azeem, W. Riaz, Y. Zhou, "Robust speech emotion recognition for Sindhi language based on deep convolutional neural network", Proceeding of the 3rd International Conference on Communications, Information System and Computer Engineering, Beijing, China, 14-16 May 2021, pp. 543-548.

[22] D. Issa, M. F. Demirci, A. Yazici, "Speech emotion recognition with deep convolutional neural networks", Biomedical Signal Processing and Control, Vol. 59, 2020, p. 101894.

[23] M. Sajjad, S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM", IEEE Access, Vol. 8, 2020, pp. 79861-79875.

[24] Y. Hifny, A. Ali, "Efficient arabic emotion recognition using deep neural networks", Proceeding of the International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12-17 May 2019, pp. 6710-6714.

[25] I. Shahin, A. B. Nassif, S. Hamsa, "Emotion recognition using hybrid Gaussian mixture model and deep neural network", IEEE Access, Vol. 7, 2019, pp. 26777-26787.

[26] M. F. Alghifari, T. S. Gunawan, M. Kartiwi, "Speech emotion recognition using deep feedforward neural network", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 10, No. 2, 2018, pp. 554-561.

[27] K. Y. Huang, C. H. Wu, Q. B. Hong, M. H. Su, Y. R. Zeng, "Speech emotion recognition using a convolutional neural network with audio word-based embedding", Proceeding of the 11th International Symposium on Chinese Spoken Language Processing, Taipei, Taiwan, 26-29 November 2018, pp. 265-269.

[28] S. Renjith, K. G. Manju, "Speech-based emotion recognition in Tamil and Telugu using LPCC and hurst parameters", Proceeding of the International Conference on Circuit, Power and Computing Technologies, Kollam, India, 20-21 April 2017, pp. 1-6.

[29] J. Rybka, A. Janicki, "Comparison of speaker dependent and speaker independent emotion recognition", International Journal of Applied Mathematics, Vol. 23, No. 4, 2013, pp. 797-808.

[30] S. Mohanty, B. K. Swain, "Emotion recognition using fuzzy K-means from Oriya speech", International Journal of Computer and Communication Technology, Vol. 2, No. 1, 2010, pp. 24-28.

[31] S. K. Nayak, A. K. Nayak, S. Mishra, P. Mohanty, A. Pati, A. Panigrahi, "Speech data collection system for KUI, a Low resourced tribal language", Journal of Autonomous Intelligence, Vol. 7, No. 1, 2024, p. 1121.

[32] S. Khamlich, F. Khamlich, I. Atouf, M. Benrabh, "Performance evaluation and implementations of MFCC, SVM and MLP algorithms in the FPGA board", International Journal of Electrical and Computer Engineering Systems, Vol. 12, No. 3, 2021, pp. 139-153.

[33] Z. Tu, B. Liu, W. Zhao, R. Yan, Y. Zou, “A Feature Fusion Model with Data Augmentation for Speech Emotion Recognition”, Applied Sciences, Vol. 13, No. 7, 2023, p. 4124.

[34] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe, S. Misra, "Data augmentation and deep learning methods in sound classification: A systematic review", Electronics, Vol. 11, No. 22, 2022, p. 3795.

[35] Z. Yang, Y. Huang, "Algorithm for speech emotion recognition classification based on Mel-frequency cepstral coefficients and broad learning system", Evolutionary Intelligence, Vol. 15, No. 4, 2022, pp. 2485-2494.

[36] C. Hema, F. P. G. Marquez, “Emotional speech recognition using CNN and deep learning techniques”, Applied Acoustics, Vol. 211, 2023, p. 109492.

[37] A. A. Anthony, C. M. Patil, “Speech emotion recognition systems: A comprehensive review on different methodologies”, Wireless Personal Communications, Vol. 130, No. 1, 2023, pp. 515-525.

[38] M. D. Pawar, R. D. Kokate, "Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients", Multimedia Tools and Applications, Vol. 80, 2021, pp. 15563-15587.