2nd International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM 2025)
27-28 June 2025, Gazipur-1707, Bangladesh

2025 2nd International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM) | 979-8-3315-5542-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/NCIM65934.2025.11160135

Speech Emotion Recognition using Branched-based Convolutional Neural Network Technique

Md. Enamul Haque
Institute of Energy Technology
Chittagong University *of Engineering and Technology*
Chittagong, Bangladesh
mdenamulhaquem@gmail.com

Md. Masud Karim
Dept of Electrical and Electronic Engg.
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
masudkarim521@gmail.com

Syed Munimus Salam
Dept of Electrical and Electronic Engg.
Chittagong University of Engineering and Technology
Chittagong, Bangladesh
munim53@gmail.com

Md. A. Hamid Howlader
Dept of Electrical and Electronic Engg.
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
hamid.eee06@gmail.com

Hasanur Zaman Anonto
Dept of Electrical and Electronic Engg.
American International University-Bangladesh
Dhaka, Bangladesh
hasanur.eee.reged@gmail.com

Abu Shufian
Dept of Electrical and Electronic Engg.
American International University-Bangladesh
Dhaka, Bangladesh
shufian.eee.reged@gmail.com

***Abstract*— Speech Emotion Recognition is a significant pattern recognition of human speech using feature extraction for communication media. This paper aims to recognize speech emotion through the CNN branch-based model. Machine learning strategy for reliable speech emotion identification in the face of limited data and cultural differences. This study proposed to analyze three different datasets and observe the most useable emotion accuracy using the CNN model for the best outcome accuracy performance. We highlight the speech operations of RAVDESS, TESS, and SAVEE Datasets using deep learning libraries like Keras- Dense, Conv1D, MaxPooling1D, Flatten, Dropout, and Batch Normalization to build branches of convolutional neural network architectures. The result performance of the model Accuracy is 87.12%. Overall, emotional recognition shows potential for improving human-machine communication and providing insights into specific emotional states.**

Keywords*— Speech Emotion Recognition (SER), Convolutional Neural Network (CNN), Feature Extraction, Human-Computer Interaction.

I. INTRODUCTION

Speech Recognition is a dynamic area of study in Human-Computer Interaction (HCI) that examines speech signals to determine the speaker's emotional state. In human communication, emotions are vital as they can convey important information that transcends the literal meaning of words. By incorporating Speech Emotion Recognition (SER) technology, machines can better understand user intent and respond appropriately, leading to more natural and effective human-computer interactions. Machine learning (ML) approaches have proven to be effective tools for SER due to their ability to understand complex patterns in voice data. This study explores the use of machine learning techniques for identifying emotions expressed through spoken language. It will investigate the feature extraction method, which involves identifying significant characteristics from speech signals, and various ML classification models used for SER tasks. In addition to advancing SER technology, this study aims to assess the effectiveness of different ML methods in distinguishing emotions. SER technology is essential for enhancing customer service, mental health monitoring, and human-computer interaction. The branch-based technique, which employs a convolutional neural network, has yielded promising results in accurately recognizing emotions from speech data, paving the way for future applications.

A background study on speech emotion recognition using CNN revealed the effectiveness of deep learning approaches in extracting complex characteristics from audio inputs. These experiments have paved the way for future advancements in emotion recognition technology, suggesting the potential for greater accuracy and real-world applications. Developing techniques that extract features invariant to extraneous factors while maintaining a high level of discriminative ability for emotion categorization is crucial for SER. Recent advances in deep learning, particularly CNNs, have shown effectiveness in creating hierarchical feature representations, which is especially valuable in situations with minimal labeled data. However, past SER research has largely concentrated on extracting discriminative features from inputs, with insufficient attention given to disentangling variability aspects [1]. Speech recognition methods utilize audio data sequences and neural networks. A system that employs a convolutional subnetwork with multiple dilated convolutional neural network layers provides alternative representations to enhance voice recognition by analyzing each input alongside its preceding inputs [2]. The study [3] utilized spectrogram images to train CNN architectures for speech-emotion recognition. Spectrograms are a visual representation of signal intensity over time at various frequencies, offering a novel approach to analyzing speech data. Research indicates that the spectrogram-based model outperforms the MFCC-based model [4], achieving an accuracy of 82.5%, which is 3.75% higher than the MFCC

model when compared to CNN models. The SER architecture [5], DCTFB (Deep CNN with Time-Distributed Flatten and BLSTM Layers), consists of a Deep Convolutional Neural Network (DCNN) with four convolutional blocks, each featuring two-dimensional convolutional layers, batch normalization, ELU activations, and max-pooling. A time-distributed flattening (TDF) layer follows, applying the same weights and biases to each temporal step to facilitate sequential analysis.

A bidirectional long short-term memory (BLSTM) layer captures past and future context and sends its output to a fully connected layer with softmax activation for emotion categorization. Additionally, the study [6] employs a three-channel feature fusion strategy for identifying speech emotions, utilizing BiLSTM networks and CNNs. This strategy proves more effective than single-feature methods, as demonstrated by the CASIA and EMO-DB corpora. A dynamic SER system with CNNs and BiLSTM integration retrieves global dynamic emotion information more efficiently, thereby improving performance; however, enhancing multimodal data for recognition systems remains a challenge [7]. The study [8] investigated Long Short-Term Memory (LSTM) and two convolutional neural networks (CNN LSTM) with 1D and 2D architectures. Both networks shared an LSTM layer and four Local Feature Learning Blocks (LFBs) for extracting speech emotions, including happiness, surprise, disgust, neutrality, sorrow, fear, and rage. The model performed well across various datasets and techniques.

The author [9] highlighted a method for recognizing emotions in spontaneous speech divided into prosodic phrases, nonverbal pauses, and silences. By utilizing CNN-generated feature vectors and an LSTM-based sequence-to-sequence model, the system enhances traditional techniques in emotion recognition. The study [10] compares CNN-LSTM with CNN-Transformer for improving emotion categorization accuracy in voice recognition. An accuracy of 82% is achieved with CNN-Transformer, compared to 74% with CNN-LSTM. A hybrid deep convolutional neural network (CNN) with a 1D branch for extracting raw audio features and a 2D branch for log-mel spectrogram analysis benefits from spatially local connections and shared weights, facilitating efficient filter learning with less pre-processing compared to other deep learning architectures [11-12]. The author [13] presents a Convolutional Recurrent Neural Network (CRNN) that combines LSTMs with CNNs. It emphasizes the effectiveness of linearly spaced spectrograms as input features. By comparing the CRNN's performance to human perceptual evaluation, the study showcases strong alignment, with accuracies of 84.55% at the file level and 77.51% at the segment level. Similar to BranchNet, recurrent neural networks can similarly forecast hard-to-predict branching [14].

However, we limit the scope of this research, which explores the use of CNNs for branch prediction, underscoring the potential for low-latency and storage-efficient predictors. The results demonstrate the CRNN's robustness and its potential to effectively identify emotions, highlighting its valuable applications. Diverse CNN architectures have shown varying degrees of effectiveness and efficiency in identifying emotional characteristics in voice signals. Understanding the advantages and limitations of each design can assist in creating more reliable and accurate emotion identification systems. The literature examines recent developments in speech-emotion recognition, emphasizing advantages as well as challenges. To effectively address real-world scenarios, SER classification systems must prioritize overall accuracy in multi-class challenges.

The rest of the research paper is arranged as follows. The methodology, the researched model architecture, SER benchmark datasets, and evaluation are presented in Section 2. Results and Discussion are covered in Section 3. Section 4 discusses the conclusion and directions for the future.

II. Methodology

Speech emotion detection has become essential for efficient communication in a variety of industries. Using machine learning approaches allows for smooth access to human-machine interfaces. In machine learning approaches, we employ supervised techniques for speech recognition, such as convolutional neural networks (CNN).

A. Architecture of the Proposed Model

A feedforward CNN network consists of layers that serve certain purposes, such as convolutional, normalization, activation, and down-sampling. A CNN's structure includes an input layer for multidimensional tensor data, hidden layers for feature extraction, and an output layer for classification tasks, which is commonly a fully connected layer. The CNN architecture is shown in Fig. 1.

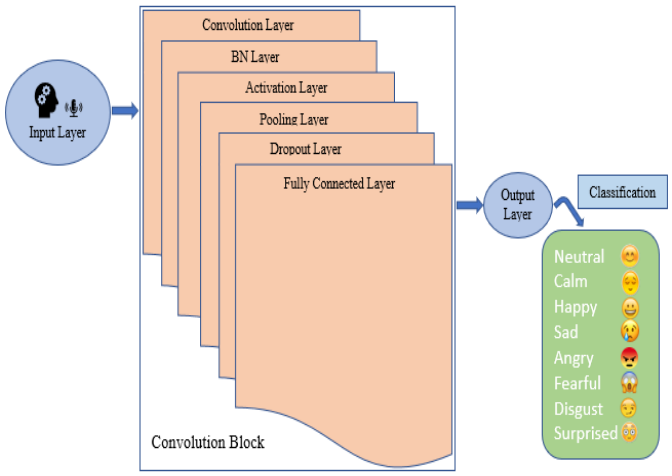

Fig. 1. Proposed Model of SER Framework.

B. Data collection and preprocessing techniques:

In our research, firstly we use a Python library for audio and music analysis called librosa. It can extract info from audio files. The files we collected from different datasets are given a short description below:

The Ryerson Audiovisual Database of Emotional Speech and Songs (RAVDESS) [15] is developing emotion recognition models for vital tools. This British English dataset contains 1,440 high-quality utterances expressing eight emotions, spoken by 24 professional actors (12 male and 12 female). Each recording, captured at 48 kHz, lasts around 3 seconds. RAVDESS's thorough annotations and thoughtful portrayal of

emotions make it a popular choice for both practical and scholarly purposes.

Toronto Emotional Speech Set (TESS) [16] is a prominent English-language dataset used in studies on emotion identification. It has 2,800 recordings of two native English-speaking females, ages 26 and 64, expressing seven different emotions: anger, contempt, fear, pleasure, pleasant surprise, sorrow, and neutral. Each recording, captured at 22,050 Hz, has an average duration of 2.06 seconds. The extensive emotional expressions and superior recordings of TESS make it a popular tool for both practical and scholarly uses.

Surrey Audio-Visual Expressed Emotion (SAVEE) [17] dataset, fifteen native speakers, ten male and five female, express seven different emotions through 1,200 English utterances. Each speaker provided 15 utterances, precisely tagged with speaker IDs and emotions. SAVEE is a valuable resource for speech emotion recognition research.

After collecting the dataset, we stacked the datasets both horizontally and vertically for the augmentation process. We applied noise, stretching, shifting, and pitch techniques to experiment with the data. Then, we defined the extracted features, such as zero crossing rate, Chroma stft, MFCC, Root Mean Square Value, and Mel Spectrogram.

C. *Description of branch-based CNN model architecture*

A branch-based Convolutional Neural Network for SER employs multiple parallel branches to analyze different varieties of information. The range includes temporal, spectral, and prosodic. Branches independently extract complementary information, which is then combined, fed into fully linked layers, and categorized. This approach captures a broader collection of features than a single stream CNN, which improves accuracy and resilience. The architecture is flexible, allowing the incorporation of various kinds of speech information and attention mechanisms, allowing the dynamic prioritization of the focus of control characteristics. Hence, it efficiently handles the complexity of emotional speech.

D. *Finding Total Parameters*

The summary approach yields the data shown in Table I. Each row represents a layer; therefore, we may refer to them just by their row names to avoid misunderstandings. It categorizes the layers into four logical groups, Conv1D + Pooling, BatchNorm + Activation, Add + Flatten, and Dense + Dropout—and highlights their respective output shapes, parameter counts, and functional roles. This abstraction simplifies understanding of the overall model structure by focusing on the major computational and functional stages, from feature extraction to classification.

TABLE I. MODEL SPECIFICATION

Layer Type	Output Shape	Parameters Count	Remarks
Conv1D + Pooling	(None, 20, 128)	94,576	Sequential convolution with pooling
BatchNorm + Activation	(None, 20, 128)	1,536	Used after conv layers for stability
Add + Flatten	(None, 1280)	0	Skip connection & feature flattening
Dense + Dropout	(None, 8)	922,632	Final classification layers

E. *Training and Testing:*

After feeding the dataset into the Kaggle platform, we observed the training and testing loss and accuracy graphs of our model represents in Fig. 2. Fig. 3 illustrate the training performance of the proposed system. These figures show the training and testing accuracies, along with their respective losses, for the benchmark dataset.

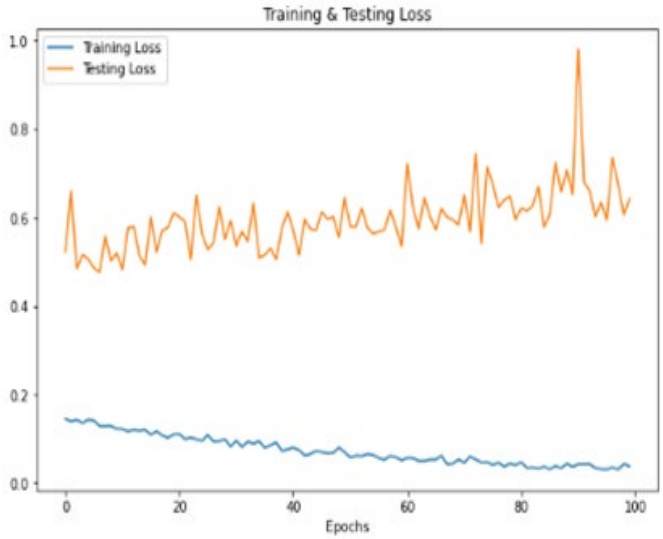

Fig. 2. Training and Testing loss graph.

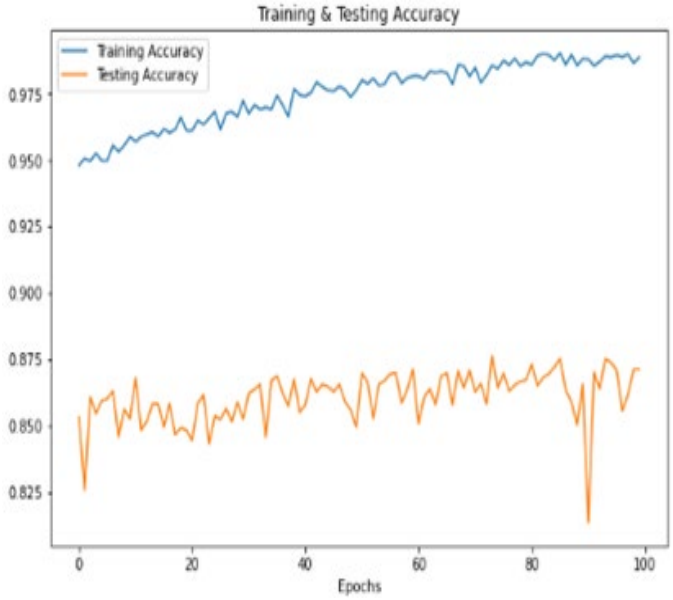

Fig. 3. Training and Testing accuracy graph.

F. Performance Evaluation

1) Precision:

Precision refers to the percentage of test images correctly predicted by the CNN as belonging to a specific class.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

2) Recall:

Recall is the percentage of test images that belong to a certain class and that the CNN predicts are from that class.

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

3) Accuracy:

The percentage of test images that CNN properly identifies is known as accuracy.

$$Accuracy = \frac{Number\ of\ Correct\ prediction}{Total\ number\ of\ prediction} \quad (3)$$

4) F1-Score:

F1-score is a widely recognized metric for evaluating the effectiveness of classification models, especially in scenarios with uneven class distributions. It represents the harmonic mean of precision and recall, achieving a balance between the two. The F1-Score formula is as follows:

$$F1 = \frac{2\times(\text{Precision}\times\text{Recall})}{\text{Precision}+\text{Recall}} \quad (4)$$

III. Results and Discussion

We observe the results of Various datasets and combine the results for maximum emotional accuracy on Kaggle performance. The taking time of our model is 0s 3ms/step, loss: 0.6423 and the Accuracy of our model on test data is 87.118 %.

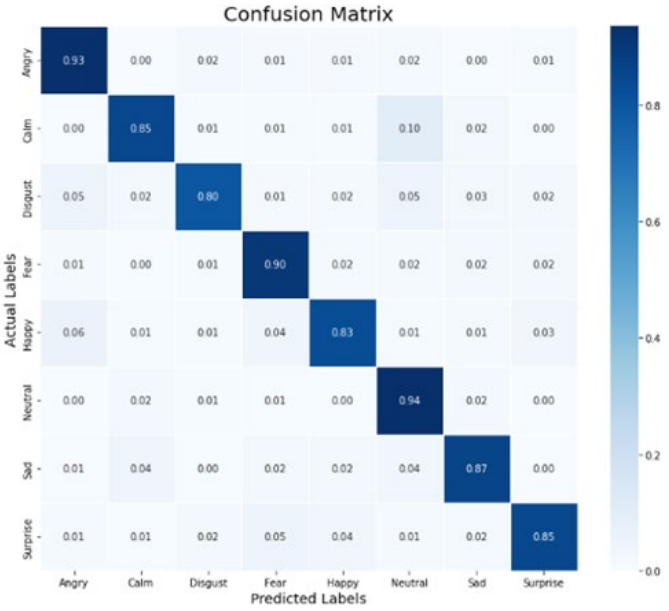

Fig. 4. Confusion Matrix of Proposed Model.

A. Confusion Matrix

The model's results are reflected in the confusion matrix shown in Fig. 4. Table II describes the model performance, which presents the classification performance metrics for a model evaluated on a test set of 3,540 samples. The accuracy of the model is 87%, indicating the overall proportion of correct predictions. The macro average yields a precision of 0.86, recall of 0.87, and F1-score of 0.86, treating all classes equally. Meanwhile, the weighted average (which accounts for class imbalance) shows consistent values of 0.87 across precision, recall, and F1-score, reflecting balanced performance across different classes.

TABLE II. Report For Emotion Recognition Classification.

	precision	***recall***	***F1-score***	***support***
accuracy	-	-	0.87	3540
macro avg	0.86	0.87	0.86	3540
weighted avg	0.87	0.87	0.87	3540

B. Comparison with existing approaches

Convolutional neural network processes of learning are the most known and effective when compared to previous model assessments for speech emotion recognition. Table III presents the previous stare of art SER. Based on this comparison with various models using different combination datasets, we find that branch-based CNN performance is superior. It also presents a comparative analysis of different models and methodologies used for emotion recognition. It lists three approaches: Ref [18] used CNN-BLSTM on the RAVDESS + IEMOCAP datasets with Spectrograms, achieving an accuracy of 84.55%. Ref [19] applied a CNN model on IEMOCAP + EMO-DB using MFCC + Chromagram + Spectrogram, with an accuracy of 71.61%. The proposed method employs a CNN model using RAVDESS + TESS + SAVEE datasets and combines MFCC + MelSpectrogram features, achieving the highest accuracy of 87.118%.

TABLE III. Comparition of Emotion Recognition Classification.

Ref	**Datasets**	**Model**	**Method**	**Accuracy %**
[18]	RAVDESS + IEMOCAP	CNN-BLSTM	Spectrograms	84.55
[19]	IEMOCAP + EMO-DB	CNN	MFCC + Chromagram + Spectrogram	71.61
Proposed	**RAVDESS +TESS +SAVEE**	**CNN**	**MFCC+ MelSpectogram**	**87.118**

C. Challenges and Future directions for improving research

The study focused on the need for tuning hyperparameters and balancing accuracy and computing economy in branch-based speech emotion recognition using CNNs. The integration of data from several modalities to improve system performance should be the subject of future research on fusion techniques. Overall, this study advances the further research and improvement of CNN-based branch-based speech emotion

identification in speech analysis. The model's hyperparameters had to be optimized, and striking a balance between accuracy and computing efficiency presented difficulties during the study. During data preparation, we utilize a one-hot encoding algorithm to address the challenges of multiclass classification effectively. To enhance the effectiveness of the emotion detection system, future research could focus on exploring different fusion algorithms for integrating data from multiple modalities.

IV. Conclusion

In this paper, we introduced a framework for speech emotion recognition, focusing on using branch-based convolutional neural networks (CNNs). It emphasizes improvements in feature data extraction and processing from speech signals to improve the differentiation between emotional voices. These enhancements attempt to boost CNN's expressive capabilities in recognizing uttered target emotion classes. Also, we observe three different datasets and give the best output result for summaries of the best outcome result among them. The performance of this model accuracy is 87.12%. For this model, we analyzed all input data in a data frame path to extract features matching the model and classifier emotion, which has eight classes. Conclusively, the findings from the study validate our model's benefits over the conventional CNN.

References

[1] Huang, Z., Dong, M., Mao, Q. and Zhan, Y.,, November. Speech emotion recognition using CNN. In Proceedings of the 22nd ACM international conference on Multimedia , 2014, (pp. 801-804).

[2] van den Oord, A.G.A., Dieleman, S.E.L., Kalchbrenner, N.E., Simonyan, K., Vinyals, O. and Espeholt, L., DeepMind Technologies Ltd, 2020.

[3] Mukherjee, S., Mundra, S. and Mundra, A.. Speech Emotion Recognition Using Convolutional Neural Networks on Spectrograms and Mel-frequency Cepstral Coefficients Images. In Information and Communication Technology for Competitive Strategies (ICTCS 2022) Intelligent Strategies for ICT , 2023, (pp. 33-41).

[4] M. E. Haque, S. M. Salam and M. S. Islam, "Human Speech Emotion Recognition Using Artificial Neural Networks Technique," 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS), Dhaka, Bangladesh, 2024, pp. 1-6.

[5] Sultana, S., Iqbal, M.Z., Selim, M.R., Rashid, M.M. and Rahman, M.S.,. Bangla speech emotion recognition and cross-lingual study using deep CNN and BLSTM networks. IEEE Access, 10, 2021, pp.564-578.

[6] Huang, L., Dong, J., Zhou, D. and Zhang, Q., May. Speech emotion recognition based on three-channel feature fusion of CNN and BiLSTM. 4th International Conference on Innovation in Artificial Intelligence 2020, (pp. 52-58).

[7] Lin, Z., Hu, Z. and Zhu, K.,. Speech emotion recognition based on dynamic convolutional neural network. Journal of Computing and Electronic Information Management, 10(1), 2023, pp.72-77.

[8] Zhao, J., Mao, X. and Chen, L.. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical signal processing and control, 47, 2019, pp.312-323.

[9] Huang, K.Y., Wu, C.H., Hong, Q.B., Su, M.H. and Chen, Y.H., May. Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5866-5870). 2019, IEEE.

[10] Vijayan, D.M., Arun, A.V., Ganeshnath, R., SA, A.N. and Roy, R.C.. Development and Analysis of Convolutional Neural Network based Accurate Speech Emotion Recognition Models.19th India Council International Conference (INDICON) (pp. 1-6). 2022, IEEE.

[11] Zhao, J., Mao, X. and Chen, L.. Learning deep features to recognise speech emotion using merged deep CNN. IET Signal Processing, 2018, 12(6), pp.713-721.

[12] Larisa, P.M. and Tapu, R., November. Speech Emotion Recognition Using 1D/2D Convolutional Neural Networks. In 2022 International Symposium on Electronics and Telecommunications (ISETC) (pp. 1-4). 2022, IEEE.

[13] Nagajyothi, D. and Siddaiah, P,. Speech recognition using convolutional neural networks. Int. J. Eng. Technol, 7(4.6), 2018, pp.133-137.

[14] Zangeneh, S., Pruett, S., Lym, S. and Patt, Y.N., October. Branchnet: A convolutional neural network to predict hard-to-predict branches. 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO) (pp. 118-130). 2020, IEEE.

[15] Livingstone, S.R. and Russo, F.A.,. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 2018, 13(5), p.e0196391.

[16] Dupuis, K. and Pichora-Fuller, M.K.,. Toronto emotional speech set (TESS). University of Toronto, Psychology Department , 2010.

[17] Jackson, P. and Haq, S.. Surrey audio-visual expressed emotion (savee) database. University of Surrey: Guildford, UK. 2014

[18] Mustaqeem and Kwon, S.,. CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network. Mathematics, 2020, 8(12), p.2133.

[19] Kwon, S.,. MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. Expert Systems with Applications, 2021, 167, p.114177.