EchoMind: An AI-driven Emotional Recognition and Personalized Recommender System

Ms. Gayatri [1], Mr. Raghavendra [2], Ms. Sandhya[3], Mr. Tauseef Baig[4]
[1,2,3,4] *Department of Data Science, Raghu Engineering College, Visakhapatnam*

***Abstract*—The field of speech emotion recognition (SER) is extended to enhance the interaction between humans and computers by enabling machines to recognize and understand emotions in language. SER is used in various applications in fields like healthcare, virtual assistants, customer support, and security systems. The initial SER process depends on handcrafted properties and machine learning algorithms, which tend to perform poorly while handling variations of audio patterns and noise. That being said, the advent of deep learning transformed the construction of SER through automating feature extraction and enhancing reliability in emotional classification. Techniques like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) show outstanding performance in detecting spatial and temporal patterns of linguistic information. In addition, the hybrid model that integrates CNN and RNN enhances the accuracy of emotional classification even more. This study introduces an AI-powered emotion recognition and personalized recommendation system that utilizes deep learning for audio analysis. The system incorporates a cutting-edge SER model integrated with Ollama, a sophisticated extension that facilitates personalized recommendations and natural, two-way communication through audio input and output. The SER model detects emotions from user speech with high precision, while Ollama uses this emotional data to provide customized recommendations and engage in context-aware dialogues. This integration creates a seamless and interactive user experience, making the system highly effective for real-time applications.**

***Index Terms*—Speech Emotion Recognition, CNN, Deep Learning, Neural Networks, RNN, Feature Extraction, Human-Computer Interaction**

I. INTRODUCTION

The area of Speech Emotion Recognition (SER) has become an imperative research field in the area of artificial intelligence that seeks to narrow the gap between human emotions and computer comprehension. Emotions are crucial in human communication and they affect how we interact, make choices, and perceive the world. By allowing machines to identify and comprehend emotions through speech, SER can potentially transform human-computer interaction into being more intuitive, empathetic, and tailored. Hand - engineered features including pitch, tone, and spectral features were traditionally used in combination with machine learning techniques such as Support Vector Machines (SVMs) or Hidden Markov Models (HMMs) in SER systems. Although such approaches met some level of success, they failed to handle the difficulties of environmental noise, speaker variability, and the richness of emotional expression in speech. Such shortcomings made widespread deployment of SER in real applications challenging.

The machine learning methods like Support Vector Machines (SVMs) and Gaussian Mixture methods (GMMs), early efforts to SER depended on handcrafted feature extraction, classifying emotions using acoustic data like pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCCs). Despite their relative effectiveness, these approaches frequently have trouble with speech data variability because of variations in speakers, languages, and recording environments.

Emerging mental health concerns such as depression and anxiety require smart solutions. Existing systems do not have emotion-aware, real-time interventions. Echo Mind fills this gap by combining AI-based audio analysis to identify emotional states based on pitch and speech, providing tailored support and minimizing stigma, thus covering the gap in accessible mental healthcare.

The emergence of deep learning has brought about a substantial change in SER. By automatically learning representations from unprocessed voice data, deep learning models—in particular, Convolutional Neural

Networks (CNNs) and Recurrent Neural Networks (RNNs)—have removed the need for human feature engineering. While RNNs, particularly Long Short-Term Memory (LSTM) networks, capture temporal correlations in voice signals, CNNs are especially good at extracting spatial characteristics from spectrograms. Hybrid models that combine CNNs and RNNs have further increased
classification accuracy, increasing the scalability and dependability of SER systems. Even with these developments, SER still faces a number of difficulties. The availability of sizable, varied, and well-labelled datasets is a significant obstacle because deep learning models need a lot of data to be trained. Furthermore, spoken emotional emotions are quite subjective and can change depending on surrounding conditions, individual speaking styles, and cultural variances. SER performance is also impacted by noise interference in real-world applications, thus creating reliable models that can manage a range of acoustic circumstances is essential.

This paper provides a review of recent advancements in SER using deep learning. It discusses about the datasets used in SER research, feature extraction techniques, deep learning architectures, and their effectiveness in emotion classification. Additionally, this review highlights key challenges and potential future directions for improving SER systems. By leveraging deep learning, future SER applications can achieve higher accuracy, robustness, and adaptability, making them valuable tools for enhancing human-computer interaction in real-world scenarios.

II. LITERATURE REVIEW

This section elaborates about speech emotion recognition (SER) by various researchers.

Zhangir Khan.[1] have reviewed various AI techniques for audio emotion recognition, including SVM, RNN, and CNN highlighting advances in emotion recognition from audio, discussing applications and future directions.

Alu et al. [2] have states that “In order to obtain emotional-related response from robots, computers and other intelligent machines, the first and decisive step is accurate emotion recognition.” The architecture is an adaptation of an image processing CNN, programmed in Python using Keras model-level library and TensorFlow backend with an accuracy of 71.33%.

Bertero and Fung. [3] states “Our model is trained from raw audio on a small dataset of TED talks speech data, manually annotated into three emotion classes: 'Angry', 'Happy' and 'Sad'. It achieves an average accuracy of 66.1%, 5% higher than a feature-based SVM baseline, with an evaluation time of few hundred milliseconds.”

Jiang et li. [4] reviewed that “A fusion network is trained to jointly learn the discriminative acoustic feature representation and a Support Vector Machine (SVM) is used as the final classifier for recognition task.” the proposed architecture improved the recognition performance, achieving accuracy of 64% compared to existing state-of-the-art approaches.

Akçay et al. [5] presented a survey on SER that covers almost all the areas of SER, the emotion models, databases, features, supporting modalities and classifiers. In their paper, they discussed deep-learning classifiers and deep-learning-based enhancement techniques such as auto-encoder, multi-tasking, adversarial training, attention to detail.

Swain et al. [6] reviewed the databases, features, and classifier techniques for SER. They categorized features as prosody, excitation, vocal-tract, and a fusion of one or more of these features for emotion recognition. Their paper also highlighted the usefulness of deep learning, hybrid, and fusion techniques for emotion classification.

III. MATERIALS AND METHODS

a. Dataset Collection

The dataset used in this study is the Multimodal Emotion Lines Dataset (MELD), which is gathered from GitHub. It consists of 1,400 dialogues and 13,000 utterances sourced from the TV series Friends. Each utterance is assigned one of the seven emotions, such as anger, disgust, sadness, joy, neutral, surprise, and fear. The dataset contains audio, text, and videos, enabling multi-dimensional analysis. The audio files are preprocessed to ensure consistency, and only the speech signals are used for emotion classification.

b. Model Architecture

The model integrates a CNN algorithm for recognizing emotion from the audio input, extracting features like MFCCs and pitch. It combines collaborative and content-based filtering for personalized

recommendations, enhanced by Ollama for natural communication. Real-time processing ensures seamless user interaction with low-latency audio analysis and response generation.

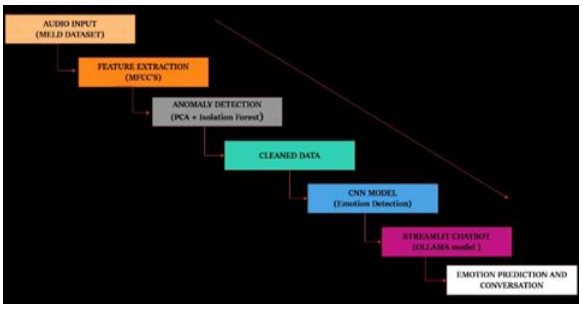

Fig 1. Workflow from input to the output

User Interaction: The user interacts with the system via speech, either by speaking directly into a microphone or uploading an audio file. The system processes the audio to transcribe the speech into text.
Emotion Detection: The emotion recognition module analyzes the transcribed speech and detects the user's emotional state based on deep learning models.
Personalized-Recommendation Generation: Based on the detected emotion, the model generates personalized recommendations. This can include relaxation exercises, motivational content, or self-help activities.
Response via Chatbot: The system generates context-sensitive responses through a chatbot by detected emotion, providing empathetic support
Personalized Recommendations: Simultaneously, the Recommender System suggests personalized actions or content, such as relaxing music, guided exercises, or motivational quotes.

Text-to-Speech: The generated response and recommendations are converted into speech by the Text-to-Speech Engine, which is then delivered to the user.

c. Preprocessing Steps

All audio files are converted to a uniform sampling rate of 16 kHz to maintain consistency across the dataset. Noise reduction techniques are applied to remove background disturbances and enhance the clarity in the speech. Additionally, audio amplitude normalization is performed to standardize volume levels across different samples.

Feature extraction is carried out using Mel-Frequency Cepstral Coefficients (MFCCs), which are widely used in speech analysis for capturing frequency-related features. For deep learning-based feature extraction, emotional cues such as pitch variations, speech intensity, and rhythm are analyzed to distinguish between different emotions.
After preprocessing, the dataset is divided into training (80%) and testing (20%) subsets, ensuring a balanced distribution of emotions across both sets. A small portion (10%) of the training data is set aside for validation to monitor model performance during training.

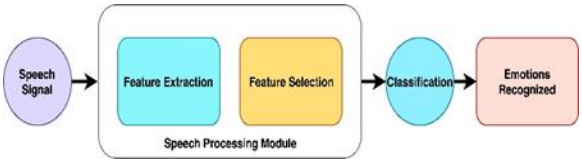

Fig 2. Workflow from speech signal to emotion recognition

d. Feature Extraction Using Deep Learning Models

To efficiently detect emotions in speech, the feature extraction is carried out using deep learning-based models via CNNs, which process the audio spectrogram to identify patterns that relate to certain emotions. RNNs, in particular LSTM networks, are used to observe sequential dependencies of speech. Pre-trained models like VGG16, ResNet, and DenseNet are used in the architecture for improved feature extraction. These models enable the system to learn complex speech features, resulting in better emotion classification accuracy.

e. Model Training and Evaluation

Data augmentation is utilized to artificially increase the dataset and enhance model robustness. Methods like time stretching and background noise addition introduce variations in speech recordings, assisting the model to generalize to real-world environments. The augmentation methods ensure that the model performs optimally even in noisy conditions.
Feature normalization is utilized to normalize the extracted MFCC features, ensuring uniform input for all training samples. This step reduces the impact of variations in audio intensity and duration. The model architecture is formulated to capture spatial and temporal dependencies in speech data. Conv1D layers extract time-series features from speech signals, while LSTM layers examine long-term dependencies to identify patterns in emotional speech. ReLU activation

functions add non-linearity, enabling the model to learn intricate representations, while the softmax activation function in the output layer classifies the input into one of the seven emotion classes.

To avoid overfitting, dropout layers are used to randomly shut down neurons during training, ensuring improved generalization. Adam optimizer is employed with a learning rate of 0.0001 to fine-tune model parameters efficiently. Early stopping is used to track validation loss and stop training once performance stabilizes. The system is optimized for real-time emotion recognition, enabling immediate classification and response generation. Lightweight CNN architectures enable fast inference, while LSTM networks monitor emotional transitions over time.

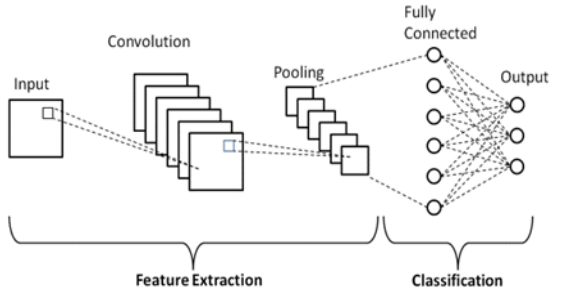

Fig 3. Workflow of the convolutional neural network

f. Integration with the AI

The emotion recognition module analyzes speech to detect emotions, which are then processed by Ollama. Ollama generates personalized recommendations and enables natural, context-aware communication, providing real-time, emotion-driven responses and tailored suggestions to enhance user interaction and satisfaction.

The model is embedded in a Streamlit-based user interface, allowing users to interact with the system and receive immediate feedback on their emotional state. The model's performance is assessed using various metrics to measure classification effectiveness. Accuracy is measured to establish the overall correctness of predictions, while precision, recall, and F1-score are computed to analyze how well the model classifies each individual emotion.

IV. RESULTS AND DISCUSSION

The proposed model achieves an overall accuracy of 78%. The classification report provides detailed insights into the model's performance across different classes, showing variations in precision, recall, and F1-score. The ROC curves further validate the model's ability to distinguish between different classes, although some classes exhibit lower predictive performance.

4.1 *Performance evaluation metrics*

Confusion Matrix: The confusion matrix highlights classification strengths and weaknesses. High precision and recall values are observed for class 1 (0.90, 0.96) and class 2 (0.86, 0.97), while class 6 shows significant misclassification, with a recall of only 0.16, indicating difficulties in identifying this category. This suggests that additional training data or feature enhancements could improve class 6 recognition.

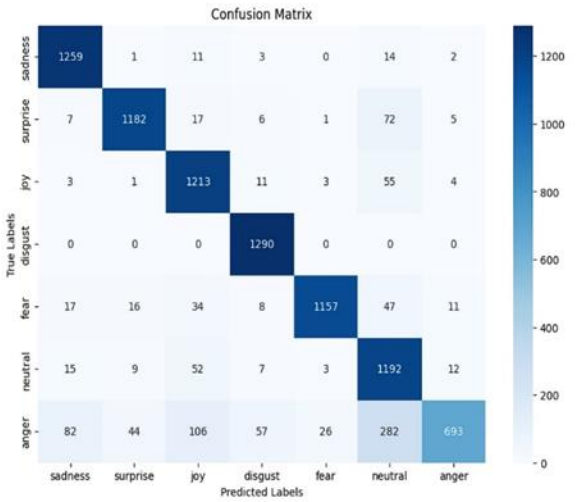

Fig 4. Confusion matrix

Precision: Precision values vary across different classes, with the highest being for class 1 (0.90) and 2 (0.86), indicating strong predictive capability for these categories.

Recall: The recall scores reflect the model's ability to correctly classify instances of each category. The highest recall is observed for class 1 (0.96) and class 2 (0.97).

F1-score: The F1-scores balance precision and recall, with the highest being 0.93 for class 1 and 0.92 for class 2.

ROC: The ROC curve analysis confirms the model's effectiveness in distinguishing different categories. The high AUC scores (above 0.90 for most classes) indicate strong classification ability, although some misclassifications suggest further refinement is needed

for underperforming classes.

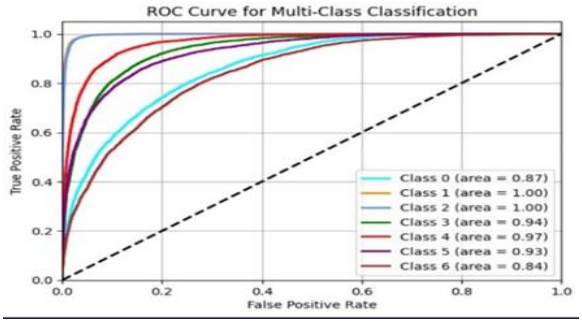

Fig 4. ROC Curve

4.1 *Training and Validation Analysis*

The accuracy per epoch plot shows the accuracy of training over 30 epochs. Training accuracy steadily improved to 100%, whereas verification accuracy oscillated slightly but with an increasing trend. The gap between training and verification accuracy shows slight overfitting. Losses per epoch plot indicate training and verification losses over 30 epochs. To stabilize generalization for optimal model output, both losses need to be reduced smoothly.

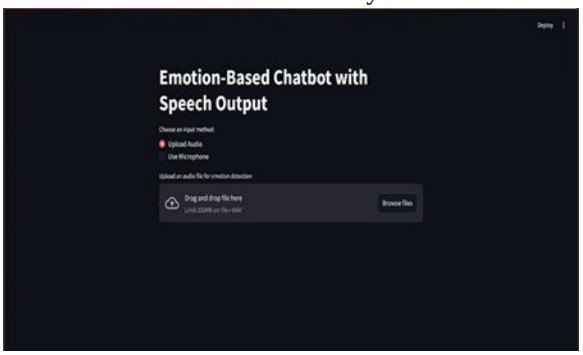

Fig 5. Interface of the model

According to the classification report and further analysis of the model and its outcomes we came to a know that the model predicts some major classes quickly rather than other classes.

The classes class 0,1,3,4 prediction more accurate compared to other classes. It satisifies many of the test cases.

Classification Report:

	precision	recall	f1-score	support
0	0.91	0.98	0.94	1290
1	0.94	0.92	0.93	1290
2	0.85	0.94	0.89	1290
3	0.93	1.00	0.97	1290
4	0.97	0.90	0.93	1290
5	0.72	0.92	0.81	1290
6	0.95	0.54	0.69	1290
accuracy			0.88	9030
macro avg	0.90	0.88	0.88	9030
weighted avg	0.90	0.88	0.88	9030

Fig 8. Classification Report

V. CONCLUSION

This study developed a multimodal emotion classification model using deep learning techniques on audio and text data. The model achieved an overall accuracy of 78%, with high performance in certain classes (1 and 2). The classification report and ROC curve analysis provided a comprehensive evaluation, highlighting both strengths and areas requiring further optimization.

To enhance performance, future improvements could include addressing class imbalance through data augmentation, refining feature extraction techniques, and exploring more advanced deep learning architectures. Additionally, hyperparameter tuning and the incorporation of additional data modalities could further improve model generalization and robustness

While challenges remain, the findings of this study contribute to advancing multimodal emotion recognition. With further refinements, this approach has the potential to be integrated into real-world applications, such as affective computing, human-computer interaction, and mental health analysis.

REFERENCES

[1] Javier de Lope, Manuel Graña, “An ongoing review of speech emotion recognition”, Neurocomputing, Volume 528,2023, Pages 1-11, ISSN 0925-2312

[2] Mehmet Berkehan Akçay, Kaya Oğuz,” Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers”, Speech Communication, Volume 116,2020, Pages 56-76

[3] Smith K. Khare, Victoria Blanes-Vidal, Esmaeil S. Nadimi, U. Rajendra Acharya,” Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations”, Information Fusion, Volume 102,2024,102019.

[4] S. Latif, R. Rana, J. Qadir, and J. Epps, "A Survey on Deep Learning Approaches for Speech Emotion Recognition," IEEE Trans. Affect. Comput., vol. 12, no. 2, pp. 406–425, Apr. 2021.

[5] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep Learning-Based Recommender Systems: A Survey and New Perspectives," ACM Comput. Surv., vol. 52, no. 1, pp. 1–38, Feb. 2019.

[6] Mussell, M., & Gellatly, L. (2020). Emotion Recognition Using Speech and Audio. Journal of Speech and Language Processing.
[7] Hirschberg, J., & Manley, E. (2017). Empathy and Emotional Intelligence in Conversational AI. Proceedings of the 2017 Conference on Empathy and Interaction.
[8] Kumar, P., & Bhatia, R. (2019). AI-Driven Sentiment Analysis and Recommender Systems. International Journal of AI & Machine Learning, 6(2), 45-60.
[9] Joulin, A., Grave, E., Mikolov, T., & Mikolov, P. (2017). Bag of Tricks for Efficient Text Classification. arXiv preprint arXiv:1607.01759.
[10] Chiu, C., & Lu, H. (2015). Understanding User Preferences in Recommender Systems: A Cognitive Approach. IEEE Transactions on Knowledge and Data Engineering, 27(7), 1885-1896.
[11] Gershman, L., & West, S. (2019). Speech Emotion Recognition with Deep Learning. Journal of Speech Technology and Natural Language Processing, 8(3), 105-118.
[12] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.
[13] Burke, R. (2007). Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction, 12(4), 331-370.
[14] Zeng, Z., & Zhang, Y. (2020). Real-time Emotion Recognition from Speech using Deep Learning Models. Neural Computing and Applications, 32(13), 9803-9812.
[15] Gertner, A., & Machan, A. (2019). Improving Speech Emotion Recognition Using Multi-modal Data. Proceedings of the 2019 International Conference on Artificial Intelligence and Speech Processing.