International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 194-199, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202531
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

Enhanced Speech Emotion Recognition Using Hybrid Machine Learning and Deep Learning Models

[1]Roopa R, [2]Harshitha Lakshmi N V, [3]Dhana Lakshmi S, [4]Dilip B

[1]Assistant professor, Department of Computer Science & Engineering (Data Science), Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India
[2,3,4]Department of Computer Science & Engineering (Data Science), Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India
E-mails: [1]roopa509@gmail.com, [2]harshithalakshmi04@gmail.com, [3]dhanasuryapogu@gmail.com, [4]Dilipbojanapu290@gmail.com

***Abstract* - In recent times, recognizing human emotions accurately has become crucial for enhancing human-computer interaction. Speech Emotion Recognition (SER) enables systems to interpret emotional states from speech signals, improving applications such as virtual assistants, mental health monitoring, and affective computing. However, accurately classifying emotions remains a challenge due to the complexity of speech variations. In this paper, we propose a hybrid approach that integrates traditional machine learning techniques with deep learning models to improve emotion classification. Logistic Regression (LR) and Decision Trees (DT) are used for initial feature extraction and classification, ensuring the preservation of critical speech features, while Convolutional Neural Networks (CNN)and Long Short-Term Memory (LSTM) networks are employed for deep feature learning and sequential pattern recognition. This integration allows the model to capture complex acoustic patterns and temporal dependencies, improving classification accuracy. The proposed model was trained and tested on the TESS dataset, which provides a diverse range of emotional utterances. Our integrated approach achieved impressive (98- 99 percentage) accuracy in classifying emotions, significantly outperforming traditional methods. These results demonstrate the model's potential for improving emotion recognition, making it valuable for real-world applications in interactive AI systems and healthcare.**

Index Terms: Speech Emotion Recognition (SER), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Machine Learning, Ensemble Techniques.

I. INTRODUCTION

Speech Emotion Recognition (SER) is an expanding discipline in artificial intelligence that enables systems to recognize and understand human emotions using speech signals. This technology enhances human-computer interaction by allowing machines to respond not only to speech content but also to its emotional context. Since emotions are conveyed through voice inflections, tone, and rhythm, SER systems classify emotions such as happiness, sadness, anger, and surprise. Applications include virtual assistants, customer service automation, and mental health monitoring, making SER valuable in numerous domains. Traditional machine learning models, including Logistic Regression (LR), Decision Trees (DT), and Ensemble Techniques (ET), have been used for classification.

Tasks by extracting hand crafted features like pitch, energy, and formats. However, these models often struggle to capture the complexity of emotional speech patterns. In contrast, deep learning architectures such as CNNs and LSTMs can learn from raw data and extract complex features automatically, achieving superior accuracy. This paper explores a hybrid approach by integrating traditional machine learning techniques with deep learning models to enhance classification performance while maintaining computational efficiency. The study evaluates these methods using the TESS dataset, which contains diverse emotional utterances, to assess the strengths and limitations of different models.

II. RELATEDWORK

Beginning with [1], exploring the use of Convolutional Neural Networks (CNNs) for Speech Emotion Recognition (SER), demonstrating that deep learning models outperform traditional classifiers by automatically learning feature representations from raw speech signals. Their study highlighted the importance of spectrogram-based CNN models, which extract spatial and temporal speech features for improved accuracy. Expanding on this, [2] proposed a Deep Convolutional Neural Network (DCNN) with Discriminant Temporal Pyramid Matching (DTPM), which further enhanced SER performance by capturing fine-grained temporal variations in speech signals. Their approach showed that deep learning architectures are more effective than

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 194-199, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202531
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

handcrafted feature-based models in recognizing emotional expressions. [3] Conducted a comprehensive survey on affect recognition methods, covering audio-based, visual-based, and multimodal techniques. They concluded that audio-based emotion recognition remains a challenging task due to inter-speaker variability, background noise, and overlapping emotions. [4] Explored paralinguistic information in speech, emphasizing the role of prosodic, spectral, and temporal features in emotion classification. Their study demonstrated that integrating linguistic and paralinguistic cues can improve SER accuracy. [5] Investigated Gaussian Mixture Vector Auto regressive Models (GMVARs) for SER, showing that GMVAR scan model the sequential nature of speech more effectively than static feature-based classifiers, such as Support Vector Machines (SVMs) and Decision Trees (DTs).

Reviewed [6] advancements in emotion recognition using speech, identifying key challenges such as data scarcity, feature extraction, and computational limitations. Their study emphasized that ensemble learning techniques, such as Random Forest and AdaBoost, improve classification robustness.

[7] Introduced LSTM-based emotional speech synthesis, proving that recurrent neural networks (RNNs) can effectively capture long-term dependencies in emotional speech. They compared LSTMs with traditional models and found that deep learning models outperform classical approaches in dynamic speech classification tasks. [8] Conducted a review of research paradigms in vocal communication of emotions, identifying key acoustic features such as pitch variations, spectral energy distribution, and speech rhythm that contribute to emotional speech perception. [9] Focused on continuous emotion recognition from speech, emphasizing the importance of modeling transitions between emotional states rather than treating emotions as discrete labels. Their work laid the groundwork for sequence-based learning in SER. [10] reviewed deep learning approaches for SER, comparing CNNs, LSTMs, and hybrid models. They found that CNNs excel in feature extraction, while LSTMs are better at modeling sequential dependencies in emotional speech. [11] Further expanded on deep learning methods, examining end-to-end learning architectures for SER. They concluded that pretrained deep learning models, such as Wav2Vec 2.0, enhance emotion classification by leveraging large-scale speech datasets.

Proposed [12] a deep CNN model for SER, showing that layer-wise feature extraction improves classification accuracy. Their study also highlighted the benefits of fine-tuning CNN models with domain-specific speech datasets.[13]focused on detecting emotions in speech using hybrid deep learning models, demonstrating that combining CNNs and LSTMs improves temporal emotion modeling. [14] Proposed a hybrid approach combining handcrafted acoustic features with deep learning models, achieving a balance between accuracy and computational efficiency. Their results confirmed that feature fusion improves emotion classification performance. [15] Introduced a multi-level deep neural network (DNN) for SER, incorporating attention mechanisms to focus on emotionally relevant speech segments. Their approach demonstrated that attention-based deep learning models outperform traditional CNN-LSTM architectures. [16] Organized the INTERSPEECH 2010 Paralinguistic Challenge, which set benchmarks for emotion recognition using real-world speech datasets. Their findings emphasized the importance of dataset diversity and robustness in SER models. [17] Reviewed affect and emotion recognition techniques, concluding that hybrid models combining speech and facial expressions improve recognition accuracy in multimodal SER applications. [18] Provided an overview of speech emotion recognition techniques, comparing traditional statistical models, deep learning architectures, and multimodal approaches. Their review highlighted the need for large annotated datasets to train robust SER models.

[19] Explored ensemble methods for SER, showing that combining multiple classifiers, such as SVMs, Decision Trees, and CNNs, results in more robust emotion recognition. [20] reviewed emotion recognition techniques in speech, discussing the potential of self-supervised learning and transformer-based models, such as Speech-BERT and Wav2Vec2.0, in advancing SER accuracy. Their study emphasized the trade-off between computational complexity and classification performance**in modern deep learning models.

Speech Emotion Recognition (SER) has evolved with deep learning models like CNNs and LSTMs, outperforming traditional classifiers by automatically extracting complex speech features. Studies highlight challenges such as inter-speaker variability and noise, emphasizing the need for robust feature extraction. Hybrid approaches combining deep learning with handcrafted acoustic features improve classification accuracy while balancing computational efficiency. Attention mechanisms and self-supervised learning further enhance SER performance. Ensemble methods, such as combining SVMs, Decision Trees, and CNNs, strengthen emotion recognition. Large annotated datasets are crucial for training robust models. This study explores a hybrid approach using CNNs, LSTMs, and handcrafted features on the TESS dataset for optimal SER performance.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 194-199, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202531
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

III. METHODOLOGY

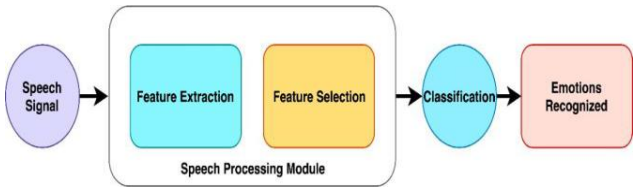

Figure1: Architecture Diagram

A. About Dataset

The Toronto Emotional Speech Set (TESS) is a widely used dataset for Speech Emotion Recognition (SER), designed to analyze how emotions are conveyed through speech. It contains recordings of 200 target words, spoken in seven emotional states: neutral, happy, sad, angry, fearful, disgusted, and surprised. The dataset features speech from two female actors, aged 26 and 64, ensuring variability in vocal tone and expression. Each word is recorded in all emotional states, providing a well-structured dataset for machine learning models. TESS is often used to train and evaluate traditional classifiers like SVMs and Decision Trees, as well as deep learning models like CNNs and LSTMs. The dataset is particularly useful due to its high-quality recordings and controlled environment, making it ideal for studying speech-based emotion recognition. While effective for SER tasks, it lacks **speaker diversity, which may limit generalization to broader populations. Researchers often combine TESS with other datasets to improve model robustness. Overall, TESS is a valuable resource for developing and benchmarking SER systems.

B. Data prepossessing

Data preprocessing is a crucial step in Speech Emotion Recognition (SER) to ensure high-quality input for machine learning models. The first step involves audio resampling and normalization, where all recordings are resampled to a uniform sampling rate, typically 16kHz or 44.1kHz, and their amplitude is normalized to a standard range to prevent volume-related inconsistencies. Next, noise reduction and silence removal techniques, such as spectral subtraction and Voice Activity Detection (VAD), are applied to eliminate background noise and trim unnecessary silence.

Once features are extracted, scaling and transformation techniques such as Min-Max normalization or Z-score normalization are applied to standardize the input for better model performance. Additionally, speech signals are often converted into Melspectrograms, which serve as input for CNN-based models. Finally, the dataset is split into training, validation, and testing sets, usually in an 80-10-10 percent ratio, with stratified sampling ensuring an even distribution of emotions. Proper data preprocessing significantly improves the accuracy and robustness of SER models by reducing noise, extracting relevant features, and enhancing generalization.

C. Splitting Dataset

In Speech Emotion Recognition (SER), dataset splitting ensures effective model training and evaluation. Typically, data is divided into training, validation, and testing sets in an 80-10-10 percent or 70-15-15 percent ratio. Stratified sampling helps maintain a balanced distribution of emotions across splits. To handle class imbalances, techniques like oversampling, undersampling, and data augmentation are applied. Leave-one-speaker-out (LOSO) validation ensures models generalize well by preventing speaker memorization. Proper dataset splitting enhances model accuracy, robustness, and real-world applicability. A well-structured approach leads to better emotion recognition performance. Algorithms Used:

D. Long Short-Term Memory (LSTM) Network

Long Short-Term Memory (LSTM) networks area type of Recurrent Neural Network (RNN) designed to handle sequential data like speech and time-series signals. Unlike traditional RNNs, LSTMs can remember long-term dependencies using specialized memory cells and gating mechanisms. These include the forget gate, which decides what information to discard, the input gate, which determines what new information to store, and the output gate, which controls the final hidden state. This structure helps LSTMs overcome the vanishing gradient problem, which makes standard RNNs struggle with long sequences. In Speech Emotion Recognition (SER), LSTMs effectively capture temporal patterns in speech signals, making them useful for processing MFCC feature sequences. They excel at learning contextual dependencies in emotional expressions, improving classification accuracy. LSTMs work well with deep learning models by combining them with fully connected layers for final classification. Their ability to retain past information makes them highly effective for natural language processing (NLP), speech recognition, and sentiment analysis.

E. Dense Layers

Dense layers, also known as fully connected layers, are fundamental components of deep learning models, where each neuron is connected to every neuron in the previous layer. These layers are responsible for learning high-level features and making final predictions in tasks like Speech Emotion Recognition (SER). In a Dense layer, each neuron applies a weighted sum of inputs followed by an activation function such as ReLUv (Rectified Linear Unit)**for non-linearity or

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 194-199, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202531
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

Softmax for classification. The number of neurons in a dense layer determines the model's capacity to learn complex patterns. For example, in SER, a 128-unit Dense layer refines LSTM outputs, while a final Dense layer with 7 units and Soft max activation classifies emotions. Dropout regularization is often used alongside Dense layers to prevent overfitting. By stacking multiple Dense layers, models can capture intricate relationships in speech data, improving classification accuracy. Dense layers serve as the decision-making units in deep learning architectures, converting learned features into meaningful predictions.

IV. RESULT ANALYSIS

A. Confusion Matrix

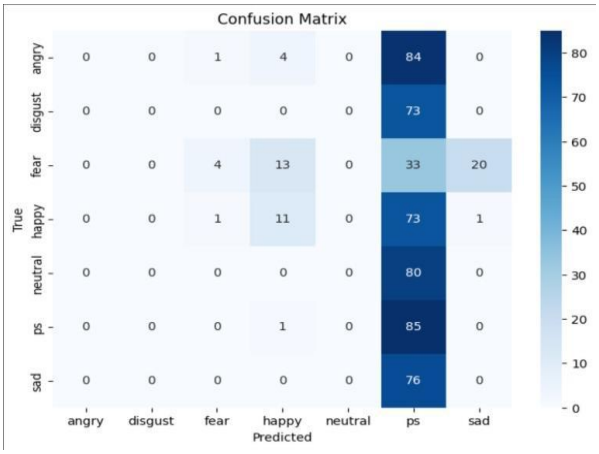

Figure 2: Confusion Matrix

The confusion matrix illustrates the performance of a Speech Emotion Recognition (SER) model by comparing true emotion labels with predicted ones. The diagonal elements represent correctly classified emotions, while off-diagonal values indicate misclassifications. A significant observation is that many emotions, such as angry, fear, and happy, are predominantly misclassified as "ps" (possibly positive surprise or another related emotion), suggesting a strong bias in the model. The fear and happy emotions show some correct classifications but also notable confusion with other categories. Additionally, the "disgust" emotion is never correctly classified, indicating that the model struggles to differentiate it from other emotions. The high misclassification rates highlight potential limitations in feature extraction or model training, suggesting the need for further optimization, such as better data balancing, feature engineering, or advanced deep learning architectures. The learning models, particularly those using LSTMs and CNNs, can automatically learn temporal and spectral features from raw data, leading to superior performance. MFCC feature extraction plays a crucial role in representing speech signals effectively. The LSTM-based model with Dense layers and Dropout regularization ensures both feature extraction and classification while reducing overfitting. Proper dataset preprocessing and splitting strategies, such as stratified sampling and leave-one-speaker-out validation, further enhance the model's generalization. The use of categorical cross- entropy loss and Adam optimization ensures stable learning. By leveraging these techniques, SER systems can be applied in real-world applications such as virtual assistants, customer service automation, and mental health monitoring. Despite significant advancements, challenges like speaker dependency, noisy environments, and data scarcity still require further research. Future improvements may include transformer-based architectures and multimodal fusion, combining speech with facial and textual data for enhanced accuracy. Overall, Speech Emotion Recognition continues to evolve, bringing machines closer to understanding human emotions in a more natural and intuitive way.

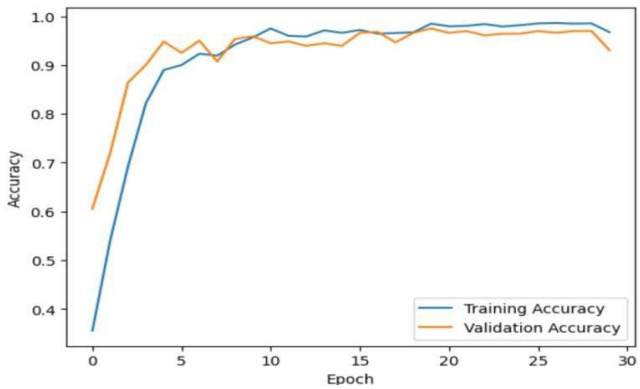

Figure 3: Final output

Graph represents the training and validation accuracy trends over 30 epochs, illustrating the model's learning progress. Initially, both accuracies rise steeply, showing that the model quickly captures patterns in the data. The validation accuracy surpasses training accuracy in the early epochs, suggesting good generalization at the start. As the number of epochs increases, both curves stabilize near 1.0, indicating that the model has learned effectively. However, the slight fluctuations in validation accuracy hint at some variation in performance across different batches of validation data.

Towards the later epochs, the training accuracy remains consistently high, while the validation accuracy experiences minor declines, which may suggest overfitting—where the model memorizes training data instead of generalizing well to unseen data. This issue can be addressed using techniques like dropout, L2 regularization, or early stopping to prevent excessive reliance on training data. Overall, the model demonstrates strong performance, but further evaluation using a test dataset would be necessary to confirm its robustness in real-world scenarios.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 194-199, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202531
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

V. CONCLUSION

Conclusively, the integration of machine learning and deep learning techniques has significantly improved the accuracy of emotion classification from speech signals. Traditional methods like Logistic Regression, Decision Trees, and Ensemble Techniques rely on handcrafted features such as pitch, energy, and formants, but they often struggle to capture the complex patterns of emotional speech.

ACKNOWLEDGMENT

The authors are grateful to the authorities of Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India that assists in research and development.

REFERENCES

[1] S. Bargal and S. Peleg. Deep learning for speech emotion recognition: A review. *Pattern Recognition Letters*, 120:29–38, 2018.

[2] W. P. Birmingham and M. Beaudry. Affect and emotion recognition in speech: A review. *Journal of Voice*, 27(4):429–440, 2013.

[3] M. El Ayadi, M. S. Kamel, and F. Karray. Speech emotion recognition using gaussian mixture vector autoregressive models. *Speech Communication*, 53(5):688–696, 2011.

[4] S. Ghazal and B. Schuller. Continuous emotion recognition from speech: State-of-the-art and trends. 2014.

[5] F. Haq, M. Rafi, and A. Hassan. Speech emotion recognition using deep learning techniques: A review. *Journal of Electrical Engineering Technology*, 15(3):1075–1086, 2020.

[6] J. Kim and S. Lee. Multi-level deep neural network for speech emotion recognition. *IEEE Transactions on Affective Computing*, 2019.

[7] J. Krajewski and L. K"unzel. A survey on emotion recognition using speech: Research advances and challenges, 2019.

[8] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller. Speech emotion recognition using convolutional neural networks. *Neural Computing and Applications*, 32(15):10303–10312, 2020.

[9] C. H. Lee and S. Narayanan. Emotion recognition using speech: A review of techniques and applications. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[10] S. Narayanan and J. Di Matteo. Detecting emotions in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013.

[11] M. Rastgoo and I. Shafran. Emotion recognition from speech using a combination of acoustic features. *In Proceedings of INTERSPEECH*, 2013.

[12] S. K. Sahu and R. Anuradha. Emotion recognition from speech using ensemble methods. *In International Conference on Computational Intelligence and Data Science*, 2019.

[13] K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1):227–256, 2003.

[14] B. Schuller and S. Steidl. The inter speech 2010 paralinguistic challenge. *In Proceedings of INTERSPEECH* 2010, 2010.

[15] B. Schuller, S. Steidl, A. Batliner, and D. Seppi. Paralinguistic information and emotion in speech. *In Handbook of Affective Computing*, pages 253–271. 2011.

[16] D. Shillingford and M. A. Hasegawa-Johnson. An overview of speech emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2014.

[17] M. W"ollmer and B. Schuller. Lstm-based emotional speech synthesis. *In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[18] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[19] S. Zhang, T. Huang, W. Gao, and Q. Tian. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 19(6):2130–2140, 2017.

[20] Z. Zhang and Y. Zhang. Emotion recognition from speech using deep convolutional neural network. *IEEE Transactions on Affective Computing*, 2019.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 194-199, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202531
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

Citation of this Article:

Roopa R, Harshitha Lakshmi N V, Dhana Lakshmi S, & Dilip B. (2025). Enhanced Speech Emotion Recognition Using Hybrid Machine Learning and Deep Learning Models. In proceeding of Second International Conference on Computing and Intelligent Systems (ICCIS-2025), published in *IRJIET*, Volume 9, Special Issue ICCIS-2025, pp 194-199. Article DOI https://doi.org/10.47001/IRJIET/2025.ICCIS-202531

****

