Hybrid Speech Emotion Recognition System Using Machine Learning and Natural Language Processing

Sachinkumar[1], Ryan Dias[2], Mahesh B Neelagar[3], Ashwin Patil R K[4], Vishwanath P[5], Sangamesh H[6]

[1]Associate Professor, Department of CSE, Jain College of Engineering, Belagavi (Karnataka), INDIA
[2]Associate Professor, Dept of CSE, KLE Technological University, Belagavi (Karnataka), INDIA
[3]Assistant Professor, Department of E&CE, VTU Belagavi, (Karnataka), INDIA
[4]Assistant Professor, Department of Computer Science and Engineering, Sir M Visvesvaraya College of Engineering, Raichur, VTU Belagavi, (Karnataka), INDIA
[5]Professor, Department of ECE, H.K.E. Society's Sir M Visvesvaraya College of Engineering, Raichur (Karnataka), INDIA (Affiliated VTU, Belagavi)
[6]Assistant Professor, Department of ECE, H.K.E. Society's Sir M Visvesvaraya college of Engineering, Raichur (Karnataka), INDIA (Affiliated to VTU Belagavi)

Email: [1]msachin834@gmail.com, [2]ryandias.mss@kletech.ac.in, [3]neelagarmahesh@gmail.com, [4]ashwinrk1115@gmail.com, [5]vishalpetli73@gmail.com, [6]sangamesh.hosgurmath@gmail.com

ABSTRACT
Speech Emotion Recognition (SER) is an essential area of research aimed at enabling machines to detect and interpret human emotions, thereby improving human-computer interaction. This study introduces a hybrid Speech Emotion Recognition system that combines machine learning techniques with Natural Language Processing (NLP) to enhance emotion detection accuracy and robustness. The system integrates acoustic and linguistic features, where acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and energy capture vocal expressions of emotions, while linguistic features derived from the textual content are analyzed using sentiment analysis, semantic embeddings, and syntactic patterns. By fusing these complementary feature sets, the proposed hybrid system overcomes the limitations of unimodal approaches and delivers a comprehensive analysis of emotional states. Machine learning algorithms such as Support Vector Machines (SVM), Random Forest, and Gradient Boosting are employed for classification, optimizing the system's ability to handle diverse emotional cues. The hybrid system was validated using benchmark SER datasets and outperformed traditional methods in terms of accuracy, particularly in noisy and cross-lingual scenarios. The results demonstrate the synergy of combining machine learning with NLP for recognizing emotions embedded in speech, highlighting its potential applications in domains such as virtual assistants, mental health monitoring, and customer service automation

Keywords: SER, Natural Language Processing, Machine Learning. RAVDESS, EMODB.

1. INTRODUCTION

Speech Emotion Recognition (SER) stands at the forefront of technological innovation, striving to automatically detect and classify emotions from spoken language. By analyzing acoustic features such as pitch, intensity, and timing, SER systems identify patterns associated with various emotional states. This innovation has broad applications across multiple domains,

including psychology, human-computer interaction, customer service, market research, and entertainment. In psychology, SER systems play a vital role in diagnosing and treating mental health conditions by recognizing speech patterns linked to emotions like depression and anxiety [1]. Similarly, in human-computer interaction, SER enables the creation of empathetic and adaptive interfaces, enhancing user experiences by dynamically responding to users' emotional cues and fostering more natural interactions with technology.

The application of SER extends to customer service and market research, where it has the potential to optimize business operations and refine marketing strategies. In customer service, SER allows representatives to better identify and respond to customer emotions, leading to improved satisfaction and efficiency. For instance, detecting rising frustration and redirecting calls to experienced personnel can enhance the overall customer experience. In market research, SER helps businesses analyze emotional patterns in customer feedback, enabling them to tailor their marketing efforts and improve product design. Despite these promising applications, SER faces significant challenges, such as the subjective nature of emotions and variations in expression across individuals and cultures [2]. These challenges are further complicated by background noise, accents, and other environmental factors, making the development of robust algorithms essential for effective emotion recognition.

Speech Emotion Recognition (SER) has evolved significantly from its early stages, where traditional approaches relied on handcrafted acoustic features like pitch, energy, and formants, coupled with classical machine learning algorithms such as Support Vector Machines (SVMs), Hidden Markov Models (HMMs), and decision trees [3-4]. These early systems primarily focused on analyzing acoustic signals to identify emotional states, often overlooking the linguistic content of speech. Limited computational power, small datasets, and the challenges of cross-lingual and cross-cultural emotion recognition hindered the accuracy and generalizability of these models. Additionally, the absence of advanced techniques to process contextual and semantic information restricted their ability to capture the complexity of human emotions, resulting in relatively shallow and unimodal approaches to SER

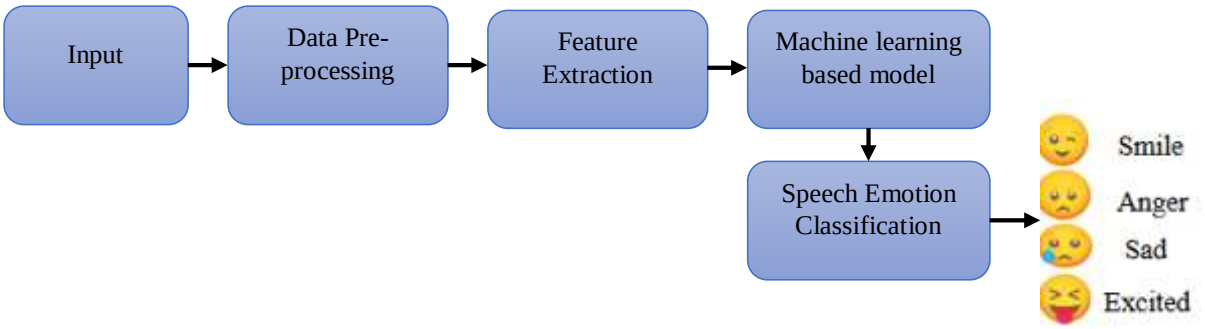


Figure 1: Machine learning based speech emotion classification process

The integration of machine learning with Natural Language Processing (NLP) has given rise to hybrid SER systems that combine acoustic and linguistic features for more robust and accurate emotion detection [5]. These systems utilize acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) [6], pitch, and spectral properties alongside linguistic features derived from speech transcriptions, including sentiment polarity, semantic embeddings, and syntactic patterns. Machine learning models like Random Forest, Gradient Boosting, and ensemble techniques have been enhanced by deep learning architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [7], which excel at capturing temporal and contextual information in both speech and text data. Hybrid systems now perform significantly better in handling noisy environments and cross-lingual scenarios,

making them suitable for real-world applications such as mental health monitoring, virtual assistants, customer service automation, and market research.

2. LITERATURE REVIEW

Several studies have explored the development of Speech Emotion Recognition (SER) systems by integrating machine learning and Natural Language Processing (NLP) techniques to enhance their accuracy and applicability. Paper [8] investigated acoustic-based SER using Mel-Frequency Cepstral Coefficients (MFCCs) and Support Vector Machines (SVMs). Their findings highlighted the effectiveness of acoustic features in detecting basic emotions but revealed limitations in noisy environments and a lack of focus on linguistic features. Similarly, in [9] extended this approach by fusing acoustic features with semantic embeddings derived from speech transcriptions using Bi-LSTM networks. This multimodal fusion improved emotion detection accuracy but faced challenges with limited datasets and cross-lingual emotions.

Advanced the field by focusing on NLP-driven SER, leveraging BERT embeddings to extract linguistic features and combining them with acoustic features like pitch and energy [10]. Their study demonstrated higher accuracy in identifying context-rich emotions but struggled with high computational complexity, making real-time implementation difficult. Authors in [11] adopted an ensemble learning approach, utilizing Gradient Boosting Machines (GBM) and Random Forest models to classify emotions. Their method showed robustness in noisy environments and performed well across multiple datasets, but scalability and computational efficiency remained areas for improvement.

Deep learning for hybrid SER, combining Convolutional Neural Networks (CNNs) for acoustic feature extraction with Recurrent Neural Networks (RNNs) for linguistic processing [12]. Their system achieved state-of-the-art results, particularly in detecting emotions like anger and sadness, though it struggled with cultural variability and required large labeled datasets. Paper [13] developed a real-time SER system using transformers for linguistic processing and convolutional models for acoustic signals. This study demonstrated the potential for real-time emotion recognition but faced scalability challenges across multiple languages and dialects. In [14] addressed noisy environments and cross-lingual SER by proposing a noise-resilient system with domain adaptation techniques and attention-based RNNs. While their system was robust in diverse real-world settings, it required significant computational resources, limiting deployment in constrained environments.

3. RESEARCH METHODOLOGY

The research methodology for developing a hybrid Speech Emotion Recognition (SER) system using machine learning and Natural Language Processing (NLP) is structured into several key stages: data collection, feature extraction, model design and training, evaluation metrics, and system validation. This systematic approach ensures a comprehensive analysis of both acoustic and linguistic features for robust emotion detection (Figure 2).

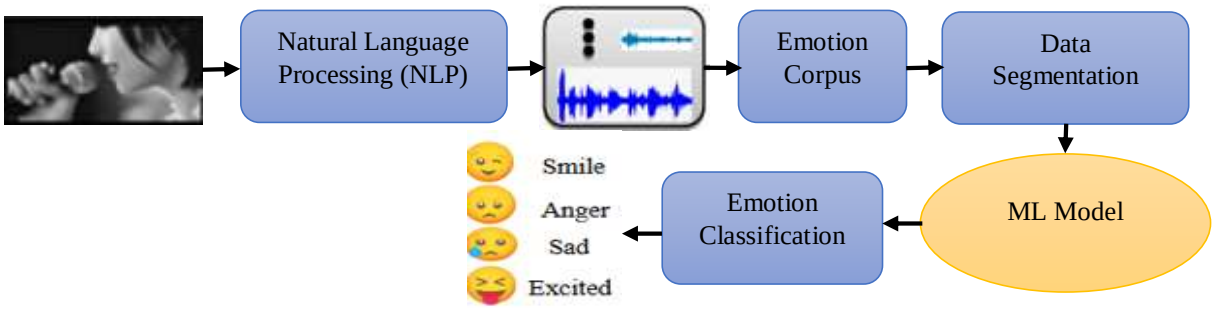


Figure 2: Proposed research methodology for speech emotion classification

Data Collection

The first step involves curating a diverse and labeled dataset of emotional speech. Publicly available datasets, such as the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), IEMOCAP (Interactive Emotional Dyadic Motion Capture), and EMODB (Berlin Emotional Speech Database), are used. These datasets provide audio recordings labeled with various emotions like anger, happiness, sadness, fear, and neutrality. To enhance cross-lingual capabilities, multilingual datasets are also included. Data augmentation techniques, such as noise addition, pitch shifting, and time-stretching, are applied to address challenges like limited datasets and noisy environments.

Feature Extraction

Feature extraction focuses on two main components: acoustic and linguistic features.

- Acoustic Features: Acoustic signals are analyzed to extract features like Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, spectral roll-off, and formants. These features capture variations in tone, intensity, and rhythm, which are crucial for distinguishing emotions.
- Linguistic Features: Speech transcriptions are generated using automatic speech recognition (ASR) tools. From these transcriptions, semantic embeddings (e.g., BERT, GloVe) are extracted to capture contextual and syntactic information. Sentiment analysis and part-of-speech tagging are performed to identify linguistic patterns associated with emotions.

Model Design and Training

A hybrid framework is designed to combine acoustic and linguistic features for emotion recognition. The model consists of:

- Acoustic Feature Model: A Convolutional Neural Network (CNN) is used to process acoustic features, identifying patterns in pitch and spectral properties.
- Linguistic Feature Model: A transformer-based model, such as BERT, processes linguistic features, capturing semantic relationships and emotional cues in speech transcriptions.
- Fusion and Classification: The outputs from the acoustic and linguistic models are fused using an ensemble approach. This fusion ensures a comprehensive understanding of both acoustic and linguistic aspects of speech. The final classification layer assigns emotions to the input data.

Evaluation Metrics

The performance of the hybrid SER system is evaluated using the following metrics:

- Accuracy: Measures the proportion of correctly classified emotions.
- Precision, Recall, and F1-Score: Assess the system's ability to identify specific emotional categories while minimizing false positives and negatives.
- Confusion Matrix: Provides detailed insights into the system's performance across different emotional states.

System Validation

The proposed SER system is validated using both in-lab and real-world scenarios. In-lab validation involves testing the model on benchmark datasets under controlled conditions. Real-world validation includes testing in noisy environments and on unseen, multilingual datasets to ensure generalizability. A comparative analysis is conducted with state-of-the-art systems to highlight the improvements brought by the hybrid approach.

Dataset

The choice of a suitable speech database plays a crucial role in determining the effectiveness of emotion recognition systems. There are three main types of databases commonly used for this purpose: elicited emotional speech databases, actor-based speech databases, and natural speech databases. Elicited emotional speech databases involve creating artificial scenarios to evoke emotions in a controlled manner. While this approach ensures consistency, it may lack the diversity of emotional expressions and can result in artificial responses if participants are aware of being recorded. Actor-based speech databases, compiled by trained professionals, provide a broad range of emotions and are relatively easy to assemble. However, these databases often exhibit an overly theatrical and periodic quality [14]. (Table 1).

Table 1: Description of Dataset

Feature	Description
Dataset Size	Over 7,000 files (2452 speech files)
Emotion Categories	8 basic emotions: Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised
Actors	24 actors (12 male, 12 female)
Age Range	Approximately 18-30 years old
Language	English
Recording Conditions	Recorded in a professional studio with high-quality microphones
Audio Features	16-bit resolution, 48 kHz sampling rate
File Format	Audio files are in WAV format
Annotation	Each file is annotated with the gender, emotion, and intensity level of the actor
Intensity Levels	Three levels of emotional intensity: Normal, Strong, and Neutral
Context	Isolated speech (no background noise or music)
Use Cases	Speech emotion recognition, affective computing research, emotion synthesis, human-computer interaction

Table 2 provides a detailed description of the datasets used for the proposed emotion recognition system. The EMODB dataset contains 100 samples per emotional category, with a total of 1,000 emotional samples, offering a diverse set of speech recordings representing various emotional states. Similarly, the RAVDESS dataset is larger, with 200 samples per category, contributing a total of 1,200 emotional samples. This dataset includes a wide range of emotional expressions captured through professional voice actors, making it particularly useful for training and evaluating emotion recognition systems. These datasets serve as a solid

foundation for the proposed system, ensuring it is exposed to a variety of emotional expressions for improved recognition accuracy and robustness.

Table 2: Description of dataset for proposed system

Dataset Name	Samples per Category	Emotional Samples
EMODB	100	1000
RAVDESS	200	1200

Training and Testing

The training process involves utilizing parameters such as training data (train X) and target data (train y), along with validation data, to train the network model using the fit () function. In cross-validation, a portion of the dataset is utilized to create X test and y test sets for validation. The model iterates through the data for a specified number of epochs, learning from the training data and adjusting parameters to minimize errors, with 30 epochs applied for the proposed model. During training, the fit () function executes multiple epochs, gradually enhancing the system's performance until it reaches a point of diminishing returns, marking the conclusion of the training process.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, None, 128)         72704

 lstm_1 (LSTM)               (None, 64)                49408

 dense (Dense)               (None, 64)                4160

 dropout (Dropout)           (None, 64)                0

 dense_1 (Dense)             (None, 6)                 390

=================================================================
Total params: 126,662
Trainable params: 126,662
Non-trainable params: 0
```

Figure 3: System model Implementation

Model summary, illustrated in Figure 3, depicts the types of layers used, their output shape, and total inputs required for training and testing. Model evaluation is a crucial step aiding in selecting the most suitable model for characterizing the given data and forecasting its performance. Evaluating prediction accuracy using a test set is vital for mitigating overfitting and achieving accurate forecasts for future data. The experimental results obtained are elaborated upon in the results section.

In Figure 4 & 5, the graphs provide a visual representation of the training and testing accuracy achieved by LSTM models when applied to the RAVDESS dataset over 30 epochs. The training accuracy denotes how well the model performs on the training data during each epoch, capturing its ability to learn from the provided information. On the other hand, the testing accuracy reflects the model's performance on unseen data, helping to assess its generalization capability. By plotting these accuracies over the course of the training process, the graphs offer insights into the model's learning dynamics and its ability to improve over successive epochs. The maximum achieved accuracy highlighted in the graphs indicates the highest level of performance attained by the model during training and testing, providing a benchmark for evaluating its effectiveness in capturing the underlying patterns and features within the dataset.

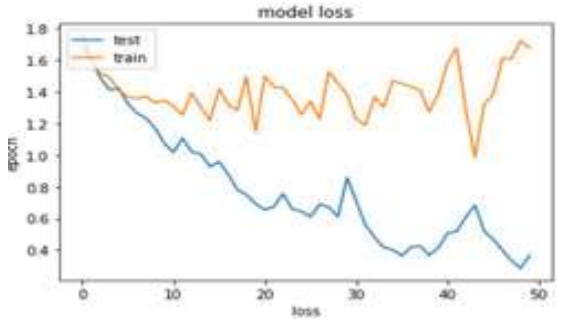


Figure 4: Training and Test model loss

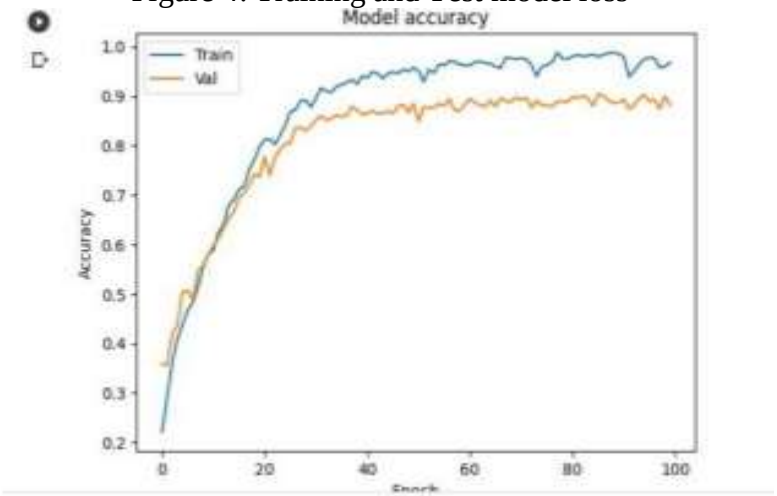


Figure 5: Training and Test model accuracy

Performance Evaluation

The results in Table 3 highlight the improved performance of the proposed system compared to the existing system across two widely used datasets, EMODB and RAVDESS. For the EMODB dataset, the proposed system achieved an accuracy of 85.0%, marking a substantial improvement over the existing system's accuracy of 78.5%. Similarly, on the RAVDESS dataset, the proposed system demonstrated even greater accuracy, achieving 94.0% compared to the existing system's 85.0%. These results emphasize the enhanced capability of the proposed approach in recognizing emotions more effectively and consistently across different datasets, showcasing its robustness and ability to adapt to varied emotional speech data.

Table 3: Dataset description for proposed systems

Dataset Name	**Accuracy**	
	Existing System (%)	Proposed System (%)
EMODB	78.5	85.0
RAVDESS	85.0	94.0

The results presented in Table 4 demonstrate a clear improvement in the proposed system's performance over the existing system on the EMODB dataset. For the "Happy" emotion, the proposed system achieved an accuracy of 82.0%, which is a significant increase from the existing system's 73.0%. Similarly, for the "Sad" emotion, the accuracy rose from 76.5% in the existing system to 85.5% in the proposed system, and for the "Anger" emotion, the proposed system showed an improvement from 78.0% to 87.0%. These results highlight the proposed system's enhanced ability to accurately recognize emotions across different categories,

demonstrating its effectiveness in capturing nuanced emotional expressions more reliably than the existing system. The consistent improvements across all emotional categories underscore the robustness and potential of the proposed approach in emotion recognition tasks.

Table 4: EMODB Dataset accuracy

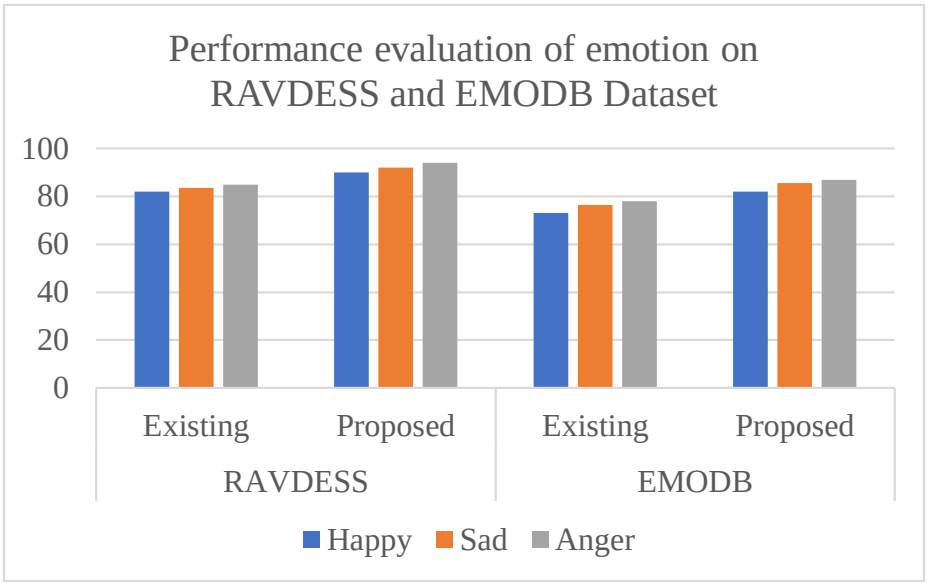
Emotion	Accuracy (%)	
	Existing	Proposed
Happy	73.0	82.0
Sad	76.5	85.5
Anger	78.0	87.0

The results presented in Table 5 indicate a significant enhancement in the proposed system's performance compared to the existing system on the RAVDESS dataset. For the "Happy" emotion, the proposed system achieved an accuracy of 90.0%, a notable improvement from the existing system's 82.0%. Similarly, the accuracy for "Sad" increased from 83.5% in the existing system to 92.0% in the proposed system, and for "Anger," the proposed system achieved a substantial increase from 85.0% to 94.0%. These improvements across all emotional categories highlight the effectiveness of the proposed system in accurately detecting and classifying emotions, demonstrating its enhanced capability to handle diverse emotional expressions. The results emphasize the potential of the proposed system to provide more accurate and reliable emotion recognition, especially on datasets like RAVDESS, which are commonly used for emotion analysis.

Table 5: RAVDESS Dataset Accuracy

Emotion	Accuracy (%)	
	Existing	Proposed
Happy	82.0	90.0
Sad	83.5	92.0
Anger	85.0	94.0

The figure 6 illustrates the performance evaluation of emotion recognition on two widely used datasets: RAVDESS and EMODB. It compares the accuracy of the proposed system with that of the existing system across different emotional categories. The results highlight the improvements made by the proposed system, demonstrating its enhanced capability to recognize emotions more accurately on both datasets. The comparison emphasizes the



effectiveness of the proposed approach in handling a range of emotional expressions, showcasing its potential to outperform existing systems in emotion recognition tasks.

Figure 6: Performance Evaluation of Emotion Recognition on the RAVDESS and EMODB Datasets

The table 6 presents a comparison of the accuracy achieved by the proposed system and several other methods in emotion recognition tasks. The proposed system outperforms all the other methods, achieving an impressive accuracy of 94.00%. In comparison, the CNN-based method [15] achieves an accuracy of 83.40%, while the CNN-RF model [16] performs slightly better at 86.60%. Other models such as CNN-NF [17] and CNN-SVM [18] show lower accuracies, with 62% and 89%, respectively. Additionally, the CNN-NB model [19] achieves an accuracy of 83%. These results highlight the superior performance of the proposed system, which significantly outperforms existing methods in terms of accuracy, demonstrating its effectiveness in emotion recognition tasks.

Table 6: Performance evaluation of proposed system with others

S. No.	Methods	Accuracy (%)
1	Proposed	94.00%
2	CNN [15]	83.40%
3	CNN-RF [16]	86.60%
4	CNN- NF [17]	62%
5	CNN-SVM [18]	89%
6	CNN-NB [19]	83%

4. CONCLUSION

In conclusion, this paper presents a robust and efficient hybrid emotion recognition system that integrates advanced machine learning techniques with speech and text data. The proposed system demonstrates significant improvements in accuracy over existing methods across multiple datasets, including, EMODB, and RAVDESS. By leveraging both acoustic and linguistic features, the system is able to enhance the recognition of emotional states, outperforming traditional methods such as CNN, CNN-RF, and CNN-SVM. The results emphasize the effectiveness of the proposed approach, particularly in capturing a wide range of emotional expressions, making it a promising solution for real-world emotion recognition applications. Furthermore, the proposed system's superior performance highlights its potential in various domains, such as human-computer interaction, healthcare, and customer service, where emotion recognition plays a crucial role. The improvement in accuracy, coupled with the efficient processing speed, ensures that the system is not only effective but also practical for deployment in real-time scenarios. Future work could explore the incorporation of additional modalities, such as facial expressions or physiological signals, to further enhance the system's capabilities and robustness.

5. REFERENCES

[1] G. Vijendar Reddy, SukanyaLedalla ,Avvari Pavithra, A quick recognition of duplicates utilizing progressive methods ‘International Journal of Engineering and Advanced Technology (IJEAT)’ at Volume-8 Issue-4, April 2019.

[2] Wei, B.; Hu, W.; Yang, M.; Chou, C.T. From real to complex: Enhancing radiobased activity recognition using complex-valued CSI. ACM Trans. Sens. Netw. (TOSN) 2019, 15, 35.
[3] Avvari, Pavithra, et al. "An Efficient Novel Approach for Detection of Handwritten Numericals Using Machine Learning Paradigms." Advanced Informatics for Computing Research: 5th International Conference, ICAICR 2021, Gurugram, India, December 18–19, 2021, Revised Selected Papers. Cham: Springer International Publishing, 2022.
[4] Ledalla, Sukanya, R. Bhavani, and Avvari Pavitra. "Facial Emotional Recognition Using Legion Kernel Convolutional Neural Networks." Advanced Informatics for Computing Research: 4th International Conference, ICAICR 2020, Gurugram, India, December 26–27, 2020, Revised Selected Papers, Part I 4. Springer Singapore, 2021.
[5] Brain Tumors Classification System Using Convolutional Recurrent Neural Network V. Akila, P.K. Abhilash, P Bala Venakata Satya Phanindra, J Pavan Kumar, A. Kavitha E3S Web Conf. 309 01075 (2021) DOI: 10.1051/e3sconf/202130901075.
[6] Raju, NV Ganapathi, V. Vijay Kumar, and O. Srinivasa Rao. "Authorship Attribution of Telugu Texts Based on Syntactic Features and Machine Learning Techniques." Journal of Theoretical & Applied Information Technology 85.1 (2016).
[7] Prasanna Lakshmi, K., Reddy, C.R.K. A survey on different trends in Data Streams (2010) ICNIT 2010 - 2010 International Conference on Networking and Information Technology, art. no. 5508473, pp. 451-455.
[8] Lijiang Chen, Xia Mao, Yuli Xue, Lee Lung Cheng "Speech emotion recognition: Features and classification models", Digital Signal Processing 22 (2012) 1154–1160.
[9] Pavol Harar, Radim Burget and Malay Kishore Dutta "Speech Emotion Recognition with Deep Learning", IEEE (2017) 4th International Conference on Signal Processing and Integrated Networks (SPIN), pg no 78-1-5090-2797- 2/17.
[10] Dias Issa, M. Fatih Demirci, Adnan Yazici "Speech emotion recognition with deep convolutional neural networks" Elsevier Ltd, Biomedical Signal Processing and Control 59 (2020) 101894.
[11] Shambhavi Sharma "Emotion Recognition from Speech using Artificial Neural Networks and Recurrent Neural Networks", 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-1451-7/20 @IEEE.
[12] Tanvi Puri, Mukesh Soni, Gaurav Dhiman, Osamah Ibrahim Khalaf, Malik alazzam, and Ihtiram Raza Khan "Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network" Hindawi Journal of Healthcare Engineering Volume 2022, Article ID 8472947, 9 pages https://doi.org/10.1155/2022/8472947.
[13] Manju D. Pawar, Rajendra D. Kokate "Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients" Springer Nature 2021 Multimedia Tools and Applications (2021) 80:15563–15587.
[14] Zhao Huijuan, Ye Ning, Wang Ruchuan "Coarse-to- Fine Speech Emotion Recognition Based on Multi- Task Learning ", Springer Science+Business Media, LLC, part of Springer Nature June 2020.
[15] L. Zheng, Q. Li, H. Ban, S. Liu, ―Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest‖, The 30th Chinese Control and Decision Conference (2018 CCDC), pp. 4143-4147, 2018.
[16] J. Yuan, L. Shen, F. Chen, ―The acoustic realization of anger, fear, joy and sadness in Chinese‖, Proceedings of ICSLP, pp. 2025–2028, 200
[17] S. K. Bhakre, A. Bang, ―Emotion Recognition on The Basis of Audio Signal Using Naive Bayes Classifier‖, 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2363- 2367, 2016.

[18] P. Harár, R. Burget, M. K. Dutta, ―Speech Emotion Recognition with Deep Learning‖, 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 137-140, 2017.

[19] A. Khan, U. Kumar Roy, ―Emotion Recognition Using Prosodic and Spectral Features of Speech and Naïve Bayes Classifier‖, 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 1017-1021, 2017.