WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Speech Emotion Recognition Using Deep Learning

Ms. M.Saravanan [1] , S.Rajalakshmi [2]

Assistant Professor , II MSC
Department of Computer Science
K.R.College of Arts & Science, K.R.Nagar,Kovilpatti- Tamil Nadu 628503

ABSTRACT :

Speech Emotion Recognition (SER) has emerged as a critical area of research in human-computer interaction, enabling systems to recognize and respond to human emotions effectively. This paper explores the implementation and comparative analysis of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) for Speech Emotion Recognition to predict emotion like Happy, Sad, Angry, Calm, Neutral, Fearful, Disgust and Excitement. CNNs are utilized for extracting spatial and spectral features from audio spectrograms, while RNNs, particularly Long Short-Term Memory (LSTM) networks, capture temporal dependencies and sequential patterns in speech data. The dataset used for experimentation consists of pre-processed audio files with labeled emotional categories. The *Ravdess* dataset is collected from K*aggle*. The experimental results indicate that CNN model attains a high level of accuracy. Experimental results indicate that each architecture excel in specific aspects of feature extraction and sequence modeling, demonstrating superior overall performance in capturing contextual emotional cues.

Keyword: CNNs, RNNs

I. INTRODUCTION :

Speech Emotion Recognition (SER) is an essential area of research that focuses on identifying and interpreting emotional states from spoken language. As communication is profoundly influenced by emotional nuances, the ability to recognize these emotions enhances human-computer interactions and improves various applications, including virtual assistants, call centers, and therapeutic tools. Traditional methods for emotion recognition relied heavily on handcrafted features and shallow learning algorithms, which often struggled to capture the complexity of vocal expressions. However, the advent of deep learning has transformed this field by enabling the extraction of high-level features directly from raw audio data, significantly enhancing recognition accuracy. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have emerged as popular architectures for SER, effectively modeling both the spatial and temporal aspects of speech signals. These models can learn from vast datasets, allowing them to generalize well across different speakers, accents, and emotional expressions. Despite the progress made, challenges such as recognizing emotions in spontaneous speech, handling noise, and adapting to diverse cultural contexts persist. Current research is focused on optimizing deep learning models, improving feature extraction techniques, and exploring semi-supervised and unsupervised learning approaches to leverage unlabeled data. As SER continues to evolve, its implications extend beyond technology, promising to enrich our understanding of emotional communication and facilitate more empathetic interactions between humans and machines. This introduction sets the stage for exploring the methodologies, challenges, and future directions in the field of speech emotion recognition using deep learning techniques.

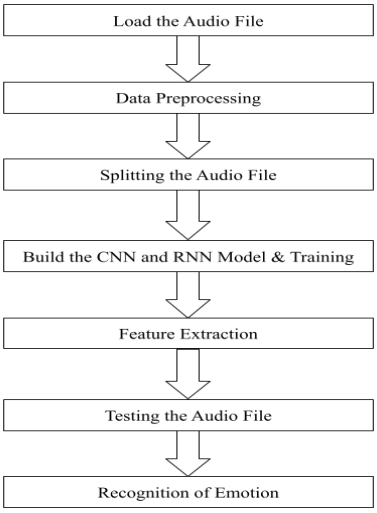
II. LITERATURE SURVEY :

Recent advancements in automatic speaker recognition models emphasize the importance of feature selection and classification methods for speech emotion recognition. The presented deep learning model demonstrates significant performance improvements, achieving accuracy rates of up to 94.21% using various audio features from the Berlin database, which includes around 500 recordings from male and female speakers. The study highlights the critical role of input quality, as it directly influences classifier performance. By integrating deep learning techniques, particularly Convolutional Neural Networks (CNN), the model effectively extracts relevant features while minimizing traditional processing complexities. Additionally, the research underscores the implications of emotional recognition in various fields, including customer service and emergency response, where understanding speaker emotions can enhance communication and decision-making. Overall, the paper contributes to the growing body of knowledge in human-computer interaction by offering a robust solution for real-time speech recognition and emotion analysis [1]. Speech Emotion Recognition (SER) has gained prominence in Human-Computer Interaction (HCI), focusing on identifying and classifying emotional states from speech signals. Despite the advances, challenges remain in extracting effective emotional features, as traditional automatic speech recognition systems often overlook paralinguistic information essential for understanding human emotions. Various methodologies have been explored, including both classical classifiers like Support Vector Machines and modern deep learning approaches, such as Convolutional Neural Networks, which show promise in enhancing recognition accuracy. The effectiveness of these systems largely depends on feature extraction techniques, which must capture salient

acoustic and prosodic characteristics. While SER has made significant progress, the field continues to face obstacles, necessitating robust algorithms and innovative feature sets to improve performance and enrich human-computer communication. Overall, the literature indicates a critical need for interdisciplinary approaches to bridge the gap between emotional understanding and machine recognition [2]. Automatic Speech Emotion Recognition (SER) has emerged as a vital research area within Human-Computer Interaction (HCI), driven by the need for more natural communication interfaces. Key acoustic features such as Mel Frequency Cepstrum Coefficients (MFCC) and Mel Energy Spectrum Dynamic Coefficients (MEDC) are extracted from speech to classify emotions using classifiers like Support Vector Machines (SVM). The Berlin Emotional Database, comprising diverse emotional utterances from professional actors, serves as a foundation for training these classifiers. SVM demonstrates high accuracy rates, achieving 93.75% classification accuracy for gender-independent cases and even higher rates for male and female speech. Various classifiers, including Neural Networks and Gaussian Mixture Models, have also been explored, but SVM's ability to construct optimal hyper planes in high-dimensional spaces has proven particularly effective. The implications of SER extend across multiple applications, including psychiatric diagnostics and educational tools, underscoring its significance in enhancing human-computer interactions [3]. The paper discusses the complexities of automatic Speech Emotion Recognition (SER), particularly the variations in emotional expression across individuals. It emphasizes the importance of feature selection for effective emotion recognition and critiques existing deep learning approaches that often rely on hand-crafted features. To address these limitations, the authors propose a multi-scale Convolutional Neural Network (MCNN) that captures features at various time scales and frequencies directly from raw speech signals. This end-to-end model incorporates multiple transformations to enhance feature extraction, leading to improved performance on the SAVEE emotion database. The study underscores the potential of learning features through multi-scale networks for better SER outcomes [4]. Emotion recognition in human communication, particularly through spoken language, remains a complex challenge for computers due to the subjective nature of emotions. Recent research emphasizes frameworks that identify emotional segments of conversations independent of semantic content, utilizing deep learning techniques like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. The effectiveness of these models has been demonstrated using Mel-frequency Cepstral Coefficients (MFCCs) on datasets such as RAVDESS and TESS, achieving impressive accuracy rates. Moreover, the integration of emotional recognition technologies has significant implications for human-machine interaction, enhancing applications in areas like mental health, customer service, and smart home devices. Despite the advancements, challenges remain, particularly in distinguishing nuanced emotions and improving model robustness [5]. Speech Emotion Recognition (SER) has emerged as a vital area within human-computer interaction, leveraging speech signals to identify emotional states across various applications such as healthcare, virtual reality, and customer service. Recent advancements in SER focus on the integration of deep learning techniques, particularly Convolutional Neural Networks (CNNs), which have shown promise in extracting salient features from speech spectrograms. However, challenges remain, including speaker variability and the computational complexity associated with traditional deep learning models. Innovative approaches, such as Dynamic Adaptive Thresholding for noise removal and specialized CNN architectures, aim to enhance recognition accuracy while reducing model complexity. The use of benchmark datasets like IEMOCAP and RAVDESS demonstrates significant improvements in performance metrics, underscoring the effectiveness of these advanced methodologies. Overall, the continuous evolution of SER techniques indicates a promising future for emotion recognition in real-world applications[6].Speech Emotion Recognition (SER) has traditionally relied on supervised learning methods, which face significant limitations due to the scarcity of labeled data. Recent advances have explored semi-supervised learning, leveraging the wealth of available unlabeled speech data to enhance recognition performance. Autoencoders have emerged as a promising tool in this context, allowing for the integration of both generative and discriminative training objectives. The proposed semi-supervised autoencoder framework demonstrates effective utilization of small labeled datasets alongside larger unlabeled corpora, achieving state-of-the-art results across various benchmarks. This approach highlights the potential of combining supervised and unsupervised methods to improve SER, addressing challenges such as data scarcity and variability in emotional expression. Overall, the literature underscores a shift towards more flexible learning paradigms that harness both labeled and unlabeled data for improved emotion recognition outcomes [7]. The literature on Speech Emotion Recognition (SER) highlights the importance of speech as a natural medium for conveying emotions and the challenges associated with classifying these emotions due to limited emotion-labeled datasets. Recent advancements have focused on combining traditional features, such as Mel-Frequency Cepstral Coefficients (MFCCs), with deep learning techniques, specifically leveraging pre-trained Convolutional Neural Networks (CNNs) for feature extraction from spectrograms. Studies have demonstrated that integrating these hybrid features into classification algorithms, like Support Vector Machines (SVMs) and Long Short-Term Memory (LSTM) networks can enhance prediction accuracy. Furthermore, real-time SER applications are emerging, providing practical solutions in fields like healthcare and customer service. While the findings suggest promising improvements in emotion recognition, ongoing challenges remain, particularly in predicting complex emotions like happiness and the need for larger datasets to improve model generalizability [8].

III. METHODOLOGY :

Fig 1. Methodology



3.1 DATASET

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset is a well-structured resource commonly used for emotion recognition research. It contains audio and video recordings of 24 professional actors (12 male and 12 female) performing emotional speech and song. The dataset includes 1440 recordings, with emotions such as neutral, calm, happy, sad, angry, fearful, surprise, and disgust. For splitting the RAVDESS dataset into training and testing sets for CNN and RNN algorithms, the data is typically divided into an 80-20 or 70-30 ratio, ensuring a balanced distribution of actors and emotions in both sets. This dataset's inclusion of professional actors ensures high-quality, emotionally consistent recordings, making it ideal for training deep learning models for speech emotion recognition tasks.

Fig 2. Ravdess Dataset

3.2. PROPOSED METHODOLOGY

In this proposed model CNN and RNN algorithms are used to recognition the emotion.

3.2.1. Convolutional Neural Network

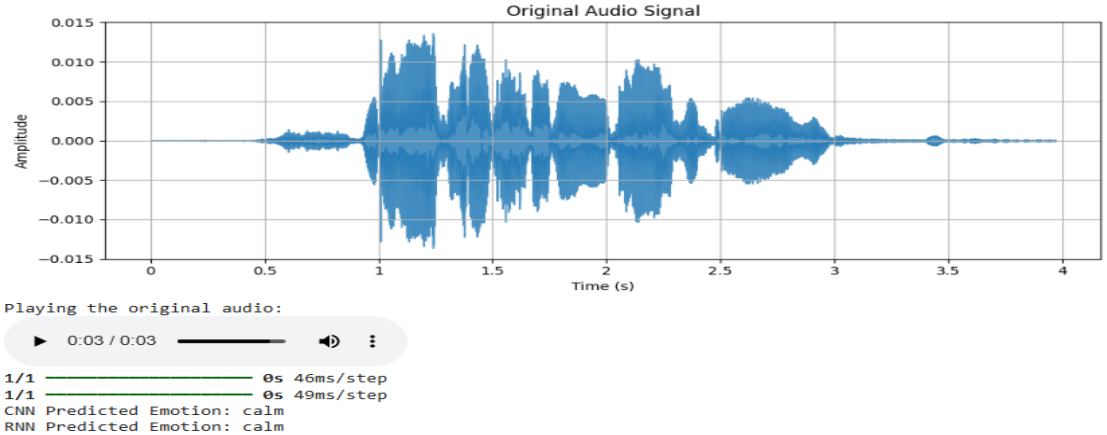
Convolutional Neural Networks (CNNs) play a crucial role in Speech Emotion Recognition (SER) by extracting spatial and spectral features from audio representations. In SER, raw audio signals are typically transformed into spectrograms or Mel-frequency cepstral coefficients (MFCCs), which provide a time-frequency representation of the speech. These representations are fed into the CNN, where convolutional layers apply filters to detect patterns such as pitch, tone, and energy variations that are crucial for distinguishing emotions. Activation functions like ReLU introduce non-linearity, enabling the network to learn complex emotional cues. Pooling layers further reduce the dimensionality, retaining essential features while reducing computational overhead. This hierarchical feature extraction allows CNNs to effectively capture the intricate spectral characteristics of speech, making them highly effective for recognizing emotions in audio data.

3.2.2. Recurrent Neural Network

Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, are highly effective for Speech Emotion Recognition (SER) due to their ability to model sequential and temporal data. Unlike CNNs, RNNs process speech signals as time-series data, capturing the dynamic changes in pitch, tone, and rhythm that convey emotions. LSTMs address the limitations of traditional RNNs by using memory cells to retain long-term dependencies, ensuring that both past and current context in speech are considered. This capability allows RNNs to understand the flow and progression of emotions throughout an utterance, making them ideal for recognizing complex and evolving emotional states in speech. By leveraging temporal dependencies, RNNs excel at capturing the contextual nuances essential for accurate emotion classification.

IV. EXPERIMENTAL RESULTS :

Fig 3. Emotion : Calm



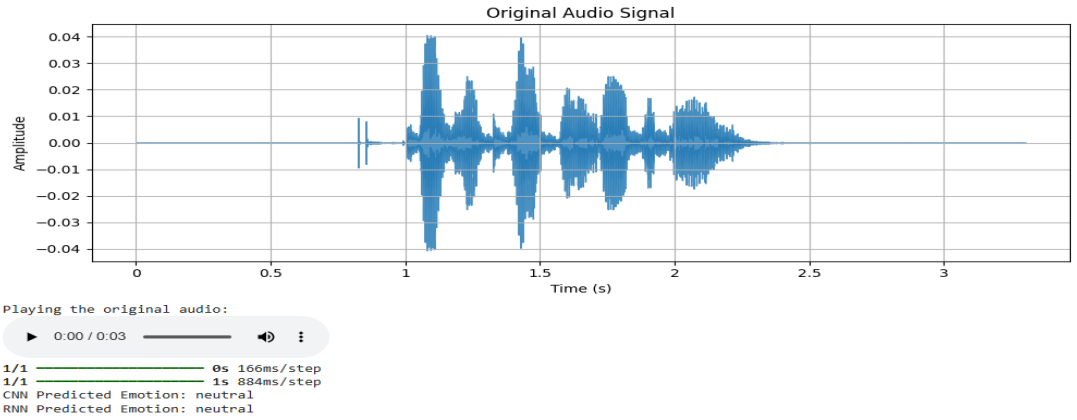


Fig 4. Emotion : Neutral

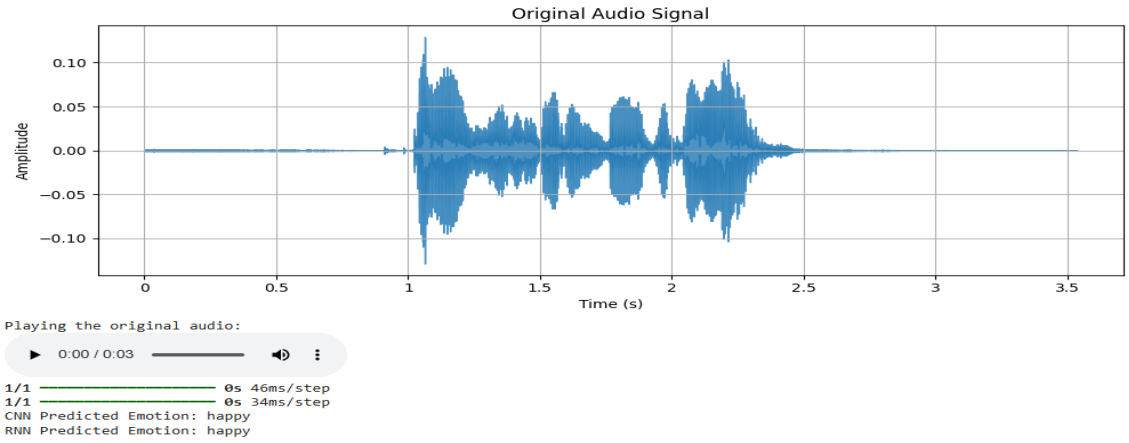


Fig 5. Emotion : Happy

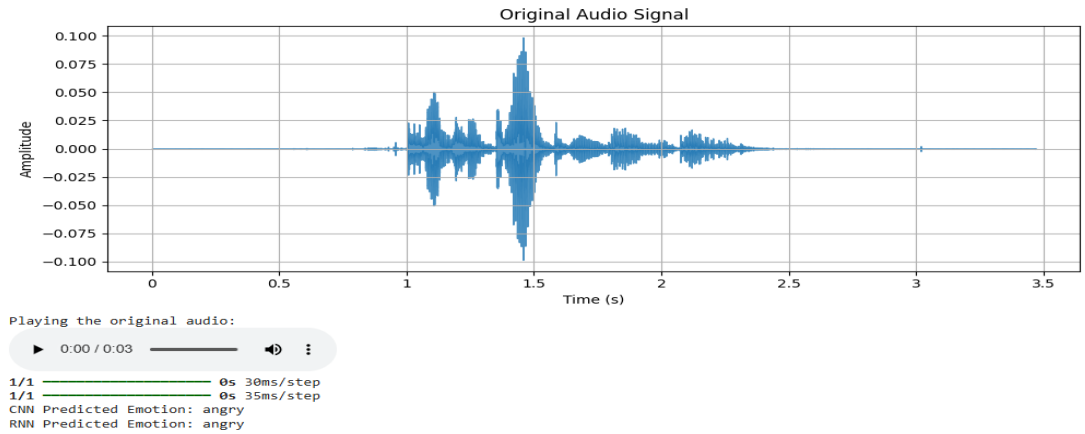


Fig 6. Emotion : Angry

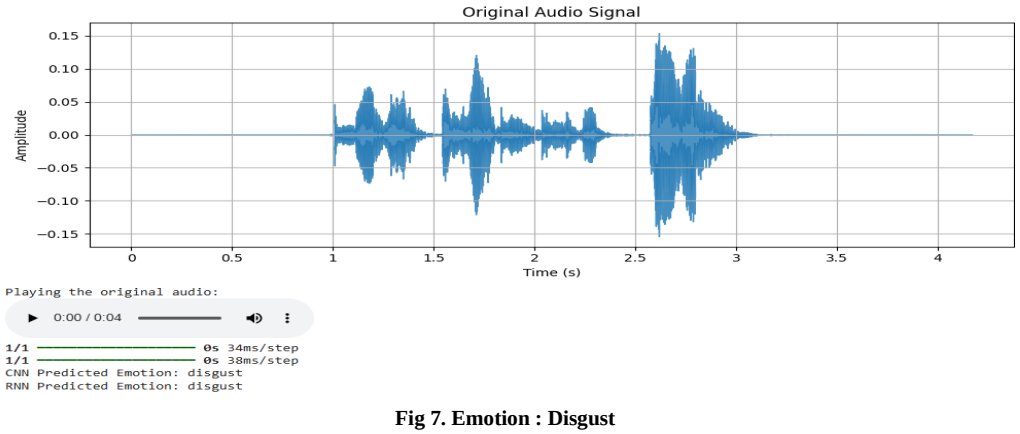


Fig 7. Emotion : Disgust

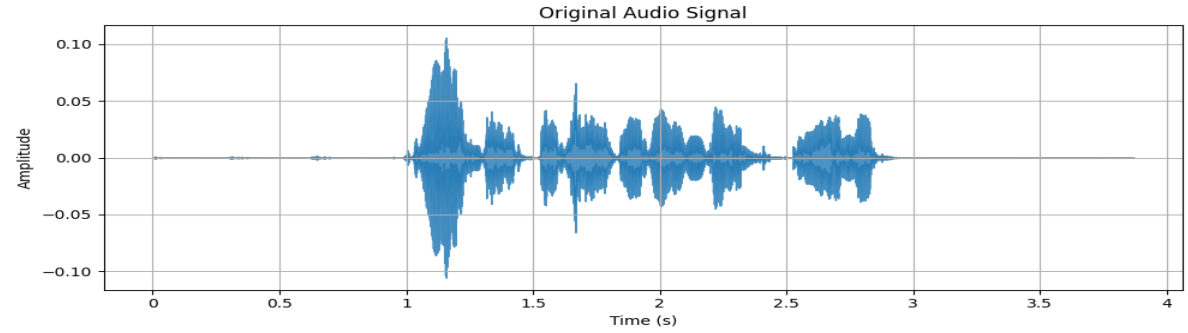


Fig 8. Emotion : Sad

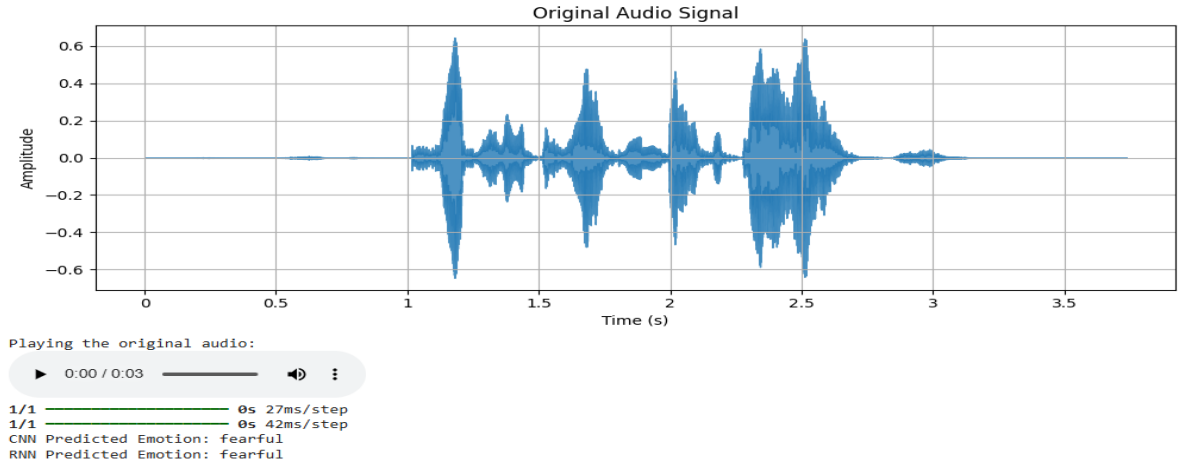


Fig 9. Emotion : Fearful

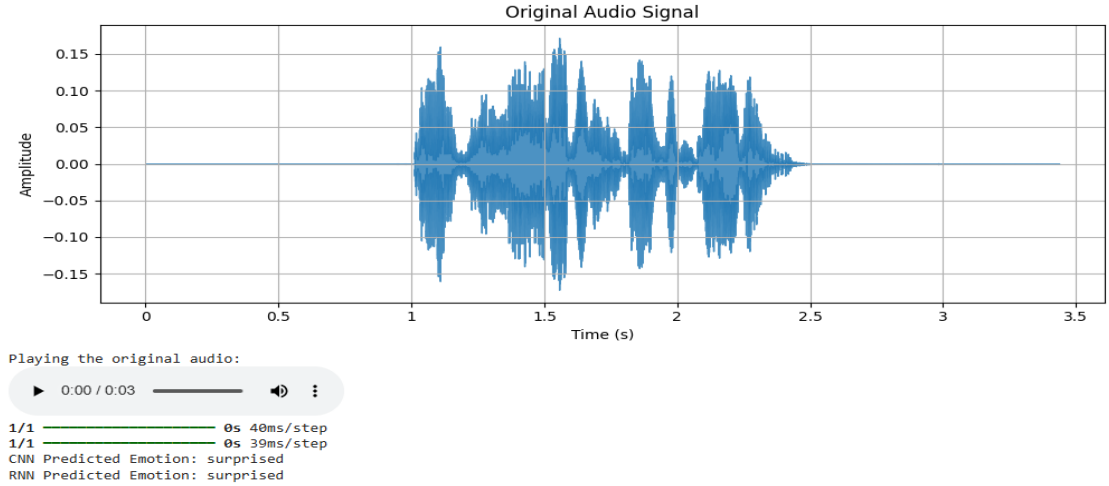


Fig 10. Emotion : Surprised

Table 1 Accuracy of models

S.NO	Algorithm	Accuracy
1	CNN	93%
2	RNN	86%

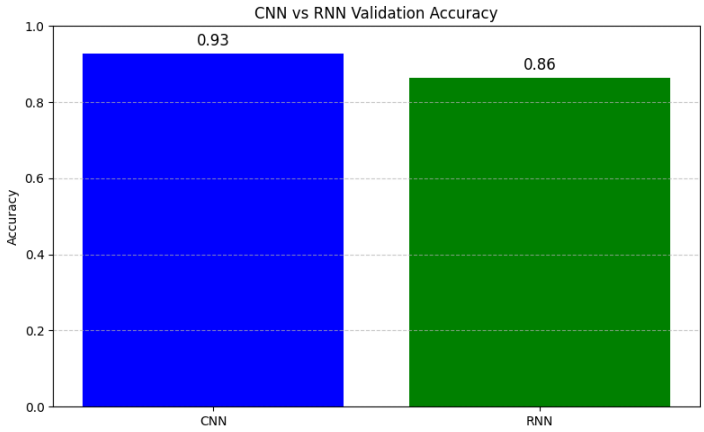


Fig 11. Accuracy of models

V. CONCLUSION :

The findings of this study highlight the effectiveness of deep learning architectures in Speech Emotion Recognition. Both CNN and RNN models demonstrated strong performance in predicting emotions such as Happy, Sad, Angry, Calm, Neutral, Fearful, Disgust, and Excitement. CNNs proved particularly adept at capturing spatial and spectral features from audio spectrograms, leveraging their ability to analyze visual patterns effectively. On the other hand, RNNs, especially LSTMs, excelled in modeling temporal dependencies, providing a robust understanding of sequential patterns in speech data. The integration of these models with the RAVDESS dataset enabled accurate recognition of emotional cues, showcasing the adaptability of the architecture. While both approaches contribute unique strengths, their combined insights reinforce the potential of deep learning in enhancing human-computer interaction by enabling systems to respond intelligently to human emotions. CNN implementation on the RAVDESS dataset demonstrated superior accuracy in Speech Emotion Recognition (SER) by effectively extracting intricate audio features than RNNs (LSTMs).

VI. REFERENCES :

1. Jermsittiparsert, K., Abdurrahman, A., Siriattakul, P., Sundeeva, L. A., Hashim, W., Rahim, R., & Maseleno, A. (2020). Pattern recognition and features selection for speech emotion recognition model using deep learning. *International Journal of Speech Technology*, *23*, 799-806. [Google Scholar]
2. Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. *IEEE access*, *9*, 47795-47814. [Google Scholar]
3. Jain, M., Narayan, S., Balaji, P., Bhowmick, A., & Muthu, R. K. (2020). Speech emotion recognition using support vector machine. *arXiv preprint arXiv:2002.07590*. [Google Scholar]
4. T. Sivanagaraja, M. K. Ho, A. W. H. Khong and Y. Wang, "End-to-end speech emotion recognition using multi-scale convolution networks," *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Kuala Lumpur, Malaysia, 2017, pp. 189-192 [Google Scholar]
5. Choudhary, R. R., Meena, G., & Mohbey, K. K. (2022, March). Speech emotion based sentiment recognition using deep neural networks. In *Journal of physics: conference series* (Vol. 2236, No. 1, p. 012003). IOP Publishing. [Google Scholar]
6. *Mustaqeem, & Kwon, S.* (2019). A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, *20*(1), 183. [Google Scholar]
7. Deng, J., Xu, X., Zhang, Z., Frühholz, S., & Schuller, B. (2017). Semisupervised autoencoders for speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(1), 31-43. [Google Scholar]
8. Araño, K. A., Gloor, P., Orsenigo, C., & Vercellis, C. (2021). When old meets new: emotion recognition from speech signals. *Cognitive Computation*, *13*(3), 771-783 [Google Scholar]