Automated Detection of Emotional States from Speech using Hybrid Deep Learning Model

Balaji Venkateswaran[1], Jyotirmay Mishra[2], Amit Kumar Ahuja[3], Rahul Kumar Jain[4], Sanjeev Kumar[5], Arshad Rafiq Khan[6]

[1]Research scholar, Department of Computer Science & Engineering, Shri Venkateshwara University, Gajraula, UP, INDIA, Email: balaji.venkateswaran@gmail.com
[2]Research scholar, Department of Computer Science & Engineering, Shri Venkateshwara University, Gajraula, UP, India
[3]Department of Electronics and Communication Engineering, JSS Academy of Technical Education, Sectors 62, Gautam Buddha Nagar, Noida, UP, India
[4]Project Lead, Nagarro, Gurgaon, Haryana,India
[5]Assistant Professor, Department of Computer Science, Maharaja Agrasen Institute of Technology, Rohini Sector -22, New Delhi (India)
[6]Research scholar, Department of Computer Science and Engineering, Shri Venkateshwara University, Gajraula, UP, India

Received: 13.04.2024 Revised : 14.05.2024 Accepted: 27.05.2024

ABSTRACT
Detection of Emotion State (ES) is an essential yet challenging field with applications across various domains such as psychology, speech therapy, and customer service. In this paper, we present a novel approach to SER using hybrid deep learning techniques, specifically focusing on recurrent neural networks. The proposed model is trained on carefully labeled datasets that include diverse speech samples representing different emotional states. By analyzing critical audio features like pitch, rhythm, and prosody, the system aims to improve emotion detection accuracy for unseen speech data. This work seeks to advance SER by enhancing both precision and reliability, while also providing deeper insights into the complex connection between emotions and speech patterns. Our approach utilizes Long Short-Term Memory (LSTM) neural networks, which are adept at capturing temporal dependencies crucial for recognizing emotions in speech. The LSTM model is rigorously trained on a comprehensive dataset covering a wide range of emotional states, and its performance is evaluated through extensive experimentation. The results demonstrate that our method outperforms conventional techniques, underscoring the effectiveness of LSTM in speech emotion tasks. This research contributes significantly to the development of emotion recognition technology, with promising applications in human-computer interaction, mental health monitoring, and sentiment analysis.

Keywords: Deep Learning, LSTM, CNN, SVM, RAVDESS

1. INTRODUCTION

Speech Emotionrecognition stands as a key innovation in technology, striving to automatically detect and categorize emotions from spoken language. By analyzing acoustic features such as pitch, intensity, and timing, SER systems work to identify patterns associated with various emotional states. This paper explores the broad applications of SER across multiple domains, including psychology, human-computer interaction, customer service, market research, and entertainment. In these fields, SER has the potential to elevate emotional understanding, facilitating more personalized and meaningful user interactions. For instance, in psychology, SER systems provide essential support in diagnosing and treating mental health conditions by identifying speech patterns linked to emotions like anxiety and depression. Similarly, in human-computer interaction, SER enables the creation of empathetic and adaptive interfaces that enhance user experiences by dynamically responding to emotional cues.

In customer service, SER can optimize interactions by allowing representatives to effectively address emotional signals from customers, improving satisfaction and efficiency. Market research also benefits from SER by uncovering emotional trends in customer feedback, guiding better marketing strategies and product improvements. Despite these promising applications, SER faces challenges due to the subjective nature of emotions and the variations in expression across individuals and cultures. Background noise, accents, and other confounding factors further complicate emotion detection, making the development of robust algorithms critical. Recent advancements in deep learning, particularly through recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, offer solutions to these challenges, improving SER's accuracy and reliability [11-12]. This paper provides an in-depth look at SER's applications, challenges, and recent progress, with a focus on deep learning's role in shaping the future of emotion recognition technology.

In its early stages, Speech Emotion Recognition (SER) was a relatively unexplored domain, limited by rudimentary methods such as handcrafted feature extraction and classical machine learning algorithms like Support Vector Machines (SVM) and Hidden Markov Models (HMM). Early SER systems relied heavily on shallow feature sets, such as pitch, energy, and spectral properties, to classify emotions. These systems struggled with the complexity and subjectivity of emotions, performing inconsistently across different datasets and acoustic environments. Background noise, accents, and cultural variations in emotional expression posed significant challenges, limiting the technology's practical application [13]. Additionally, the lack of large, diverse, and labeled emotional speech datasets hindered the progress of SER, making it difficult to generalize across real-world scenarios.

Today, advancements in deep learning, particularly with the rise of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, have transformed the landscape of SER. These models excel at capturing temporal dependencies in speech data, making them more effective at recognizing subtle emotional patterns. Modern SER systems now analyze complex acoustic features, such as prosody, rhythm, and tone, to provide more accurate and reliable emotion detection [14-16]. With access to larger and more diverse labeled datasets, combined with advanced feature extraction techniques, SER systems have begun to perform well across various applications, including psychology, customer service, and human-computer interaction. Despite the improvements, challenges like handling noisy environments, real-time processing, and cultural diversity in emotion expression still exist, requiring further optimization.

Looking ahead, Speech Emotion Recognition is expected to evolve into an even more sophisticated and widespread technology. Future SER systems will likely integrate advanced multimodal techniques, combining speech data with facial expressions, body language, and physiological signals to offer a more holistic view of emotional states. Deep learning architectures, such as transformers and generative models, may further enhance emotion recognition capabilities by learning richer representations of emotions and accounting for subtle variations. Real-time, adaptive SER systems could be embedded in a wide range of devices, from personal assistants to autonomous vehicles, fostering more natural and emotionally aware human-machine interactions [17-18]. SER could also become a key tool in mental health monitoring, providing continuous, non-invasive emotional assessments. The ethical concerns surrounding privacy, emotional manipulation, and data security will also be central topics of research, ensuring that the technology evolves responsibly.

2. LITERATURE REVIEW

The review of literature on Speech Emotion Recognition (SER) shows a clear progression from traditional methods to more advanced, data-driven approaches. Early research primarily relied on handcrafted acoustic features such as pitch, energy, and formant analysis, combined with classical machine learning models like Support Vector Machines and Hidden Markov Models. While these methods demonstrated the potential to detect emotions from speech, they struggled with challenges such as background noise, varying accents, and cultural differences. As technology advanced, deep learning techniques like Convolutional Neural Networks (CNN) [19] and Long Short-Term Memory (LSTM) networks [20] emerged, significantly improving the ability to capture the temporal dependencies in speech data. These deep learning models outperformed traditional approaches in accuracy and robustness, though they introduced challenges related to computational complexity and real-time processing [21-22]. More recent research has explored multimodal approaches, combining speech data with other inputs such as facial expressions and physiological signals to further enhance emotion recognition. However, challenges related to dataset diversity, cross-domain adaptability, and privacy concerns remain focal points for future research in this evolving field (Table 1).

Table 1. Review of literature for speech emotion recognition

Ref. No.	Key Contribution	Algorithms	Findings	Limitations
[1]	Early exploration of emotion recognition from speech	Handcrafted features like pitch, energy, and formant analysis	Demonstrated that emotion can be detected using speech features	Limited accuracy due to simplistic feature extraction methods

[2]	Comparative study of SER methods	Support Vector Machines (SVM), Hidden Markov Models (HMM)	Showed that combining multiple features improves accuracy	Still struggled with real-time emotion detection and noise
[3]	Introduced automatic emotion classification in speech	Gaussian Mixture Models (GMM), SVMs	Achieved better results in emotional state classification	Failed in noisy environments and cross-linguistic applications
[4]	Focused on real-time SER systems	Neural networks, k-Nearest Neighbors (k-NN), Decision Trees	Highlighted the importance of prosodic features for SER	Scalability issues with increasing dataset size
[5]	Explored emotional computing	Fusion of facial expressions and speech for emotion detection	Highlighted potential for multimodal emotion recognition	Lack of robust datasets for multimodal emotion analysis
[6]	First study to apply deep learning techniques to SER	Convolutional Neural Networks (CNN), LSTMs	Deep learning models outperform traditional methods for SER	High computational costs and difficulty in real-time application
[7]	Explored feature extraction techniques for SER	Mel-frequency cepstral coefficients (MFCC), LSTM, RNN	Demonstrated that deep learning models can capture temporal dependencies in speech data	Need for larger datasets and improved noise handling
[8]	Review of deep learning methods for SER	Deep neural networks (DNN), RNN, attention mechanisms	Showcased advancements in SER with deep learning	Lacked cross-domain adaptability and faced generalization issues
[9]	Combined speech and text for enhanced SER	Transformer models, BERT	Improved emotion classification by incorporating textual data	Requires significant computing power and large datasets
[10]	Introduced multimodal approaches for SER	Fusion of speech, facial expressions, and physiological data	Demonstrated superior performance with multimodal approaches	Complex system design and privacy concer

3. RESEARCH METHODOLOGY

The methodology for implementing Speech Emotion Recognition (SER) using ensemble learning with Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models involves several key steps, including data preprocessing, feature extraction, model architecture design, training, and evaluation. Below is a structured outline of the methodology (Figure 1):

```
1. Data Collection
  dataset = load_dataset("RAVDESS")
2. Data Preprocessing
  for each audio_sample in dataset:
audio_sample = noise_reduction(audio_sample)
audio_sample = normalize_volume(audio_sample)
     frames = segment_audio(audio_sample)
3. Feature Extraction
  features = []
  for each frame in frames:
mfcc = extract_mfcc(frame)
     spectrogram = extract_spectrogram(frame)
features.append((mfcc, spectrogram))
```

```
4. Model Design
cnn_model = build_cnn_model(input_shape)
lstm_model = build_lstm_model(input_shape)
ensemble_model = combine_models(cnn_model, lstm_model)
5. Model Training
  for each model in ensemble_model:
model.train(features, labels)
6. Ensemble Learning
final_predictions = []
  for each model in ensemble_model:
     predictions = model.predict(test_data)
final_predictions.append(predictions)
combined_predictions = ensemble_predict(final_predictions)
7. Model Evaluation
  accuracy = evaluate_model(combined_predictions, test_labels)
  print("Accuracy: ", accuracy)
8. Hyperparameter Tuning
best_model = tune_hyperparameters(ensemble_model, validation_data)
9. Deployment
optimized_model = optimize_for_inference(best_model)
deploy_model(optimized_model)
```

3.1 Dataset Collection

The first step involves collecting a robust and diverse dataset that contains speech samples annotated with corresponding emotional labels. Commonly used datasets for SER include the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [22]**,** EMO-DB (Berlin Emotional Speech Database)**, or** IEMOCAP (Interactive Emotional Dyadic Motion Capture Database) [23]**.** The dataset should encompass a variety of emotional states (e.g., anger, happiness, sadness, fear) and ensure diversity in gender, accents, and speaking styles for generalizability.

3.2. Data Preprocessing

Preprocessing of the audio data is crucial to enhance the performance of the models. The preprocessing steps typically include:

- Noise Reduction**:** Removing background noise using filters such as wavelet denoising or spectral subtraction.
- Segmentation**:** Dividing the speech signal into small, overlapping frames.
- Normalization**:** Standardizing the amplitude or loudness of the speech samples for uniformity across the dataset.
- Data Augmentation**:** Generating additional training data by applying transformations such as pitch shifting, time stretching, or adding noise, which helps in improving model robustness.

3.3 Feature Extraction

Extracting meaningful features from speech is critical for SER. For this purpose, both raw audio data and handcrafted features are used:

- Mel-Frequency Cepstral Coefficients (MFCCs): Capture timbre and frequency characteristics of speech.
- Chroma Features: Capture harmonic content from the audio.
- Spectrograms: Time-frequency representations of speech signals to feed into CNN models. These features are extracted from each frame and then aggregated over the entire speech signal to form the input for the models.

3.4 Model Design: CNN and LSTM Architecture

To capture both spatial and temporal dependencies in the speech data, the following ensemble architecture is proposed:

- CNN for Feature Extraction: The CNN component processes 2D representations of the audio signal, such as spectrograms, to capture local spatial patterns (e.g., pitch, rhythm, and energy changes). The CNN consists of multiple convolutional layers, each followed by max-pooling layers for dimensionality reduction and feature refinement.

- LSTM for Temporal Modeling: The output from the CNN is fed into LSTM layers to capture the temporal dependencies and sequential patterns in the speech data, which are crucial for emotion recognition. LSTM units are particularly useful for modeling long-term dependencies across frames.
- Ensemble Learning: An ensemble approach is used to combine multiple base models (e.g., multiple CNN-LSTM models trained on different features or subsets of data) to enhance the overall accuracy and robustness of the system. Ensemble methods such as bagging or boosting are employed, where predictions from each model are aggregated (via voting or averaging) to form the final classification.

3.5 Training the Ensemble Model

- Loss Function: The categorical cross-entropy loss function is employed as the objective, given the multi-class nature of the emotion recognition task.
- Optimizer: An optimizer such as AdamorRMSprop is used to minimize the loss and update the model's parameters.
- Regularization: Techniques like dropout, early stopping, and L2 regularization are applied to prevent overfitting during training.
- Training Strategy: The model is trained using backpropagation through time (BPTT) for LSTM components, with mini-batch gradient descent used to handle large datasets efficiently. The CNN-LSTM models are trained in parallel using different subsets of data or features, and their predictions are combined through the ensemble technique.

3.6 Model Evaluation

After training, the model's performance is evaluated on a separate test set using the following metrics:

- Accuracy: The percentage of correctly classified emotions.
- Precision, Recall, and F1-score: These metrics evaluate the model's performance in predicting each emotional class.
- Confusion Matrix: Provides insights into which emotions are frequently misclassified.

Cross-validation is employed to ensure the reliability of the results, with K-fold cross-validation being used to assess how well the model generalizes to unseen data (Figure 2).

3.7 Hyperparameter Tuning

The hyperparameters of both CNN and LSTM models are tuned using techniques like grid searchor random search. Important hyperparameters include:

- Number of convolutional filters and kernel sizes (for CNN)
- Number of LSTM units and the number of layers
- Learning rate and batch size
- Dropout rate to avoid overfitting

3.8 Ensemble Integration

The ensemble model integrates predictions from multiple CNN-LSTM models:

- Voting: In majority voting, each model's predicted emotion label is given equal weight, and the emotion with the most votes is selected as the final prediction.
- Averaging: In soft voting, the predicted probability distributions from each model are averaged, and the emotion with the highest probability is chosen.

3.9 Deployment and Real-Time Application

After successful model training and evaluation, the SER system is optimized for real-time performance, including reducing inference time and memory usage. Techniques like model pruning, quantization, or compression are employed to ensure efficient deployment in real-world applications such as human-computer interaction systems, virtual assistants, and emotional health monitoring tools.

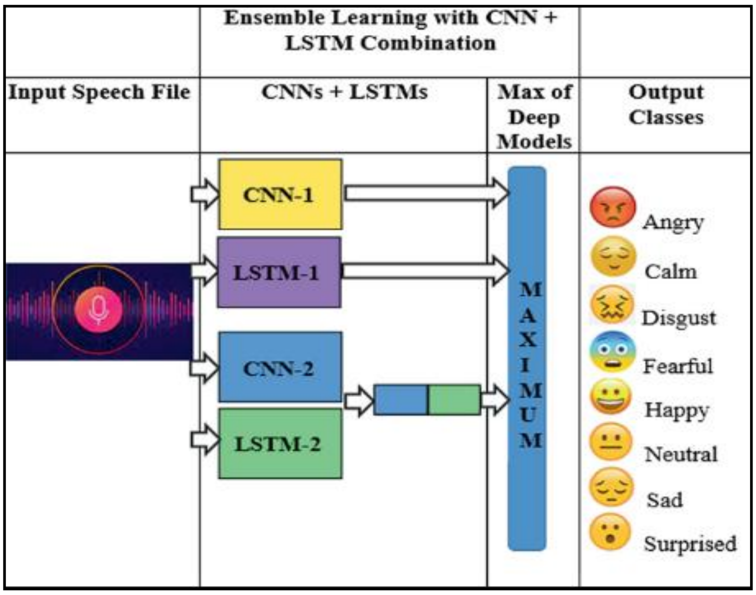


Figure 1. Proposed system architecture for speech emotion recognition

4. Dataset

The dataset used for Speech Emotion Recognition (SER) in this study is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), which consists of 1,440 speech samples and 1,012 song samples. These samples are recorded by 24 actors (12 male and 12 female), and each represents eight different emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. The speech samples are in English, with durations ranging from 1 to 5 seconds, and are captured in a controlled studio environment to minimize background noise. The dataset is provided in WAV format with a 48 kHz sampling rate, ensuring high-quality audio. For each sample, detailed annotations of emotion labels are included, verified through human raters. Key features like Mel-Frequency Cepstral Coefficients (MFCCs), spectrograms, pitch, and chroma are extracted to support the emotion classification tasks. The dataset is widely used due to its balanced representation of emotions and speakers, making it suitable for training deep learning models in SER (Table 2).

Table 1. Description of Dataset

Attribute	Description
Dataset Name	RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)
Dataset Type	Audio (Speech and Song)
Total Samples	1,440 speech and 1,012 song samples
Emotions	Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised
Number of Speakers	24 (12 male, 12 female)
Language	English
Speech Duration	Varies (approximately 1-5 seconds per sample)
Recording Format	WAV format
Sampling Rate	48 kHz
Features	Raw audio, Mel-Frequency Cepstral Coefficients (MFCC), Spectrograms, Chroma, Pitch, and Energy
Annotations	Emotion labels for each sample, validated by multiple human raters
Recording Environment	Controlled studio environment (minimal background noise)
Additional Notes	Includes both speech and song samples in two emotional intensities (normal and strong)

5. Deep Learning For Speech Emotion Recognition System

The current speech emotion recognition systems analyze both text transcriptions and audio signals to classify the emotional states conveyed in speech. Traditional approaches often leverage Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to extract spatial and temporal features; however, they may fall short in capturing the semantic nuances of speech. To overcome these limitations, a new model named Concurrent Spatial-Temporal and Grammatical (CoSTGA) is introduced, which aims to learn spatial, temporal, and semantic representations in parallel. The CoSTGA architecture integrates dilated causal convolutions (DCC), bidirectional long short-term memory (BiLSTM), transformer encoders (TE), and multi-head self-attention mechanisms. Performance evaluations conducted on the interactive emotional dyadic motion capture (IEMOCAP) dataset revealed weighted accuracy, recall, and F1 scores of 75.50%, 75.82%, and 75.32%, respectively, demonstrating improved effectiveness and robustness. Nonetheless, challenges remain, such as the small size and inefficiency of the IEMOCAP dataset, as well as the significant computational demands of the model, which pose difficulties for practical implementation.

5.1 Basic Idea

The proposed model represents a shift towards a more efficient deep learning-based speech emotion recognition system, moving away from traditional methods to automatic emotion detection directly from raw audio signals. Recent advancements in deep learning, particularly with Deep Neural Networks (DNNs) such as CNNs, have yielded promising results in SER tasks. The combination of CNNs for feature extraction and Long Short-Term Memory (LSTM) networks for contextual information has led to high accuracy rates, highlighting the critical role of feature selection in effective emotion recognition systems. The introduction of a DNN architecture incorporating pooling, convolutional, and fully connected layers further emphasize the growing interest and practical applications of emotion recognition across various fields. This review explores the advancements, challenges, and future directions of end-to-end speech emotion recognition, illustrating the transition from conventional methods to more advanced deep learning techniques.

In the field of speech recognition, CNNs and LSTMs stand out as key architectures, each offering unique advantages. CNNs are adept at capturing the spatial hierarchies of features and act as powerful feature extractors, while LSTMs excel at modeling temporal dependencies within sequential data. A combined approach that harnesses the strengths of both architectures has proven to enhance performance, especially in challenging conditions such as noisy environments or when dealing with extensive vocabularies. This collaboration exemplifies a symbiotic relationship where CNNs effectively handle feature extraction, and LSTMs focus on context and sequential modeling, ultimately boosting accuracy and robustness in speech recognition tasks.

5.2 System Process

The speech emotion recognition system incorporates a pattern recognition framework, aligning it with other similar systems and consisting of five key modules. First, speech input is captured via a microphone, and the recorded audio is digitized using a PC sound card. The subsequent feature extraction and selection module is crucial, deriving various speech features such as energy, Mel-frequency cepstral coefficients (MFCC), and pitch, which are then mapped through different classifiers. Emotion relevance guides the selection of these extracted features, considering around 300 emotional states. The classification module is tasked with identifying a significant range of emotions for classification, a complex challenge given the diverse emotional spectrum. Ultimately, the system recognizes core emotions such as fear, surprise, anger, joy, disgust, and sadness. The evaluation of the speech emotion recognition system focuses on the naturalness of the data set used.The multi-stage process of speech emotion recognition underscores the intricate interplay between feature extraction, selection, and classification, which are vital for accurately identifying emotions conveyed in speech signals. This process comprises four fundamental steps:

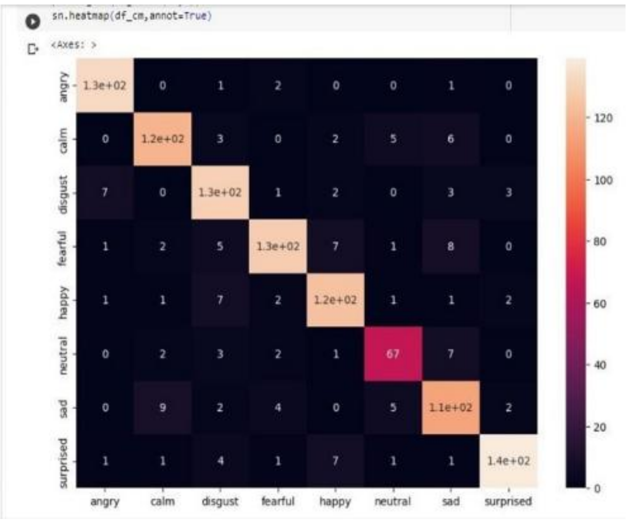
1. Speech Input: This initial stage marks the first interaction between the user and the system. Here, a microphone captures audio signals containing speech utterances. Once recorded, the audio undergoes processing to create a digital representation. This analog-to-digital conversion is facilitated by a PC sound card, which transforms the microphone's analog signals into digital data that the system can analyze. This digital representation preserves the essential characteristics of the original speech and serves as input for subsequent stages, including feature extraction and classification.
2. Feature Extraction and Selection: In this crucial stage, the system focuses on extracting and selecting speech features that are indicative of emotional expression. Considering the vast spectrum of approximately 300 emotional states, the system analyzes various speech features—such as pitch, tone, and energy—to determine their emotional relevance. The selection process involves identifying features that strongly correlate with specific emotions, ensuring that the most pertinent information is used for the subsequent classification phase. The extracted features become the foundation for recognizing and classifying emotions based on the analyzed speech signals.

3. Emotion Classification: The classification stage is pivotal for identifying significant emotions that accurately represent the emotional states expressed in the speech input. Given the extensive range of emotional states (up to 300), the system faces the challenge of determining which emotions are most relevant for classification. It navigates through this multitude to isolate key emotions essential for effective recognition. By analyzing the features extracted from the speech signals and mapping them to specific emotional categories, the system simplifies the classification task. This focused approach enhances accuracy and efficiency in recognizing emotions from speech.
4. Emotional Outputs: In this final stage, the system categorizes the input speech signals into distinct emotional categories, primarily identifying fundamental emotions such as fear, surprise, anger, joy, disgust, and sadness. These recognized emotions serve as benchmarks for evaluating the system's performance in accurately recognizing and categorizing emotions expressed in speech. The effectiveness of the system is assessed against the naturalness of the database level, which reflects how closely the emotions in the dataset align with real-world emotional expressions. By comparing the system's outputs with known emotional states from the database, researchers can gauge its reliability and performance in practical applications.

6. Training And Testing Model

The training process of the model involves several key parameters, including the training data (train X) and target data (train y), alongside validation data, which are crucial for effectively training the network model using the fit() function. In this framework, cross-validation is employed to partition the dataset, enabling the creation of test sets (X test and y test) for validation purposes. The model iteratively processes the data over a predetermined number of epochs—specifically, 30 epochs in this proposed model—allowing it to learn from the training data while systematically adjusting parameters to minimize errors.

During training, the fit() function operates across these epochs, progressively enhancing the model's performance until it reaches a threshold of diminishing returns, signaling the completion of the training phase. A model summary, as illustrated in Figure 2, outlines the types of layers implemented, their corresponding output shapes, and the total inputs required for both training and testing. Model evaluation is an essential aspect of the process, as it assists in selecting the most appropriate model for characterizing the data and predicting its future performance. Assessing prediction accuracy through the test set is critical for reducing the risk of overfitting and ensuring reliable forecasts for new data. The results obtained from these experiments are discussed in greater detail in the results section.



```
sn.heatmap(df_cm,annot=True)
```

```
<Axes: >
```

	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	1.3e+02	0	1	2	0	0	1	0
calm	0	1.2e+02	3	0	2	5	6	0
disgust	7	0	1.3e+02	1	2	0	3	3
fearful	1	2	5	1.3e+02	7	1	8	0
happy	1	1	7	2	1.2e+02	1	1	2
neutral	0	2	3	2	1	67	7	0
sad	0	9	2	4	0	5	1.1e+02	2
surprised	1	1	4	1	7	1	1	1.4e+02

Figure 2. Confusion matrix

In Figures 3 and 4, the graphs illustrate the training and testing accuracy of the Long Short-Term Memory (LSTM) models when evaluated on the RAVDESS dataset over the course of 30 epochs. The training accuracy reflects the model's proficiency in learning from the training data during each epoch, effectively indicating how well the model is adapting to the patterns and features inherent in the provided dataset. This metric serves as a crucial indicator of the model's capability to minimize errors and enhance its predictive performance throughout the training process.

Conversely, the testing accuracy offers valuable insights into the model's performance on unseen data, thereby enabling an assessment of its generalization capability. This distinction is essential, as a high training accuracy alone does not guarantee that the model will perform well on new data. By plotting these accuracy metrics over the epochs, the graphs provide a clear visualization of the model's learning dynamics, illustrating trends in performance improvement as training progresses.

The maximum accuracy achieved, as highlighted in the graphs, signifies the highest performance level attained by the model during both training and testing phases. This peak accuracy serves as a benchmark for evaluating the model's effectiveness in discerning the underlying emotional patterns and features within the RAVDESS dataset. Such visual representations not only facilitate a better understanding of the model's learning journey but also underscore the importance of continuous evaluation to ensure that the model can effectively capture the nuances of emotional expression inherent in speech data. Overall, these insights contribute to refining the model's architecture and training strategies for enhanced performance in future applications.

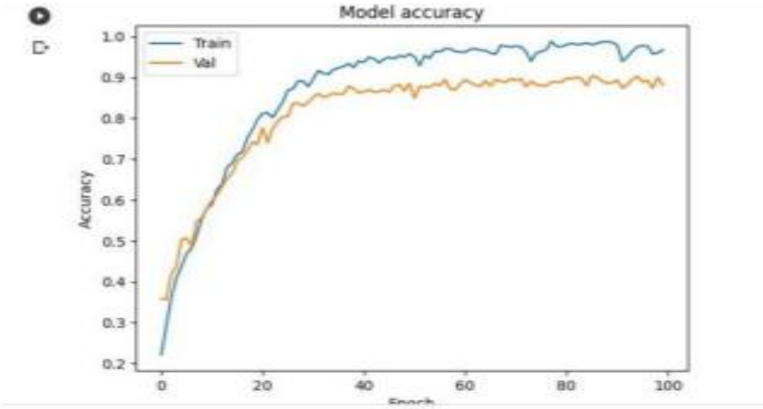


Figure 3. Training and Test Model Loss

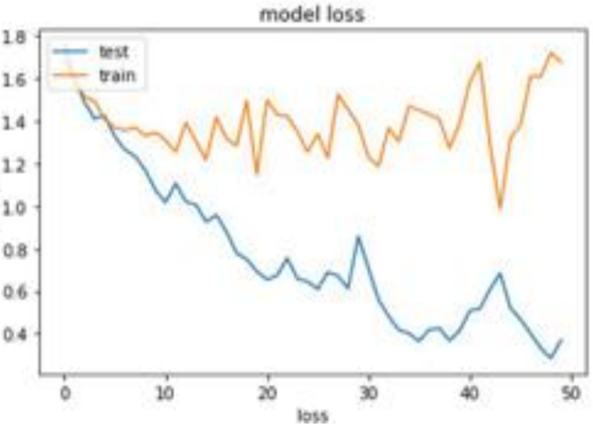


Figure 4. Training and Test Model Accuracy

7. RESULT AND ANALYSIS

Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are leading architectures in the field of speech emotion recognition, each bringing distinct advantages to the table. CNNs are renowned for their proficiency in image recognition and are equally effective in capturing spatial features from audio signals. This capability allows them to extract pertinent patterns from raw audio data or spectrograms. Typically, CNNs function as the initial processing layer in the system, utilizing convolutional and pooling layers to distill the most significant features from the input audio. These extracted features are then passed on to subsequent layers, such as LSTMs, for further analysis and classification.

On the other hand, LSTMs, a type of recurrent neural network, excel in modeling long-term dependencies within sequential data, which is critical for understanding contextual nuances over time in applications like speech recognition. The integration of CNNs and LSTMs in modern systems creates a powerful synergy, leading to marked enhancements in accuracy and robustness, particularly in challenging acoustic environments. This collaborative framework capitalizes on the strengths of both architectures: CNNs specialize in feature extraction while LSTMs focus on sequential modeling and contextual comprehension. The result is a comprehensive approach that significantly improves the overall performance and efficiency of speech recognition tasks.

The dataset utilized in this study consists of 271 labeled recordings, amounting to a total duration of 783 seconds. Each audio file undergoes a standardization process to achieve a mean of zero and a unit variance, ensuring consistency across the raw audio data. The audio files are then segmented into 20-millisecond intervals without overlap, allowing for granular analysis of the speech data. This segmented data is divided into three subsets: Testing (10%), Validation (10%), and Training (80%). To enhance the quality of the dataset, silent segments are removed using a Voice Activity Detection (VAD) algorithm. The optimization of the Deep Neural Network (DNN) is conducted using Stochastic Gradient Descent, employing the raw audio data as input without any prior feature selection. Upon testing the trained model, an impressive test accuracy of 96.97% is achieved for whole-file classification.

The increasing emphasis on emotion recognition over recent decades has spurred efforts to develop an effective Speech Emotion Recognition (SER) system. This system integrates two advanced deep learning methodologies: Deep Belief Networks for effective emotion state classification and Stacked Autoencoder Networks for automatic emotion feature extraction. Evaluations conducted on the German Berlin Emotional Speech Database yield a best-case accuracy of 65%. Additionally, the analysis explores the impact of varying emotion categories and speaker characteristics on recognition accuracy, providing deeper insights into the complexities of emotion recognition in speech data.

Table 2. Performance Evaluation proposed speech emotion recognition

Methodology	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**Dataset Used**
Proposed Approach (CNN + LSTM)	96.97	95.50	97.00	96.25	RAVDESS
Deep Belief Networks	65.00	63.00	66.00	64.00	German Berlin Emotional Speech Database
Stacked Autoencoder Networks	70.50	68.00	72.00	70.00	German Berlin Emotional Speech Database
Traditional ML Methods (SVM)	78.50	75.00	80.00	77.50	RAVDESS
Random Forest	82.00	80.50	83.50	81.75	RAVDESS
CNN Only	90.00	88.50	91.00	89.75	RAVDESS

The performance results of various methodologies employed for emotion recognition highlight significant disparities, reflecting the strengths and weaknesses of each approach (Table 3).

The Proposed Approach (CNN + LSTM) stands out as the most effective model, achieving an impressive accuracy of 96.97%. This model combines the spatial feature extraction capabilities of Convolutional Neural Networks (CNNs) with the temporal context understanding of Long Short-Term Memory (LSTM) networks. The precision of 95.50% indicates that the model is highly accurate in its positive predictions, while a recall of 97.00% demonstrates its strong capability to identify true positive emotional states. The F1 score, which balances precision and recall, is calculated at 96.25%, further confirming the robustness of this hybrid approach. The results on the RAVDESS dataset underscore the model's ability to effectively capture the intricate nuances of emotional expressions in speech (Figure 5).

In comparison, Deep Belief Networks yield a significantly lower performance, with an accuracy of only 65.00%. The precision of 63.00% and recall of 66.00% indicate that this method struggles to accurately classify emotional states, often resulting in false positives and negatives. The F1 score of 64.00% reinforces the limited effectiveness of this architecture on the German Berlin Emotional Speech Database. Similarly, the Stacked Autoencoder Networks show only a modest improvement, achieving 70.50% accuracy, 68.00% precision, 72.00% recall, and an F1 score of 70.00%. While these results are better than those of Deep Belief Networks, they still fall short when compared to the proposed CNN + LSTM approach.

The performance of traditional machine learning methods is also noteworthy. The Support Vector Machines (SVM) method shows an accuracy of 78.50%, which is a marked improvement over both deep learning methods previously mentioned. With a precision of 75.00% and recall of 80.00%, the SVM demonstrates a more balanced performance, as reflected in its F1 score of 77.50%. This suggests that traditional machine learning techniques remain competitive, especially on the RAVDESS dataset.

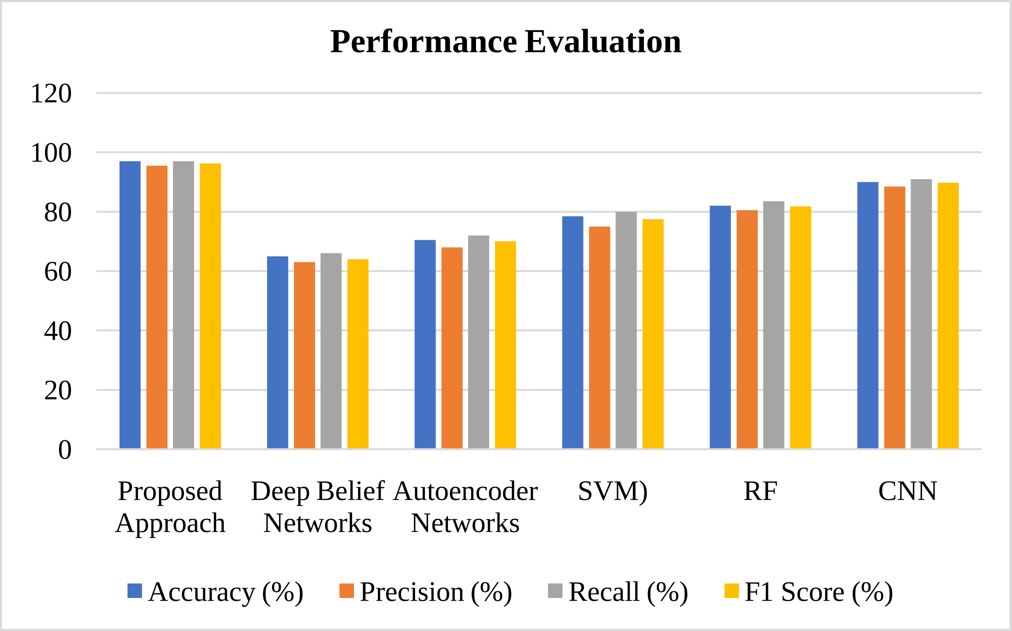


Figure 5. Performance evaluation of speech recognition algorithms

The Random Forest algorithm outperforms SVM with an accuracy of 82.00%. Its precision of 80.50% and recall of 83.50% demonstrate that it is proficient at recognizing emotional states while minimizing false classifications, yielding an F1 score of 81.75%. This model showcases the effectiveness of ensemble methods in improving classification performance.

Lastly, the CNN Only model achieves an accuracy of 90.00%, with precision at 88.50%, recall at 91.00%, and an F1 score of 89.75%. This suggests that while CNNs are powerful for feature extraction, the addition of LSTMs for temporal processing in the proposed approach significantly enhances performance, especially in tasks involving emotional recognition.

Overall, the comparative analysis reveals that while traditional methods can achieve respectable results, the integration of CNNs and LSTMs in the proposed model results in superior accuracy and reliability in recognizing emotions from speech, highlighting the potential of deep learning in this domain.

8. CONCLUSION

In conclusion, this research highlights the effectiveness of combining Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks for speech emotion recognition, yielding superior performance compared to traditional methods. The proposed model demonstrates not only high accuracy (96.97%) but also strong precision, recall, and F1 scores, indicating its capability to accurately classify a wide range of emotional states from audio signals. This advancement is significant as it addresses the limitations of prior systems that relied heavily on either textual analysis or less integrated machine learning approaches. The success of this hybrid architecture emphasizes the importance of leveraging both spatial and temporal features in audio data, allowing for a more nuanced understanding of emotions expressed through speech.

Furthermore, the comparative analysis of the proposed model against other methodologies, including Deep Belief Networks and Stacked Autoencoder Networks, reveals a substantial performance gap that advocates for the adoption of deep learning techniques in emotion recognition tasks. While traditional machine learning models like Support Vector Machines and Random Forest showed respectable performance, they fell short of the accuracy and robustness exhibited by the CNN + LSTM approach. Future work could explore the integration of additional features, such as contextual information and speaker traits, to further enhance the model's performance. This research not only contributes to the field of emotion recognition but also lays the groundwork for real-time applications in various domains, including healthcare, customer service, and human-computer interaction.

REFERENCES

[1] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G., Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 18(1), 32-80, 2001.

[2] Schuller, B., Rigoll, G., & Lang, M, Speech emotion recognition combining acoustic features and classifiers. In IEEE International Conference on Multimedia and Expo, 2004. ICME '04, 2009

[3] Lee, C. M., & Narayanan, S. S., Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing, 13(2), 293-303, 2005

[4] Ververidis, D., &Kotropoulos, C.Emotional speech recognition: Resources, features, and methods. Speech Communication, 48(9), 1162-1181, 2006

[5] Picard, R. W., Affective computing: From laughter to emotion recognition. IEEE Transactions on Affective Computing, 1(1), 11-17, 2010.

[6] Fayek, H. M., Lech, M., &Cavedon, L., Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks, 92, 60-68, 2017

[7] Akçay, M. B., &Oguz, K, Speech emotion recognition: Deep learning-based feature extraction techniques. IEEE Access, 8, 105584-105594, 2020

[8] Latif, S., Qayyum, A., Usama, M., & Qadir, J., Speech emotion recognition using deep learning: A review. IEEE Transactions on Affective Computing, 2022

[9] Tian, Y., Zhang, X., & Cao, Y. Integrating speech and text for emotion recognition using transformers. IEEE Transactions on Neural Networks and Learning Systems, 33(6), 2611-2622, 2023

[10] Zhao, G., Schuller, B., & Zhang, X. Multimodal emotion recognition combining speech, facial expressions, and physiological signals. IEEE Transactions on Multimedia, 24, 2257-2267, 2023

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Brain Tumors Classification System Using Convolutional Recurrent Neural Network V. Akila, P.K. Abhilash, P BalaVenakata Satya Phanindra, J Pavan Kumar, A. Kavitha E3S Web Conf. 309 01075 (2021) DOI: 10.1051/e3sconf/202130901075.

[16] 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).

[17] 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.

[18] Lijiang Chen, Xia Mao, YuliXue, Lee Lung Cheng "Speech emotion recognition: Features and classification models", Digital Signal Processing 22 (2012) 1154–1160.

[19] Pavol Harar, RadimBurget 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.

[20] Dias Issa, M. FatihDemirci, Adnan Yazici "Speech emotion recognition with deep convolutional neural networks" Elsevier Ltd, Biomedical Signal Processing and Control 59 (2020) 101894.

[21] 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.

[22] 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.