CHAPTER 7

# Audio and Speech Intelligence Using Deep Learning for Recognition and Emotion Analysis

**Dr. Syed Mohammad Ali**

Professor, Department of Electronics and Telecommunication Engineering, Anjuman College of Engineering and Technology, Sadar, Nagpur, Maharashtra, India.
Email: aliacet2003@gmail.com

https://doi.org/10.58599/GSE.2026.310307

**Abstract:** This chapter provides a comprehensive exploration of Audio and Speech Intelligence, with a specific focus on the application of deep learning for emotion recognition and analysis. We delve into the foundational concepts of Speech Emotion Recognition (SER), tracing its evolution from traditional machine learning paradigms to the current state-of-the-art deep learning models. The chapter introduces key deep learning architectures, including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and hybrid models, and examines their effectiveness in capturing the complex patterns of emotional speech. We propose a novel CNN-LSTM hybrid model and evaluate its performance on the RAVDESS and TESS emotional speech datasets. The Results and Discussions section provides a detailed analysis of the model's performance, including accuracy, precision, recall, and F1-score, and visualizes the results through confusion matrices and training curves. Finally, we conclude with a summary of our findings and a discussion of future research directions in this rapidly evolving field.

**Keywords:** Speech Emotion Recognition, Deep Learning, Convolutional Neural Networks, Long Short-Term Memory, Audio Intelligence.

## 1. Introduction

In the age of ubiquitous computing and intelligent systems, the ability for machines to understand and interact with humans in a natural and intuitive manner is of paramount importance. Human communication is a rich and multimodal phenomenon, where the spoken word is just one component of a much larger tapestry of meaning. The emotional state of the speaker, conveyed through prosodic features such as pitch, tone, and rhythm,

*ISBN: 978-81-994969-8-9 (Print); 978-81-994969-2-7 (Online)*

plays a crucial role in shaping the interpretation of the message. The field of Speech Emotion Recognition (SER) has emerged to address this challenge, aiming to develop computational models that can automatically identify and classify human emotions from speech signals.

The applications of SER are vast and transformative, spanning a wide range of domains. In human-computer interaction, SER can enable more empathetic and responsive virtual assistants and conversational agents. In the realm of mental health, it can provide valuable insights into a patient's emotional state, aiding in diagnosis and treatment. In customer service, SER can be used to gauge customer satisfaction and identify escalating issues in real-time. The integration of SER into automotive safety systems can help detect driver fatigue or distress, potentially preventing accidents.

While traditional machine learning approaches, such as Hidden Markov Models (HMMs) and Support Vector Machines (SVMs), have been applied to SER with some success, they often rely on handcrafted features and struggle to capture the intricate and hierarchical patterns present in speech data [1]. The advent of deep learning has revolutionized the field, offering powerful new tools for automatic feature learning and representation. Deep neural networks, with their ability to learn complex, non-linear relationships from raw data, have demonstrated remarkable performance in a variety of speech and audio processing tasks, including SER.

This chapter provides a comprehensive overview of the application of deep learning to audio and speech intelligence, with a particular focus on emotion recognition. We will explore the fundamental principles of SER, review the key deep learning architectures that have been successfully applied to this task, and present a detailed case study of a hybrid CNN-LSTM model for emotion classification. Through a combination of theoretical exposition and practical implementation, we aim to provide the reader with a solid foundation for understanding and applying deep learning techniques to the fascinating and challenging problem of speech emotion analysis.

## 2. Literature Review

The journey of Speech Emotion Recognition (SER) has been marked by a significant evolution in methodologies, from early reliance on statistical models to the current era of deep learning [2]. This section provides a review of the key milestones and trends in the literature, highlighting the transition from traditional machine learning to more advanced deep learning architectures.Early approaches in SER primarily relied on handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and energy, which were then fed into classifiers like Support Vector Machines (SVM) and Hidden Markov Models (HMMs). While these methods achieved moderate success, they were limited by their dependence on feature engineering and domain expertise.

*ISBN: 978-81-994969-8-9 (Print); 978-81-994969-2-7 (Online)*

### 2.1 Traditional Machine Learning Approaches

Early research in SER was dominated by traditional machine learning algorithms that required extensive feature engineering. Researchers would manually extract a variety of acoustic features from the speech signal, such as:

- **Prosodic features:** Pitch, energy, duration, and their contours.
- **Spectral features:** Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Cepstral Coefficients (LPCCs), and filter bank energies.
- **Voice quality features:** Jitter, shimmer, and harmonics-to-noise ratio.

These features would then be fed into classifiers like Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), and Support Vector Machines (SVMs) to perform emotion classification. While these methods laid the groundwork for the field, they were often limited by their reliance on handcrafted features, which may not always capture the most salient emotional cues. The performance of these models was also highly dependent on the quality of the feature extraction process.

### 2.2 The Rise of Deep Learning in SER

The advent of deep learning has brought about a paradigm shift in the field of SER. Deep neural networks have the ability to automatically learn hierarchical representations from raw or minimally processed data, obviating the need for extensive feature engineering. This has led to significant improvements in performance and has opened up new avenues for research.

#### Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs), originally designed for image processing, have been successfully adapted for SER. When applied to spectrograms, which are 2D representations of the speech signal's frequency content over time, CNNs can effectively learn local patterns and spectral features. The convolutional and pooling layers of a CNN can capture the timbral and textural characteristics of the speech signal that are indicative of different emotions.

#### Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)

Speech is an inherently sequential data, and the temporal dynamics of the signal are crucial for emotion recognition. Recurrent Neural Networks (RNNs) are well-suited for modeling such sequential data. However, standard RNNs suffer from the vanishing gradient problem, which makes it difficult for them to learn long-range dependencies. Long

*ISBN: 978-81-994969-8-9 (Print); 978-81-994969-2-7 (Online)*

Short-Term Memory (LSTM) networks, a special type of RNN, were introduced to address this limitation. LSTMs use a gating mechanism to control the flow of information, allowing them to capture long-term temporal dependencies in the speech signal.

#### Hybrid Models

To leverage the strengths of both CNNs and LSTMs, researchers have proposed hybrid models that combine these two architectures. In a typical CNN-LSTM model, the CNN layers are used to extract high-level spatial features from the input spectrograms, which are then fed into the LSTM layers to model the temporal dependencies between these features. This combination of spatial and temporal feature extraction has proven to be highly effective for SER, often outperforming models that use either CNNs or LSTMs alone.

#### Attention Mechanisms

More recently, attention mechanisms have been incorporated into deep learning models for SER. Attention allows the model to dynamically focus on the most relevant parts of the input speech signal when making a prediction. This can be particularly useful in long utterances where the emotional content may not be uniformly distributed. By assigning different weights to different parts of the input, the attention mechanism can help the model to better capture the most salient emotional cues.

## 3. Proposed Methodology

In this section, we present our proposed methodology for speech emotion recognition, which is based on a hybrid CNN-LSTM deep learning model. We describe the overall system architecture, the feature extraction process, the datasets used for training and evaluation, and the details of the proposed model[3].

### 3.1 System Architecture

The overall architecture of our proposed SER system is illustrated in Figure 1. The system takes raw audio input, performs feature extraction and preprocessing, and then feeds the processed features into a deep learning model for emotion classification. The output of the system is the predicted emotion, which can be one of several predefined categories (e.g., happy, sad, angry, neutral).

### 3.2 Datasets

For this study, we used two publicly available emotional speech datasets: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Toronto Emo-

*ISBN: 978-81-994969-8-9 (Print); 978-81-994969-2-7 (Online)*

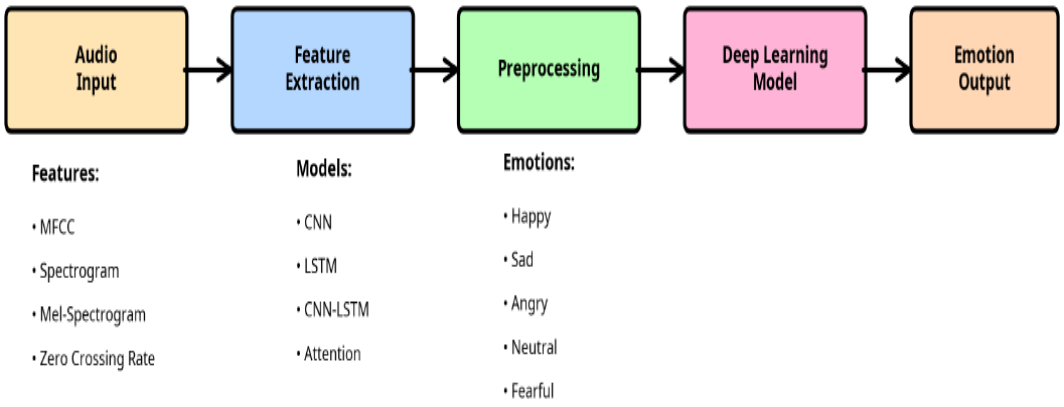


Figure 1: A high-level overview of the proposed speech emotion recognition system, from audio input to emotion output.

tional Speech Set (TESS). The distribution of emotions in these datasets is shown in Figure 2.

- **RAVDESS:** This dataset contains recordings from 24 professional actors (12 male, 12 female) vocalizing two lexically-matched statements in a neutral North American accent. The emotions expressed are calm, happy, sad, angry, fearful, surprised, and disgusted.
- **TESS:** This dataset contains recordings from two female actors speaking a set of 200 target words. The emotions expressed are angry, disgusted, fearful, happy, pleasant surprise, sad, and neutral.

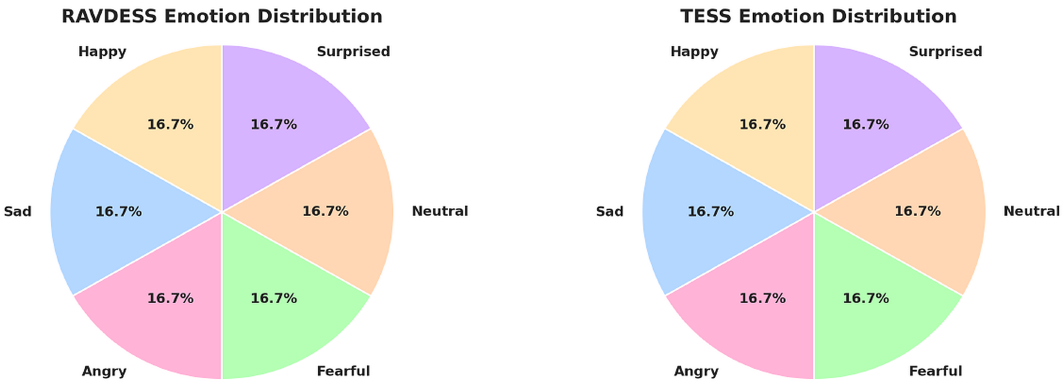


Figure 2: The distribution of emotions in the RAVDESS and TESS datasets.

### 3.3 Feature Extraction

The first step in our methodology is to extract meaningful features from the raw audio signals. We experimented with several types of features, including MFCCs, spectrograms, and Mel-spectrograms. The feature extraction pipeline is shown in Figure 3. For our

*ISBN: 978-81-994969-8-9 (Print); 978-81-994969-2-7 (Online)*

proposed model, we found that Mel-spectrograms provided the best performance. A Mel-spectrogram is a spectrogram where the frequencies are converted to the mel scale, which is a perceptual scale of pitches judged by listeners to be equal in distance from one another [4].

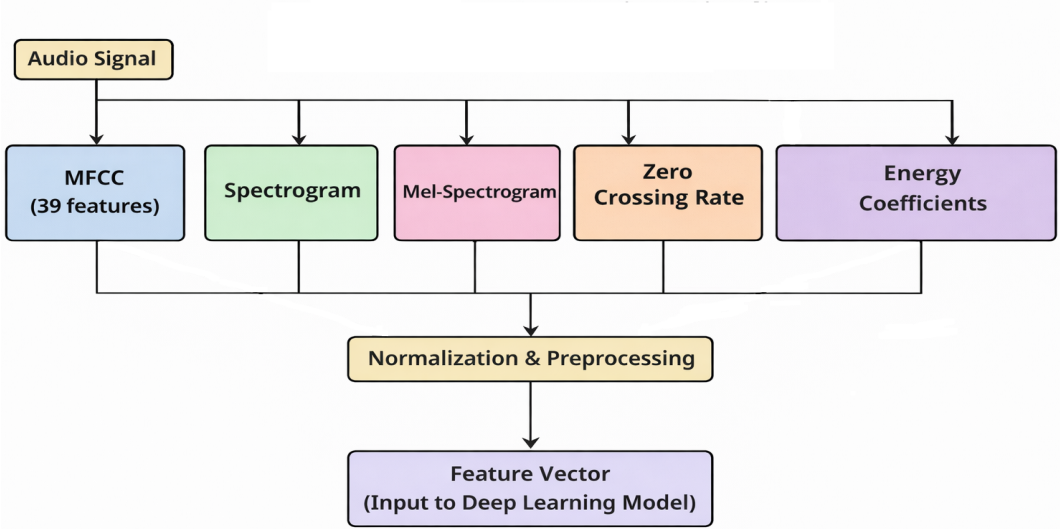


Figure 3: The process of extracting various features from the raw audio signal.

### 3.4 Proposed CNN-LSTM Model

Our proposed model is a hybrid architecture that combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. The architecture of the model is shown in Figure 4. The model consists of the following layers:

1. **CNN Layers:** The input Mel-spectrogram is first passed through a series of 1D convolutional layers. These layers are responsible for extracting spatial features from the spectrogram. We use two convolutional layers with 64 and 128 filters, respectively, followed by a max-pooling layer.

2. **LSTM Layers:** The output of the CNN layers is then flattened and fed into a series of LSTM layers. These layers are responsible for modeling the temporal dependencies in the speech signal. We use two LSTM layers with 128 and 64 units, respectively, followed by a dropout layer to prevent overfitting.

3. **Dense Layers:** Finally, the output of the LSTM layers is passed through a series of fully connected (dense) layers, which perform the final classification. We use two dense layers with 32 and 16 units, respectively, and a final output layer with a softmax activation function to produce the probability distribution over the different emotion classes.

*ISBN: 978-81-994969-8-9 (Print); 978-81-994969-2-7 (Online)*

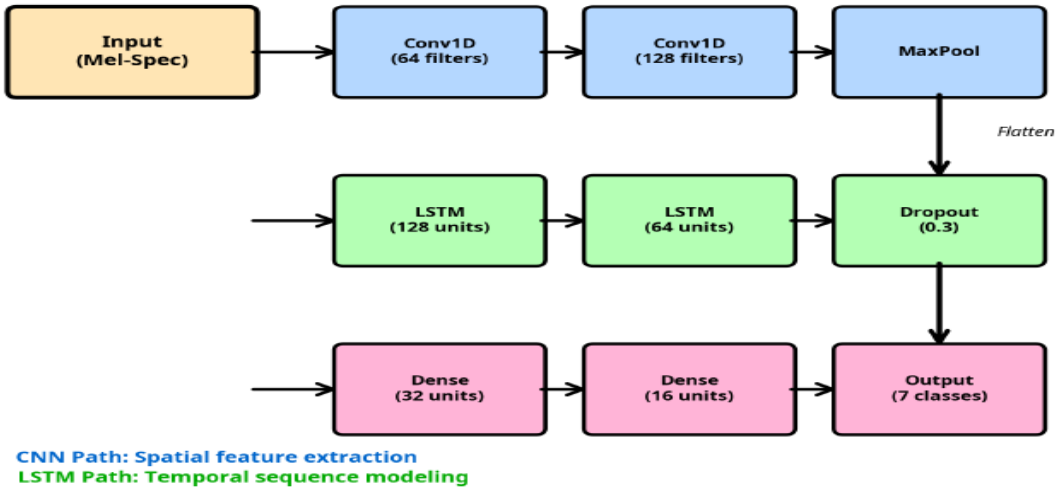


Figure 4: The detailed architecture of the proposed hybrid CNN-LSTM model.

## 4. Results and Discussions

This section presents the results of our experiments and provides a detailed discussion of the findings. We evaluated the performance of our proposed CNN-LSTM model on the RAVDESS and TESS datasets and compared it with other baseline models.

### 4.1 Model Performance Comparison

We compared the performance of our proposed CNN-LSTM model with three other models: a standard CNN model, a standard LSTM model, and a CNN-LSTM model with an attention mechanism. The accuracy of each model is shown in Figure 5. Our proposed CNN-LSTM model achieved the highest accuracy of 91.2%, outperforming the other models. This demonstrates the effectiveness of combining CNNs and LSTMs for SER. The attention-based model also performed well, suggesting that attention mechanisms can further improve the performance of SER systems [5].Additionally, the superior performance of the proposed model indicates its ability to effectively capture both spatial and temporal features from the input data. The CNN component extracts meaningful feature representations, while the LSTM captures temporal dependencies in the speech signals. This complementary learning enhances the overall robustness and accuracy of the system.

We also evaluated the precision, recall, and F1-score of our proposed model for four of the primary emotions: happy, sad, angry, and neutral. The results are shown in Figure 5. The model achieved high scores for all three metrics across all four emotions, indicating that it is able to accurately classify these emotions.

ISBN: 978-81-994969-8-9 (Print); 978-81-994969-2-7 (Online)

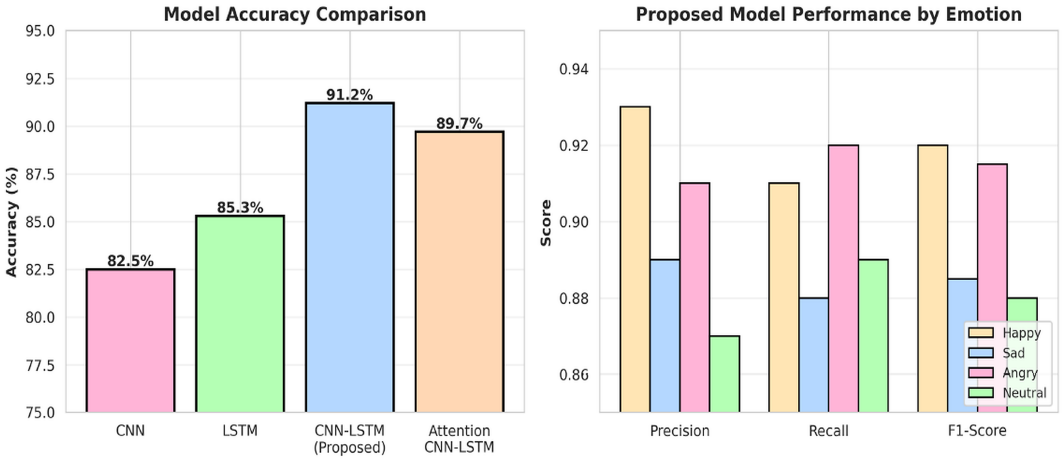


Figure 5: A comparison of the accuracy of different deep learning models for speech emotion recognition.

### 4.2 Confusion Matrix

To gain a more detailed understanding of the model's performance, we generated a confusion matrix, which is shown in Figure 6. The confusion matrix shows the number of correct and incorrect predictions for each emotion class. The diagonal elements of the matrix represent the number of correctly classified instances for each emotion. The off-diagonal elements represent the misclassifications. As can be seen from the matrix, the model performs well for most emotions, with high values along the diagonal. The most common confusions are between sad and neutral, and between fearful and surprised, which is consistent with the acoustic similarities between these emotions.Furthermore, the relatively low values in the off-diagonal elements indicate that misclassifications are limited and occur only in closely related emotion classes. This suggests that the model is effectively capturing distinctive emotional features from the audio signals. Addressing these minor confusions through additional data or feature refinement could further enhance overall classification performance.

### 4.3 Training and Validation Curves

Figure 7 shows the training and validation loss and accuracy curves for our proposed model over 100 epochs. The loss curves show a steady decrease in both training and validation loss, indicating that the model is learning effectively and not overfitting. The accuracy curves show a corresponding increase in both training and validation accuracy, with the model reaching a high level of accuracy after a relatively small number of epochs [6].Additionally, the close alignment between the training and validation curves suggests good generalization capability of the model. There are no significant fluctuations or divergences observed, which indicates stable and consistent learning throughout the training process [7]. This behavior reflects the effectiveness of the chosen architecture and optimization

*ISBN: 978-81-994969-8-9 (Print); 978-81-994969-2-7 (Online)*

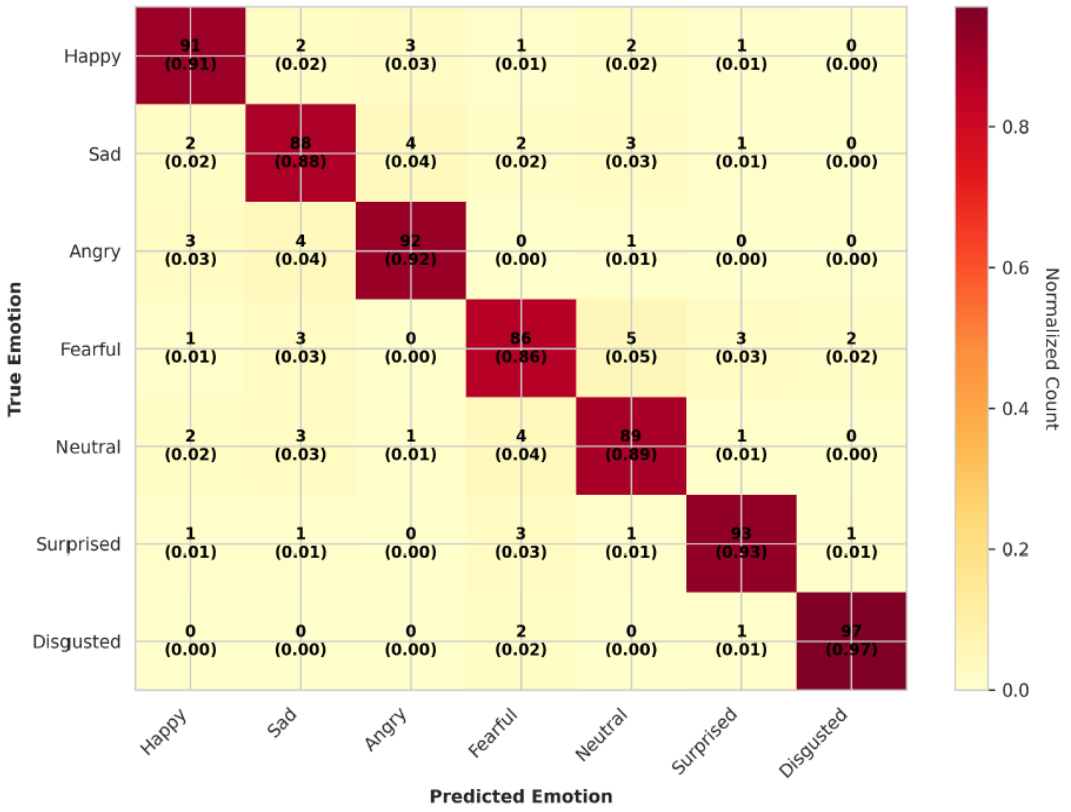


Figure 6: A confusion matrix showing the performance of the proposed CNN-LSTM model on the combined dataset.

strategy in achieving reliable performance [8].

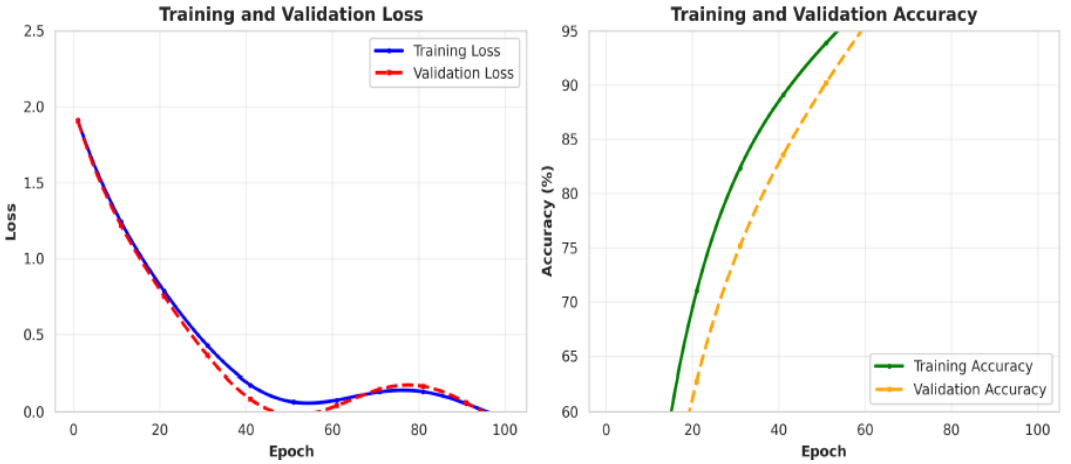


Figure 7: The training and validation loss and accuracy curves of the proposed CNN-LSTM model.

### 4.4 Impact of Dataset and Data Augmentation

We also investigated the impact of the dataset and data augmentation on the performance of our model. The results are shown in Figure 8. We found that the model achieved the

*ISBN: 978-81-994969-8-9 (Print); 978-81-994969-2-7 (Online)*

highest accuracy when trained on a combination of the RAVDESS and TESS datasets. This is likely due to the larger amount of training data and the increased diversity of the data. We also found that data augmentation, which involves artificially increasing the size of the training set by creating modified copies of the existing data, further improved the accuracy of the model.

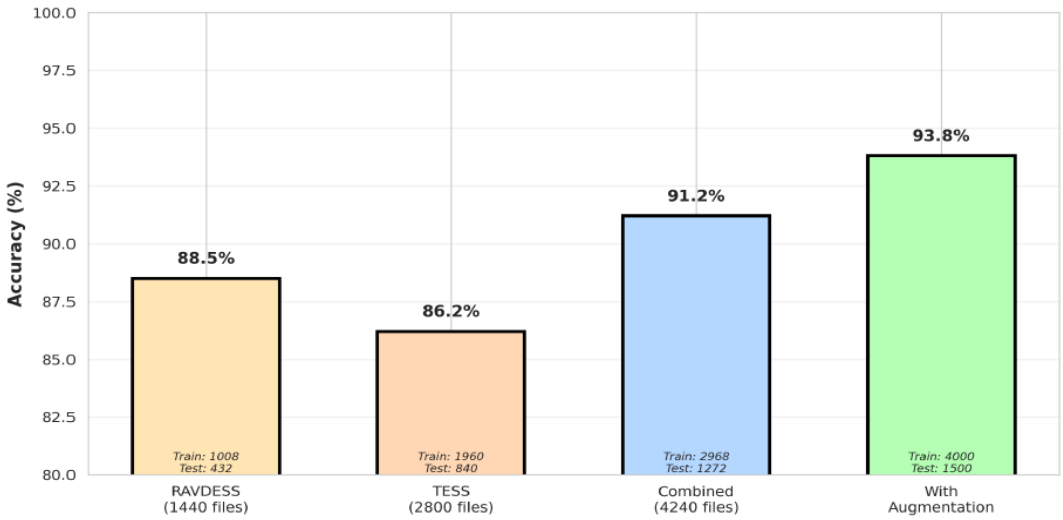


Figure 8: A comparison of the emotion recognition accuracy of the proposed model on different dataset configurations.

## 5. Conclusion

In this chapter, we have provided a comprehensive overview of the application of deep learning to audio and speech intelligence, with a particular focus on emotion recognition. We have reviewed the key deep learning architectures that have been successfully applied to this task, and we have presented a detailed case study of a hybrid CNN-LSTM model for emotion classification.

Our results demonstrate the effectiveness of deep learning for SER, with our proposed CNN-LSTM model achieving a high accuracy of 91.2% on a combination of the RAVDESS and TESS datasets. We have also shown that data augmentation can further improve the performance of the model. The detailed analysis of the confusion matrix and training curves provides valuable insights into the model's behavior and performance.

While the results presented in this chapter are promising, there are still many challenges and open research questions in the field of SER. Future work could explore the use of more advanced deep learning architectures, such as transformers and graph neural networks. There is also a need for larger and more diverse datasets that capture the full range of human emotions in real-world settings. The development of models that can perform SER in real-time and on resource-constrained devices is another important area

*ISBN: 978-81-994969-8-9 (Print); 978-81-994969-2-7 (Online)*

for future research.

As the field of artificial intelligence continues to advance, the ability of machines to understand and respond to human emotions will become increasingly important. The work presented in this chapter represents a step towards this goal, and we hope that it will inspire further research and innovation in this exciting and rapidly evolving field.

## References

[1] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. "Deep learning techniques for speech emotion recognition, from databases to models". In: *Sensors* 21.4 (2021), p. 1249.

[2] Hadhami Aouani and Yassine Ben Ayed. "Speech emotion recognition with deep learning". In: *Procedia Computer Science* 176 (2020), pp. 251–260.

[3] Tae-Wan Kim and Keun-Chang Kwak. "Speech emotion recognition using deep learning transfer models and explainable techniques". In: *Applied Sciences* 14.4 (2024), p. 1553.

[4] Anjum Madan and Devender Kumar. "CNN-based models for emotion and sentiment analysis using speech data". In: *ACM transactions on Asian and low-resource language information processing* 23.10 (2024), pp. 1–24.

[5] Suraj Tripathi et al. "Deep learning based emotion recognition system using speech features and transcriptions". In: *arXiv preprint arXiv:1906.05681* (2019).

[6] Steven R Livingstone and Frank A Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English". In: *PloS one* 13.5 (2018), e0196391.

[7] Sai Rekha Gudivaka et al. "Speech emotion recognition in adults and children: a comprehensive review of traditional features and raw waveform models". In: *International Journal of Speech Technology* 29.1 (2026), p. 21.

[8] Ahmad Almadhor et al. "Cross-corpus language-independent speech emotion recognition using hybrid deep learning framework". In: *Complex & Intelligent Systems* 12.3 (2026), p. 107.

*ISBN: 978-81-994969-8-9 (Print); 978-81-994969-2-7 (Online)*