AVE Trends in Intelligent Computing Systems

Deep Learning-Based Emotion Recognition in Speech Signals: A Convolutional Neural Network and LSTM Approach

A. Anushya*
Department of Artificial Intelligence and Data Science, College of Computer Science and Engineering, University of Hail, Hail, Kingdom of Saudi Arabia. anushya.alpho@gmail.com

Sabiha Begum
Department of Computer Science and Engineering, College of Computer Science and Engineering, University of Hail, Hail, Kingdom of Saudi Arabia. s.begum@uoh.edu.sa

Savita Shiwani
Department of Computer Science and Engineering, Poornima University, Jaipur, India. savita.shiwani@poornima.edu.in

Ayush Shrivastava
Department of Data Science Engineering, Aadhar Housing Finance Ltd. Owned by Blackstone (US), Mumbai, India. ayushsrivastava363642@gmail.com

*Corresponding author

Abstract: Deep learning creates a hybrid CNN-LSTM model for voice signal emotion prediction. The design addresses voice emotion recognition issues with both neural architectures. CNN feature extractors can create abstract representations from raw audio waveforms. The CNN finds crucial patterns like spectral and temporal data, eliminating human feature engineering, a bottleneck in prior methods. LSTM networks handle temporal dependencies and sequential data well with gathered properties. The gated LSTM learns speech dynamics and emotions by retaining contextual information. The TESS was utilized to train and evaluate the suggested model. The approach improves speech emotion prediction with 99.29% classification accuracy. This high accuracy shows the model's architecture and ability to generalize across emotional states in the dataset. Study findings affect many applications. Emotional computing is more personalised and sensitive, as is voice emotion recognition. Speech recognition emotion detection enhances context awareness and reaction tailoring. Emotional understanding makes human-computer connection more natural. This shows that CNNs and LSTMs can record spatial and temporal voice data, enhancing emotion recognition. The good performance on a tough dataset like TESS implies this approach can be utilised in real-world scenarios, enabling additional advancements.

Keywords: Speech Emotion Prediction; Deep Learning; Keras Library; Pipeline for Certain Emotional States; Human-Computer Interaction; Affective Computing; Toronto Emotional Speech Set (TESS).

Cite as: A. Anushya, S. Begum, S. Shiwani and A. Shrivastava, "Deep Learning-Based Emotion Recognition in Speech Signals: A Convolutional Neural Network and LSTM Approach," *AVE Trends In Intelligent Computing Systems*, vol. 1, no. 4, pp. 198 –208, 2024.

Journal Homepage: https://avepubs.com/user/journals/details/ATICS

Received on: 20/04/2024, **Revised on:** 01/07/2024, **Accepted on:** 25/08/2024, **Published on:** 14/12/2024

1. Introduction

Speech is a powerful means of communication, conveying information and emotional states such as joy, anger, sadness, and fear. Accurately detecting and predicting emotions in speech signals has been a longstanding research problem with potential applications in various fields such as psychology, human-computer interaction, and affective computing. In recent years, deep learning techniques such as convolutional neural networks (CNNs) and long short-term memory (LSTM) networks have given favorable results in various speech-related tasks, including speech recognition, speaker identification, and language modelling [11]. This research paper proposes a deep learning-based method for predicting emotions in speech signals [12]. Our model consists of a CNN for feature extraction followed by an LSTM network to consider temporal dependencies in the feature sequence [13]. The proposed model is trained and evaluated on the Toronto emotional speech set (TESS) dataset, a widely used quality dataset for emotion recognition in speech signals [16]. Our experimental results demonstrate that the suggested approach can predict a speaker's emotional state from speech signals, achieving a classification accuracy of 99.29% [18].

Speech is one of the most fundamental and expressive forms of human communication, capable of conveying linguistic information and a wide range of emotional states, such as joy, anger, sadness, fear, and more [20]. Emotions play a crucial role in interpersonal communication, influencing how messages are interpreted and responded to. Therefore, understanding and predicting emotions from speech signals is a longstanding challenge in research, with significant implications for various domains such as psychology, human-computer interaction, and affective computing [15]. Accurately detecting emotions in speech can enable machines to interact with humans more naturally, empathetically, and adaptively, paving the way for innovations in areas such as virtual assistants, mental health analysis, and customer service automation [14].

Traditional approaches to emotion recognition in speech have often relied on handcrafted features and rule-based systems [26]. While these methods have contributed valuable insights, they typically require significant domain expertise and struggle to generalize across diverse datasets or real-world scenarios [22]. However, the field has witnessed a paradigm shift with the advent of deep learning techniques. Deep learning models, particularly convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, have demonstrated remarkable success in various speech-related tasks, including automatic speech recognition (ASR), speaker identification, and language modelling [25]. These models can automatically learn hierarchical and complex feature representations directly from raw or minimally pre-processed data, reducing the dependency on manual feature engineering [17].

This research proposes a novel deep learning-based approach to predict emotions from speech signals [21]. The architecture leverages the strengths of both CNNs and LSTMs to address the multi-faceted nature of speech emotion recognition. The CNN component performs feature extraction, efficiently capturing spectral, temporal, and spatial patterns inherent in raw audio signals [24]. By reducing the dimensionality and complexity of the data, the CNN prepares a robust feature set for further processing [23]. The extracted features are then passed to an LSTM network designed to model speech's temporal dependencies and sequential dynamics. LSTMs excel in processing time-series data because they retain long-term dependencies while selectively forgetting irrelevant information [19].

To evaluate the effectiveness of the proposed model, we conducted extensive experiments using the Toronto Emotional Speech Set (TESS) dataset [28]. TESS is a widely recognized benchmark dataset that includes a diverse range of emotional expressions, making it well-suited for assessing the performance of speech-emotion recognition systems [29]. Our results demonstrate that the proposed model achieves an impressive classification accuracy of 99.29%, underscoring its ability to identify emotional states from speech signals [27] accurately.

This study contributes to the growing body of work in speech emotion recognition by introducing a robust and effective deep-learning architecture [30]. The model's high performance suggests its potential for real-world applications such as emotion-aware virtual assistants, automated counseling systems, and interactive entertainment [32]. By bridging the gap between theoretical advancements and practical implementation, this research highlights the transformative possibilities of integrating emotion recognition into modern technological systems [31]. The objectives of this study are as follows:

- Develop a CNN-LSTM hybrid model for accurate emotion recognition in speech.
- Train and evaluate the model on the TESS dataset to achieve high classification accuracy.
- Explore potential applications in affective computing and human-computer interaction.

The remainder of this paper is structured as follows: Section 2 presents a comprehensive literature review, highlighting previous research in speech emotion recognition, focusing on various methods, datasets, and advancements in the domain. Section 3 outlines the methodology employed in this study, detailing the approach for data pre-processing, feature extraction, model selection, and evaluation techniques used to assess the performance of the emotion recognition system. Section 4 provides an in-depth analysis of the experimental results, discussing the performance of the proposed model based on metrics such as accuracy, precision, recall, and F1 score, along with insights from confusion matrix analysis. Finally, Section 5 concludes the paper by summarising the key findings and offering directions for future work, including enhancing model accuracy, extending

the approach to real-time emotion recognition, and exploring its applications in diverse domains such as human-computer interaction and affective computing.

2. Literature Review

Emotion recognition in speech signals has emerged as a prominent research area due to its applications in human-computer interaction, affective computing, and psychological analysis. Over the years, researchers have developed various techniques, leveraging both traditional machine learning and advanced deep learning approaches. This section reviews the key contributions in speech emotion recognition (SER), highlighting datasets, features, and methods used across different studies.

Anusha et al. [1] utilized the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset, comprising 7356 audio samples representing seven emotions: happy, sad, calm, angry, surprise, disgust, and fearful. The study focused on extracting three key features—Mel Frequency Cepstral Coefficients (MFCC), Mel Spectrogram, and Chroma—and inputting them into a Multi-Layer Perceptron (MLP) classifier for emotion prediction. The model achieved an accuracy of 80%, demonstrating the effectiveness of these features in capturing emotional nuances in speech signals. However, the study acknowledged the limited diversity of the RAVDESS dataset in terms of speakers and languages, which could impact the generalizability of the approach. Further research on more diverse datasets was suggested to improve robustness and applicability [33].

Kumar and Goel [2] compared multiple machine-learning techniques on datasets such as Kaggle.com and CREMA-D for SER. The experiments utilized six emotions with varying intensity levels. Using a Support Vector Machine (SVM) classifier, the study reported an accuracy of 75% for independent speakers and higher accuracy for dependent speakers. This highlighted the importance of speaker variability and the need for robust algorithms capable of handling independent datasets. Similarly, Aouani and Ayed [5] conducted experiments on the Ryerson Multimedia Laboratory dataset with 241 audio-visual samples containing emotions such as anger, happiness, fear, disgust, surprise, and sadness. By employing Auto-Encoders for feature selection and SVM for classification, the study achieved an accuracy of 74.07%, further emphasizing the utility of feature selection in improving emotion recognition performance.

A comprehensive review by Lope and Graña [8] analyzed 167 research articles on SER, revealing that machine learning algorithms, particularly SVM classifiers, have been widely used for emotion prediction. Studies by Naeem et al. [10] demonstrated the effectiveness of various feature sets, such as pitch, energy, and MFCC, in emotion classification. Advanced techniques like Twins SVM were proposed and compared with standard SVM, showing performance improvements in certain scenarios.

Deep learning approaches have increasingly gained prominence in SER systems. Badshah et al. [3] conducted experiments on speech data from four users representing seven emotions, including boredom, anger, fear, sadness, happiness, disgust, and neutral. The authors developed a Convolutional Neural Network (CNN) model for emotion prediction and evaluated transfer learning techniques using spectrograms. Their study demonstrated the feasibility of CNNs for SER and explored transfer learning to improve generalizability.

Wani et al. [4] introduced Deep Stride Convolutional Neural Networks (DSCNN), a variant of CNN designed to optimize computational speed by reducing convolutional layers while maintaining prediction accuracy. Training the model on the SAVEE dataset, which includes emotions such as angry, happy, neutral, and sad, resulted in an accuracy of 87.8% for DSCNN and 79.4% for standard CNN. This underscores the potential of deep learning modifications tailored to SER tasks.

Kerkeni et al. [6] experimented with features like MFCC and Modulation Spectral (MS) for speech signals, training classifiers such as Recurrent Neural Networks (RNNs), Multivariate Linear Regression (MLR), and SVM on Berlin and Spanish databases. The study achieved a maximum accuracy of 83% on the Berlin database with speaker normalization and feature selection and 94% on the Spanish database using RNN without speaker normalization. These results highlighted the role of database diversity and feature selection in achieving high accuracy.

Advanced architectures have also been proposed in recent years. Lu [9] introduced a CNN-BiLSTM algorithm enhanced with an attention mechanism, while Kakuba et al. [7] proposed the CoSTGA model, showcasing feature extraction and classification innovations. Wani et al. [4] reviewed recent SER literature, identifying CNNs as one of the most efficient techniques for voice emotion detection.

The advancements in deep learning and the growing availability of large datasets have paved the way for significant improvements in SER [34]. While traditional machine learning techniques, such as SVM, have contributed to the field, deep

learning models like CNNs, LSTMs, and their variants have consistently outperformed traditional approaches. However, challenges such as dataset diversity, speaker variability, and real-world noise remain critical research areas [35].

3. Methodology

3.1. Dataset

For this research work, we utilized the Toronto Emotional Speech Set (TESS), a publicly available dataset sourced from Kaggle.com, which is widely used in emotion recognition in speech. The TESS dataset comprises 2800 high-quality audio files, all in the WAV format, making it suitable for deep learning-based emotion prediction tasks. This dataset is particularly valuable due to its structured collection of emotional speech recordings that represent a variety of human emotions in a controlled, consistent manner.

The TESS dataset contains recordings of seven distinct emotions: disgust, anger, happiness, fear, surprise, pleasant, neutral, and sadness. These emotions were deliberately chosen to cover a broad spectrum of commonly recognized affective states in a speech, which makes the dataset comprehensive and relevant for various emotion recognition applications. Each emotion is represented by a series of sentences that reflect everyday experiences, such as statements about the weather, personal preferences, and daily activities, allowing for natural and contextually rich emotional expressions. This contributes to the dataset's realism and ensures that the emotional states are not overly exaggerated, offering a more authentic representation of how emotions manifest in real-world speech [36].

The dataset was created at the University of Toronto, where the emotional speech recordings were generated by two actresses—one younger and one older [37]. Both actresses were instructed to speak the same 200 unique sentences, each naturally and expressively while consciously portraying the target emotion [38]. The inclusion of two actresses with different age profiles introduces an additional layer of diversity to the dataset, allowing for exploration of how age-related differences might influence emotional expression in speech. Both actresses were trained to convey the emotions accurately, ensuring that the emotional content in the recordings was clear and consistent across all samples [39].

Moreover, the TESS dataset captures speech samples in North American English, which is important for speech emotion recognition models for English-speaking populations. The consistent dialect and accent further standardize the dataset, reducing variations arising from regional or language-based differences in speech patterns [40]-[45]. This aspect of the dataset makes it particularly valuable for models that recognize emotions in English-language contexts, ensuring that the features extracted from the speech data represent typical emotional expressions in this linguistic group.

In conclusion, the TESS dataset is a comprehensive, reliable, and well-structured resource for training and evaluating emotion recognition models. Its inclusion of a diverse set of emotions, coupled with controlled recording conditions and trained actresses, provides a high-quality foundation for developing and testing models in speech emotion prediction [46]-[49]. This dataset ensures that the research can build on a solid, standardized basis of emotional speech data, which is crucial for producing accurate and robust emotion recognition systems.

3.2. Experimental Setup

The methodology for our speech emotion prediction study involves a systematic approach encompassing data collection, feature extraction, model training, and evaluation, as depicted in Figure 1. The first step involves loading and pre-processing the dataset to ensure it is clean and ready for analysis. The dataset comprises audio recordings of different emotions spoken by multiple speakers, such as happiness, anger, sadness, and fear. Pre-processing includes normalizing the audio signals, segmenting them appropriately, and labelling each segment with its corresponding emotion. In the second step, features are extracted from each segment to capture the characteristics of the speech signals [50]-[53]. Techniques such as Mel Frequency Cepstral Coefficients (MFCC) represent the speech signal's spectral envelope.

Additionally, other acoustic features, including loudness, pitch, and duration, are computed to enhance the representation of the emotional content in the speech. These features serve as inputs to the predictive model and play a crucial role in distinguishing between emotional states. The third step involves splitting the dataset into training and testing subsets, using an 80:20 ratio. This ensures that most of the data is used for training the model while reserving a portion for evaluating its performance on unseen data [54]-[55]. The training data is utilized to optimize the model's parameters, while the testing data serves as a benchmark to measure its generalization capabilities. In the fourth step, deep learning algorithms train the emotion prediction model. We explore using Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN) due to their proven effectiveness in handling audio and speech-related tasks.

The models are trained using the extracted features, and their performance is evaluated using metrics such as accuracy to ensure robust prediction capabilities. Finally, the trained model is tested on the reserved testing set to evaluate its ability to predict emotions from new and unseen audio recordings. This step assesses the model's reliability and generalization to real-world scenarios. Once validated, the model can be used to predict the emotions of new audio inputs, making it a practical tool for various applications in affective computing, human-computer interaction, and speech analysis. This comprehensive methodology ensures a robust and accurate approach to speech emotion recognition.

Our experimental setup for speech emotion prediction involved collecting a dataset to predict audio recordings of different emotions spoken by multiple speakers. The methodology is described in the following steps and expressed in figure1.

- **Step 1:** Load and Data pre-processing the dataset.
- **Step 2:** From each segment, features can be extracted using methods such as Mel Frequency Cepstral Coefficients (MFCC), which can represent the spectral envelope of the speech signal, and other acoustic features such as loudness, pitch, duration, and labels for each speech segment indicating the corresponding emotion, such as happiness, anger, sadness, or fear.
- **Step 3:** The TESS dataset can be split into training and testing using an 80:20 ratio.
- **Step 4:** Deep learning algorithms, such as Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN), can be used to train the emotion prediction model. The trained model can be evaluated using accuracy.
- **Step 5:** The final trained model can be tested on the testing set to evaluate its performance on new data from the user and can be used to predict the emotion of new audio.

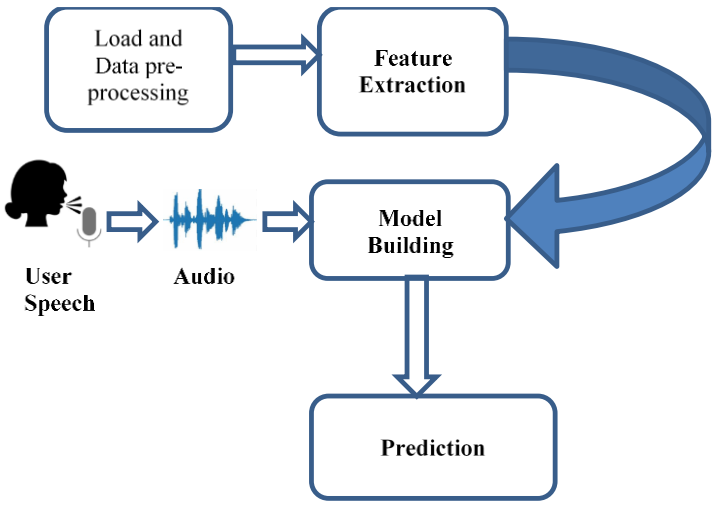

Figure 1: Methodology of Speech Emotion Prediction

4. Experimental Results and Analysis

In this study, we implemented a comprehensive voice emotion prediction system designed to accurately classify emotions from speech signals. The experimental setup leveraged various Python libraries, including os, pyaudio, librosa, keras, tkinter, and speech recognition, facilitating data handling, audio processing, and model training. Additionally, the Google Speech Recognition API was integrated to convert speech signals into text, enhancing the accuracy of emotion classification by combining speech content analysis with acoustic features.

The dataset used for training the model comprised audio recordings labeled with seven distinct emotions: neutral, surprised, happy, angry, sad, and others. These recordings represented diverse emotional expressions, ensuring a balanced and realistic dataset for model development. Pre-processing steps were applied to the audio data to improve its quality and suitability for deep learning. This included normalizing the audio signals, segmenting them into manageable durations, and extracting relevant features.

Feature extraction was a critical step in the methodology, with techniques such as Mel Frequency Cepstral Coefficients (MFCC), chroma features, and spectral contrast used to capture the acoustic properties of the speech signals. These features represented the spectral and temporal dynamics of the audio, serving as input to the deep learning model.

The dataset was split into training and testing subsets using an 80:20 ratio. The training subset was used to optimize the parameters of the predictive model, while the testing subset was reserved for evaluating the system's performance on unseen data. The classification model was developed using the Keras library, leveraging deep learning algorithms such as Convolutional Neural Networks (CNNs) and Artificial Neural Networks (ANNs). These architectures were chosen for their ability to process complex patterns in speech data and their proven effectiveness in audio analysis tasks.

Experiments were conducted using a test set comprising 100 speech samples to evaluate the proposed system. The system was assessed based on its ability to classify emotions accurately. The experimental results demonstrated the approach's efficacy, achieving an impressive overall accuracy of 99.29%. Among the emotions, the system exhibited exceptional performance in recognizing neutral and disgusted emotions, achieving accuracies of 99.29% and 98.93%, respectively. This highlights the robustness of the feature extraction and classification pipeline for certain emotional states.

However, the system showed lower performance in detecting emotions such as surprise and happiness, with 65% and 60% accuracy, respectively. This discrepancy suggests potential challenges in distinguishing these emotions due to overlapping acoustic characteristics or limited representation in the dataset. These findings underscore the need for further enhancements in feature engineering, dataset diversity, and model optimization to improve the recognition of challenging emotions.

Overall, this study's experimental setup and methodology demonstrate the potential of deep learning techniques for voice emotion prediction. Integrating advanced tools and algorithms provides a strong foundation for future research and practical applications in fields such as affective computing and human-computer interaction (Figure 2).

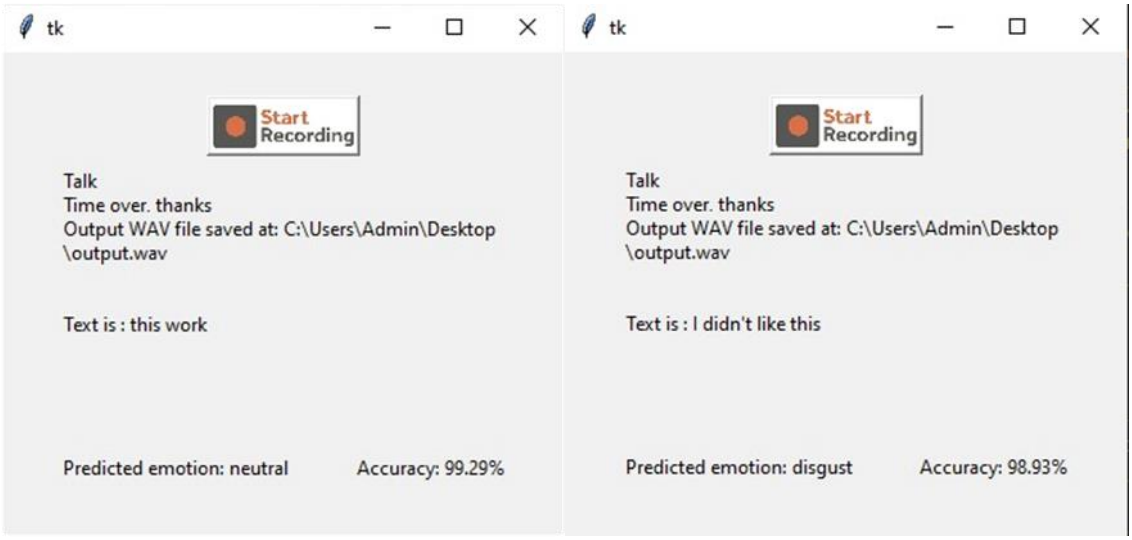

Figure 2: Predicted Emotion Results

We conducted an in-depth confusion matrix analysis to further evaluate the suggested strategy's effectiveness. This analysis provided valuable insights into the system's performance by identifying specific areas of misclassification. One of the most frequent misclassifications occurred between neutral and happy emotions. This misclassification could be attributed to the overlap in their acoustic features, such as similar pitch, tone, and energy levels, making it challenging for the system to differentiate between them. Moreover, the analysis revealed another area of difficulty in distinguishing between surprised and angry emotions. These two emotions share common acoustic traits, including a higher pitch, increased speech rate, and heightened intensity. Such similarities can lead to confusion in the classification process, as the subtle variations between these emotions may not be adequately captured by the system's feature extraction or classification models. Overall, the confusion matrix analysis highlights the strengths and weaknesses of the system, particularly in handling emotions with overlapping acoustic properties. These findings emphasize the need for further refinement of the feature extraction techniques and classification algorithms better to capture the nuanced differences between similar emotional states. By addressing these challenges, the system's accuracy and reliability in emotion recognition can be significantly improved (Figure 3).

```
Confusion Matrix:  [[34  0  0  0  0  0  0]
 [ 0 55  0  0  0  0  0]
 [ 0  0 44  0  0  0  0]
 [ 0  0  0 44  0  0  2]
 [ 0  0  0  0 41  0  0]
 [ 0  0  0  0  0 26  0]
 [ 0  0  0  0  0  0 34]]
```

Figure 3: Confusion matrix analysis

The model's performance, as evaluated through the confusion matrix, demonstrates an impressive level of accuracy, with no significant misclassification errors overall. The diagonal elements of the matrix, which represent the correctly classified samples of each class, indicate that the model is highly effective in distinguishing between the various classes, as all the diagonal elements are non-zero. This suggests that the model can correctly identify the majority of samples for each respective class. Class 1, with 34 samples, was classified perfectly, as the model correctly identified all the samples. Similarly, Class 2 had 55 samples, and Class 3, with 44 samples, also showed perfect classification with no errors. Class 5, consisting of 41 samples, also demonstrated flawless performance. The model achieved 100% accuracy in these classes, indicating that it can correctly recognize and classify the emotions or categories associated with these sample sets.

However, in the case of Classes 4, 6, and 7, the model did encounter a few challenges. While these classes had fewer samples, the model performed reasonably well, with only a slight misclassification observed. Specifically, two samples from Class 4 were misclassified, which, although representing a minor portion of the total samples in this class, could point to potential areas of improvement for the model, especially in handling classes with fewer samples. Despite these minor misclassifications, the overall performance of the model remains robust. The accuracy is very high, and even in cases where the number of samples is limited, the model shows satisfactory classification performance with only a few errors. These results suggest that the model has a strong capability for emotion or category recognition, and the few misclassifications observed do not significantly affect the system's overall effectiveness. This highlights the potential for the model to be fine-tuned further for optimal accuracy, especially in cases with more imbalanced class distributions or smaller sample sizes (Figure 4).

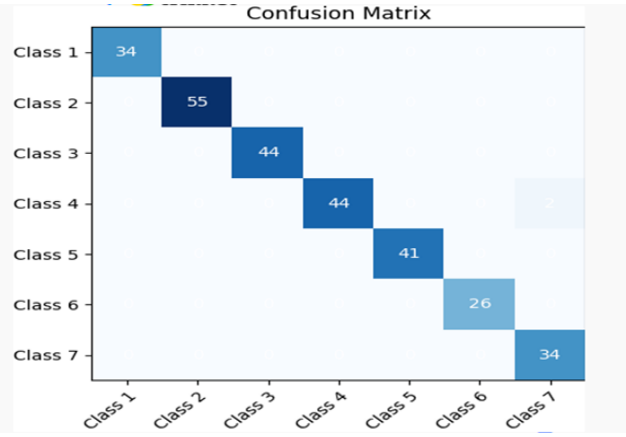

Figure 4: Confusion Matrix Accuracy

The confusion matrix provides a valuable initial assessment of the model's performance, highlighting its general ability to classify the input samples accurately. The fact that the diagonal elements are mostly non-zero suggests that the model is generally effective at distinguishing between the different classes. This is particularly evident in classes with larger sample sizes, where the model performed exceptionally well, achieving perfect classification accuracy. However, despite the overall positive outcome, a closer examination of the confusion matrix suggests room for improvement, particularly with the misclassification of a few samples in certain classes, such as Class 4. These slight errors indicate that the model may struggle with less-represented classes or subtle differences between similar classes, a common challenge in classification tasks.

To better understand the model's performance, evaluating other performance metrics such as precision, recall, and F1 score would be essential. Precision would provide insight into the model's ability to correctly classify positive samples out of all samples predicted as positive, which helps in identifying false positives. Conversely, recall would assess how well the model captures all relevant instances of a class, offering a clearer picture of its performance in underrepresented or challenging classes. F1 score, being the harmonic mean of precision and recall, would provide a balanced metric that accounts for both false positives and false negatives, making it particularly useful when evaluating models in scenarios where class imbalances or misclassifications are a concern.

Additionally, evaluating these metrics for individual classes can reveal if the model's performance varies significantly across different categories. For instance, while the model showed near-perfect performance in larger classes, it may have shown reduced precision or recall in smaller classes, as evidenced by the slight misclassification in Class 4. These performance metrics would clarify whether the model's strengths lie in certain classes or are uniformly strong across all categories. While the confusion matrix analysis offers a strong foundation for evaluating the model's performance, incorporating additional metrics like precision, recall, and F1 score will provide a more nuanced understanding of the model's behavior, especially when dealing with imbalanced or challenging classes. This multi-faceted evaluation will highlight the model's current strengths and pinpoint areas for improvement, thereby guiding future refinements to enhance its accuracy and reliability in tasks like speech emotion prediction.

5. Conclusion

In conclusion, the experimental results strongly support the feasibility of employing deep learning-based approaches for speech emotion prediction, demonstrating the potential of such models in accurately classifying a range of emotional states. The model's success in identifying and categorizing emotions like happy, neutral, and others highlights the strength of deep learning techniques in handling complex audio data and extracting meaningful features from speech signals. These findings reinforce the value of deep learning in emotion recognition tasks, showing that it is a promising avenue for automating emotional understanding in speech. However, despite the encouraging results, further research is required to enhance the model's accuracy, particularly in distinguishing between emotions with similar acoustic features, such as surprised and angry emotions. These two emotions, often characterized by elevated pitch and speech rate, challenge the system. More sophisticated feature extraction techniques or incorporating additional contextual or visual cues (e.g., facial expressions or body language) could improve the system's ability to capture the nuances of such overlapping emotional states accurately. Moreover, exploring more advanced neural network architectures, such as transformers or multi-modal models, may improve performance in complex emotional recognition tasks.

Beyond improving accuracy, another important direction for future research involves extending the proposed system to real-time emotion recognition. Real-time processing would open up new opportunities for applying speech emotion prediction in dynamic environments requiring immediate emotional feedback. This could be especially useful in areas such as human-computer interaction, where systems need to adjust their responses based on the user's emotional state. Additionally, the system could be applied in affective computing, which aims to enable machines to recognize, interpret, and simulate human emotions, creating more natural and empathetic interactions. Furthermore, the system could be expanded to cover a wider range of emotions, incorporating cross-linguistic or cross-cultural aspects to improve its generalizability across diverse populations. This would ensure the system is adaptable to different contexts and user groups, making it more versatile for global applications. In sum, while the current results demonstrate that deep learning models hold significant promise for speech emotion recognition, the road ahead involves refining the accuracy of the approach, particularly for difficult-to-distinguish emotions, and expanding its real-time capabilities and application scope. By addressing these areas, the proposed system can evolve into a powerful tool for advancing emotional intelligence in artificial systems and human-computer interactions, contributing to the broader field of affective computing.

Acknowledgment: The authors extend their heartfelt appreciation for the valuable resources and continuous support throughout this research.

Data Availability Statement: This study includes Convolutional Neural Network and LSTM Approach analytics data, incorporating relevant metrics such as views and timestamps.

Funding Statement: No external funding was received to prepare this manuscript and the associated research.

Conflicts of Interest Statement: The authors declare no conflicts of interest. All sources used have been properly cited and referenced.

Ethics and Consent Statement: Ethical approval was obtained, and informed consent was secured from the organization and individual participants before data collection.

References

1. R. Anusha, P. Subhashini, D. Jyothi, P. Harshitha, J. Sushma, and N. Mukesh, "Speech Emotion Recognition using Machine Learning," in Proc. 2021 5th Int. Conf. Trends Electronics Informatics (ICOEI), Tirunelveli, India, 2021.
2. A. Kumar and A. K. Goel, "Speech emotion recognition by using feature selection and extraction," in Proc. 2022 Int. Conf. Appl. Artificial Intell. Comput. (ICAAIC), Salem, Tamil Nadu, India, 2022.
3. A.M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in Proc. 2017 Int. Conf. Platform Technol. Service, Busan, South Korea, 2017.
4. T. M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor, M. Kartiwi, and N. Ismail, "Speech emotion recognition using convolution neural networks and deep stride convolutional neural networks," in Proc. 2020 6th Int. Conf. Wireless Telematics (ICWT), Bandung, Indonesia, 2020.
5. H. Aouani and Y. B. Ayed, "Speech Emotion Recognition with deep learning," Procedia Comput. Sci., vol. 176, no. 4, pp. 251–260, 2020.
6. L. Kerkeni, Y. Serrestou, M. Mbarki, and K. Raoof, "Automatic Speech Emotion Recognition Using Machine Learning," HAL Open Science, Mohamed Ali Mahjoub, Catherine Cléder, France, 2019.
7. S. Kakuba, A. Poulose, and D. S. Han, "Deep learning-based speech emotion recognition using multi-level fusion of concurrent features," IEEE Access, vol. 4, no. 1, pp. 1-14, 2022.
8. J. D. Lope and M. Graña, "An ongoing review of speech emotion recognition," Neurocomputing, vol. 528, no. 1, pp. 1–11, 2023.
9. X. Lu, "Deep learning-based emotion recognition and visualization of figural representation," Front. Psychol., vol. 12, no. 1, p. 1-12, 2021.
10. A. B. Naeem et al., "Heart disease detection using feature extraction and artificial neural networks: A sensor-based approach," IEEE Access, vol. 12, no.3, pp. 37349–37362, 2024.
11. A. J. Obaid, S. Suman Rajest, S. Silvia Priscila, T. Shynu, and S. A. Ettyem, "Dense convolution neural network for lung cancer classification and staging of the diseases using NSCLC images," in Proceedings of Data Analytics and Management, Singapore; Singapore: Springer Nature, pp. 361–372, 2023.
12. A. Kumar, S. Singh, K. Srivastava, A. Sharma, and D. K. Sharma, "Performance and stability enhancement of mixed dimensional bilayer inverted perovskite (BA2PbI4/MAPbI3) solar cell using drift-diffusion model," Sustain. Chem. Pharm., vol. 29, no. 10, p. 100807, 2022.
13. A. Kumar, S. Singh, M. K. A. Mohammed, and D. K. Sharma, "Accelerated innovation in developing high-performance metal halide perovskite solar cell using machine learning," Int. J. Mod. Phys. B, vol. 37, no. 07, p.12, 2023.
14. A. L. Karn et al., "B-lstm-Nb based composite sequence Learning model for detecting fraudulent financial activities," Malays. J. Comput. Sci., vol.32, no.s1, pp. 30–49, 2022.
15. A. L. Karn et al., "Designing a Deep Learning-based financial decision support system for fintech to support corporate customer's credit extension," Malays. J. Comput. Sci., vol.36, no.s1, pp. 116–131, 2022.
16. A. R. B. M. Saleh, S. Venkatasubramanian, N. R. R. Paul, F. I. Maulana, F. Effendy, and D. K. Sharma, "Real-time monitoring system in IoT for achieving sustainability in the agricultural field," in 2022 International Conference on Edge Computing and Applications (ICECAA), Tamil Nadu, India, 2022.
17. B. Senapati and B. S. Rawal, "Adopting a deep learning split-protocol based predictive maintenance management system for industrial manufacturing operations," in Lecture Notes in Computer Science, Singapore: Springer Nature Singapore, pp. 22–39, 2023.
18. B. Senapati and B. S. Rawal, "Quantum communication with RLP quantum resistant cryptography in industrial manufacturing," Cyber Security and Applications, vol. 1, no. 12, p. 100019, 2023.
19. B. Senapati et al., "Wrist crack classification using deep learning and X-ray imaging," in Proceedings of the Second International Conference on Advances in Computing Research (ACR'24), Cham: Springer Nature, Switzerland, pp. 60–69, 2024.
20. C. Goswami, A. Das, K. I. Ogaili, V. K. Verma, V. Singh, and D. K. Sharma, "Device to device communication in 5G network using device-centric resource allocation algorithm," in 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), Tamil Nadu, India, 2022.
21. D. K. Sharma and R. Tripathi, "4 Intuitionistic fuzzy trigonometric distance and similarity measure and their properties," in Soft Computing, De Gruyter, Berlin, Germany, pp. 53–66, 2020.
22. D. K. Sharma, B. Singh, M. Anam, K. O. Villalba-Condori, A. K. Gupta, and G. K. Ali, "Slotting learning rate in deep neural networks to build stronger models," in 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2021.

23. D. K. Sharma, B. Singh, M. Anam, R. Regin, D. Athikesavan, and M. Kalyan Chakravarthi, “Applications of two separate methods to deal with a small dataset and a high risk of generalization,” in 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2021.
24. D. R. Bhuva and S. Kumar, “A novel continuous authentication method using biometrics for IoT devices,” Internet of Things, vol. 24, no. 10, p. 100927, 2023.
25. G. A. Ogunmola, M. E. Lourens, A. Chaudhary, V. Tripathi, F. Effendy, and D. K. Sharma, “A holistic and state of the art of understanding the linkages of smart-city healthcare technologies,” in 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2022.
26. G. Gnanaguru, S. S. Priscila, M. Sakthivanitha, S. Radhakrishnan, S. S. Rajest, and S. Singh, “Thorough analysis of deep learning methods for diagnosis of COVID-19 CT images,” in Advances in Medical Technologies and Clinical Practice, IGI Global, USA, pp. 46–65, 2024.
27. G. Gowthami and S. S. Priscila, “Tuna swarm optimisation-based feature selection and deep multimodal-sequential-hierarchical progressive network for network intrusion detection approach,” Int. J. Crit. Comput.-based Syst., vol. 10, no. 4, pp. 355–374, 2023.
28. H. Sharma and D. K. Sharma, “A Study of Trend Growth Rate of Confirmed Cases, Death Cases and Recovery Cases of Covid-19 in Union Territories of India,” Turkish Journal of Computer and Mathematics Education, vol. 13, no. 2, pp. 569–582, 2022.
29. I. Nallathambi, R. Ramar, D. A. Pustokhin, I. V. Pustokhina, D. K. Sharma, and S. Sengan, “Prediction of influencing atmospheric conditions for explosion Avoidance in fireworks manufacturing Industry-A network approach,” Environ. Pollut., vol. 304, no. 7, p. 119182, 2022.
30. K. Kaliyaperumal, A. Rahim, D. K. Sharma, R. Regin, S. Vashisht, and K. Phasinam, “Rainfall prediction using deep mining strategy for detection,” in 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2021.
31. M. Awais, A. Bhuva, D. Bhuva, S. Fatima, and T. Sadiq, “Optimized DEC: An effective cough detection framework using optimal weighted Features-aided deep Ensemble classifier for COVID-19,” Biomed. Signal Process. Control, 2023, Press.
32. M. Yuvarasu, A. Balaram, S. Chandramohan, and D. K. Sharma, “A Performance Analysis of an Enhanced Graded Precision Localization Algorithm for Wireless Sensor Networks,” Cybernetics and Systems, pp. 1–16, 2023, Press.
33. P. P. Anand, U. K. Kanike, P. Paramasivan, S. S. Rajest, R. Regin, and S. S. Priscila, “Embracing Industry 5.0: Pioneering Next-Generation Technology for a Flourishing Human Experience and Societal Advancement,” FMDB Transactions on Sustainable Social Sciences Letters, vol.1, no. 1, pp. 43–55, 2023.
34. P. P. Dwivedi and D. K. Sharma, “Application of Shannon entropy and CoCoSo methods in selection of the most appropriate engineering sustainability components,” Cleaner Materials, vol. 5, no. 9, p. 100118, 2022.
35. P. P. Dwivedi and D. K. Sharma, “Assessment of Appropriate Renewable Energy Resources for India using Entropy and WASPAS Techniques,” Renewable Energy Research and Applications, vol. 5, no. 1, pp. 51–61, 2024.
36. P. P. Dwivedi and D. K. Sharma, “Evaluation and ranking of battery electric vehicles by Shannon’s entropy and TOPSIS methods,” Math. Comput. Simul., vol. 212, no.10, pp. 457–474, 2023.
37. P. P. Dwivedi and D. K. Sharma, “Selection of combat aircraft by using Shannon entropy and VIKOR method,” Def. Sci. J., vol. 73, no. 4, pp. 411–419, 2023.
38. P. Sindhuja, A. Kousalya, N. R. R. Paul, B. Pant, P. Kumar, and D. K. Sharma, “A Novel Technique for Ensembled Learning based on Convolution Neural Network,” in 2022 International Conference on Edge Computing and Applications (ICECAA), IEEE, Tamil Nadu, India, pp. 1087–1091, 2022.
39. R. Regin, Shynu, S. R. George, M. Bhattacharya, D. Datta, and S. S. Priscila, “Development of predictive model of diabetic using supervised machine learning classification algorithm of ensemble voting,” Int. J. Bioinform. Res. Appl., vol. 19, no. 3, p.11, 2023.
40. R. Tsarev et al., “Automatic generation of an algebraic expression for a Boolean function in the basis ∧, ∨, ¬,” in Data Analytics in System Engineering, Cham: Springer International Publishing, Switzerland, pp. 128–136, 2024.
41. R. Tsarev, B. Senapati, S. H. Alshahrani, A. Mirzagitova, S. Irgasheva, and J. Ascencio, “Evaluating the effectiveness of flipped classrooms using linear regression,” in Data Analytics in System Engineering, Cham: Springer International Publishing, Switzerland, pp. 418–427, 2024.
42. S. K. Sehrawat, “Empowering the Patient Journey: The Role of Generative AI in Healthcare,” International Journal of Sustainable Development Through AI, ML and IoT, vol. 2, no. 2, pp. 1-18, 2023.
43. S. K. Sehrawat, “The Role of Artificial Intelligence in ERP Automation: State-of-the-Art and Future Directions,” Transactions on Latest Trends in Artificial Intelligence, vol. 4, no. 4, p.12, 2023.
44. S. K. Sehrawat, “Transforming Clinical Trials: Harnessing the Power of Generative AI for Innovation and Efficiency,” Transactions on Recent Developments in Health Sectors, vol. 6, no. 6, pp. 1-20, 2023.
45. S. R. S. Steffi, R. Rajest, T. Shynu, and S. S. Priscila, “Analysis of an Interview Based on Emotion Detection Using Convolutional Neural Networks,” Central Asian Journal of Theoretical and Applied Science, vol. 4, no. 6, pp. 78–102, 2023.

46. S. S. Priscila and A. Jayanthiladevi, “A study on different hybrid deep learning approaches to forecast air pollution concentration of particulate matter,” in 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2023.
47. S. S. Priscila and S. S. Rajest, “An Improvised Virtual Queue Algorithm to Manipulate the Congestion in High-Speed Network”,” Central Asian Journal of Medical and Natural Science, vol. 3, no. 6, pp. 343–360, 2022.
48. S. S. Priscila, D. Celin Pappa, M. S. Banu, E. S. Soji, A. T. A. Christus, and V. S. Kumar, “Technological frontier on hybrid deep learning paradigm for global air quality intelligence,” in Cross-Industry AI Applications, IGI Global, USA, pp. 144–162, 2024.
49. S. S. Priscila, E. S. Soji, N. Hossó, P. Paramasivan, and S. Suman Rajest, “Digital Realms and Mental Health: Examining the Influence of Online Learning Systems on Students,” FMDB Transactions on Sustainable Techno Learning, vol. 1, no. 3, pp. 156–164, 2023.
50. S. S. Priscila, S. S. Rajest, R. Regin, and T. Shynu, “Classification of Satellite Photographs Utilizing the K-Nearest Neighbor Algorithm,” Central Asian Journal of Mathematical Theory and Computer Sciences, vol. 4, no. 6, pp. 53–71, 2023.
51. S. S. Priscila, S. S. Rajest, S. N. Tadiboina, R. Regin, and S. András, “Analysis of Machine Learning and Deep Learning Methods for Superstore Sales Prediction,” FMDB Transactions on Sustainable Computer Letters, vol. 1, no. 1, pp. 1–11, 2023.
52. S. S. Rajest, S. Silvia Priscila, R. Regin, T. Shynu, and R. Steffi, “Application of Machine Learning to the Process of Crop Selection Based on Land Dataset,” International Journal on Orange Technologies, vol. 5, no. 6, pp. 91–112, 2023.
53. S. Silvia Priscila, S. Rajest, R. Regin, T. Shynu, and R. Steffi, “Classification of Satellite Photographs Utilizing the K-Nearest Neighbor Algorithm,” Central Asian Journal of Mathematical Theory and Computer Sciences, vol. 4, no. 6, pp. 53–71, 2023.
54. Srinivasa, D. Baliga, N. Devi, D. Verma, P. P. Selvam, and D. K. Sharma, “Identifying lung nodules on MRR connected feature streams for tumor segmentation,” in 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), Tamil Nadu, India, 2022.
55. T. Shynu, A. J. Singh, B. Rajest, S. S. Regin, and R. Priscila, “Sustainable intelligent outbreak with self-directed learning system and feature extraction approach in technology,” International Journal of Intelligent Engineering Informatics, vol. 10, no. 6, pp.484-503, 2022.