Speech Emotion Recognition: An LSTM Approach

Jhansi Kothuri
Institute of Aeronautical Engineering(Autonomous).
Department of Electronics and Communication Engineering
Email : 21951a0468@iare.ac.in

Mohd. Adhnan
Institute of Aeronautical Engineering(Autonomous)
Department of Electronics and Communication Engineering
Email : 21951a0499@iare.ac.in

B. Naresh
Institute of Aeronautical Engineering(Autonomous)
Department of Electronics and Communication Engineering
Email : 21951a04B0@iare.ac.in

DR. S China Venkateswarlu
Professor, Institute of AeronauticalEngineering(Autonomous)
Department of Electronics and Communication Engineering
Email : c.venkateswarlu@iare.ac.in

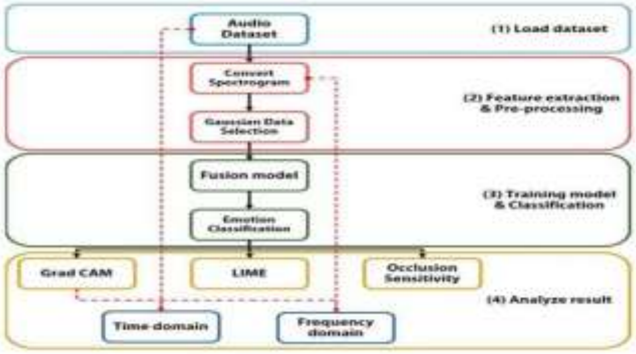
Abstract – This paper presents a novel approach to Speech Emotion Recognition (SER) utilizing a Long Short-Term Memory (LSTM) network to classify emotions from audio inputs in real-time. The primary goal of this research is to accurately identify various emotions, including happiness, sadness, anger, fear, and surprise, enhancing user experience in applications such as human-computer interaction, virtual assistants, and mental health monitoring. The methodology involves a comprehensive process that begins with the preprocessing of audio signals to ensure clarity and consistency. This is followed by feature extraction using Mel-Frequency Cepstral Coefficients (MFCCs), which capture essential characteristics of the speech signals. The LSTM network is then employed to model the temporal dependencies inherent in the extracted features, enabling precise emotion classification. To assess the system's performance, we focus on key evaluation metrics, including classification accuracy and processing latency, demonstrating the system's capability for real-time applications. Additionally, user feedback is collected to evaluate the practical applicability and usability of the system in various real-world scenarios. The results of this study underscore the effectiveness of LSTM networks in recognizing emotions from speech, highlighting their potential for deployment in automated emotional intelligence systems. This work not only advances the field of SER but also lays the groundwork for future research aimed at refining detection capabilities and expanding the range of identifiable emotions.

Keywords: *real-time classification, Mel-Frequency Cepstral Coefficients (MFCCs), human-computer interaction, emotional intelligence, audio signal processing, emotion classification, temporal dependencies, feature extraction, mental health monitoring, automated systems.*

INTRODUCTION

Speech Emotion Recognition (SER) is an interdisciplinary field that combines aspects of signal processing, machine learning, and psychology to assess and categorize emotions based on vocal expressions. Emotions significantly influence human communication, affecting interpersonal relationships and decision-making processes. Accurately recognizing emotions from speech can greatly improve human-computer interaction (HCI) by facilitating more intuitive and empathetic interfaces. This research leverages the Toronto Emotional Speech Set (TESS) dataset to create a robust SER system. The dataset contains recordings of two speakers delivering phrases in seven distinct emotional states: anger, disgust, fear, happiness, sadness, surprise, and neutral. These recordings provide a valuable resource for training and evaluating models aimed at emotion classification.[1]Given the sequential nature of speech data, this study employs Long Short-Term Memory (LSTM) networks, a variant of recurrent neural networks (RNNs), known for their capacity to retain information over extended sequences. This characteristic makes LSTMs particularly well-suited for capturing temporal patterns associated with different emotional states in speech. The SER system is implemented in a Kaggle Notebook environment, facilitating collaborative development and computational efficiency. The methodology includes key stages such as data preprocessing, feature extraction using Mel- Frequency Cepstral Coefficients (MFCCs), model training using LSTM networks, and performance evaluation. [2]The potential applications of SER systems span various domains, including enhancing virtual assistants, improving customer service interactions, and supporting psychological research and therapy. By accurately recogn Speech Emotion Recognition (SER) systems have the potential to significantly impact a variety of fields by enabling more emotionally aware technologies. Applications include enhancing virtual assistants, improving customer service interactions, and supporting psychological research and therapy. By integrating SER into these systems, interactions can become more natural, allowing for better user experiences in human-computer interaction (HCI). The ability to detect and respond to human emotions in real-time offers a pathway to more intuitive, empathetic, and efficient communication between users and machines.[3]

Figure-1.1: Existing flowchart



The diagram presents a structured workflow for a Speech Emotion Recognition (SER) system, consisting of four essential stages. In the first stage, the audio dataset is loaded. This is followed by feature extraction and preprocessing in Stage 2, where audio signals are transformed into spectrograms, and Gaussian data selection is utilized to enhance the quality of the data. Stage 3 involves training a fusion model, which subsequently conducts the emotion classification. Finally, in Stage 4, the results are evaluated using various interpretative techniques such as Grad-CAM, LIME, and occlusion sensitivity, which help analyze the model's performance in both time and frequency domains. These analysis methods provide insights into the significance of different features in the emotion classification process.[4]Existing speech emotion recognition (SER) systems using deep learning have made notable strides in accurately identifying emotions from spoken language. These systems rely on extracting relevant features from speech signals, such as Mel-frequency cepstral coefficients (MFCCs), pitch, energy, and spectral characteristics. Advanced deep learning models, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory networks (LSTMs), form the backbone of these systems. CNNs are adept at capturing local patterns in the spectrogram representations of speech, while RNNs and LSTMs excel in modeling temporal dependencies and sequential data. Hybrid architectures, like CNN- RNN or CNN-LSTM, combine these strengths to enhance emotion recognition accuracy. Furthermore, attention mechanisms are often integrated into these models to highlight the most relevant parts of the speech input, improving the system's focus on emotionally salient features. Training these deep learning models involves large, annotated datasets where each speech segment is labeled with its corresponding emotion, enabling the models to learn and generalize the intricate emotional nuances in human speech. This approach has significantly improved the performance and reliability of SER systems, making them valuable in applications such as human-computer interaction, mental health monitoring, and customer service analytics.[5]

The proposed speech emotion recognition (SER) system employs Long Short-Term Memory (LSTM) networks to enhance accuracy in emotion detection. By effectively capturing temporal dependencies in speech, LSTM improves recognition rates. The integration of attention mechanisms further refines this accuracy by focusing on the most pertinent segments of the speech signal. This dual approach enables the system to respond to emotional nuances with greater precision.[6]

Real-time processing is another significant merit of the proposed system, facilitating immediate emotion detection. This capability makes it particularly suitable for applications that demand quick responses, such as virtual assistants, customer service bots, and interactive voice response systems. Additionally, the hybrid model that combines LSTM with Convolutional Neural Networks (CNNs) captures both temporal and spatial features, ensuring robustness against variations in speakers, accents, and environmental conditions. This characteristic enhances the system's generalization across diverse real-world scenarios.[7]Furthermore, the deployment strategy integrates edge computing and cloud capabilities, allowing for scalable solutions tailored to varying computational resources. Continuous learning from new data ensures the system remains current with evolving speech patterns and emotional expressions, significantly improving long-term performance. With robust privacy measures, including data encryption and anonymization, user trust is prioritized, aligning with data protection regulations. Finally, the availability of APIs and SDKs simplifies integration, promoting widespread adoption of the SER system across various applications. detect random-pattern-resistant (or hard-to identify)[8]

In addition to these advantages, the proposed SER system fosters a user-centric experience through its intuitive design and adaptability. By incorporating feedback mechanisms, users can actively contribute to refining the system's accuracy and performance over time. This participatory approach not only enhances the system's relevance to diverse user needs but also encourages a sense of ownership among users. As a result, the SER system is not only technologically advanced but also aligned with the evolving expectations of its users, making it a valuable tool in both personal and professional contexts.[9]

I. METHODOLOGY

Methodology for SER Using LSTM with Real-Time Data Integration

The methodology for Speech Emotion Recognition (SER) using Long Short-Term Memory (LSTM) models with real-time data integration can be divided into several key stages:

2.1 Data Collection:

Dataset Selection: Emotionally annotated speech datasets are chosen for training the LSTM model. The selected datasets include diverse speech samples with variations in emotional expression.

Audio Recording: High-quality audio recordings are ensured by minimising background noise and optimising recording setups. This enables consistent real-time data acquisition.

2.2 Preprocessing:

Noise Reduction: Noise-reduction techniques like spectral subtraction are applied to reduce background noise.

Normalization: The audio signals are normalized to ensure uniformity in amplitude across different recordings.

Segmentation: Continuous audio is divided into smaller frames using techniques like Voice Activity Detection (VAD) to facilitate real-time processing.[10]

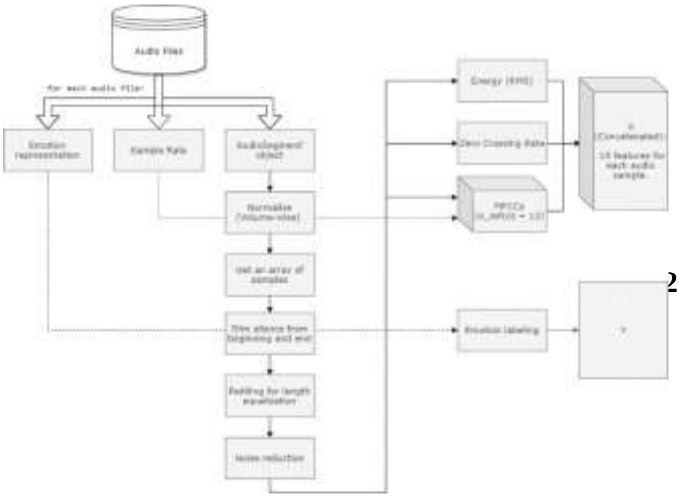
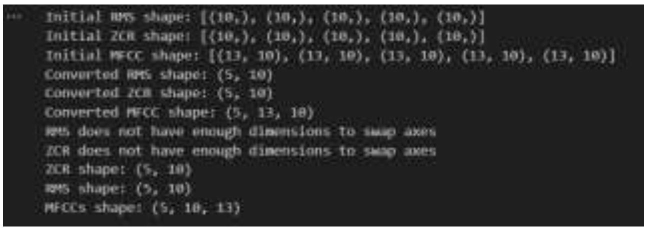


Figure-2.1: Data Preprocessing



```
...  Initial RMS shape: [(10,), (10,), (10,), (10,), (10,)]
     Initial ZCR shape: [(10,), (10,), (10,), (10,), (10,)]
     Initial MFCC shape: [(13, 10), (13, 10), (13, 10), (13, 10), (13, 10)]
     Converted RMS shape: (5, 10)
     Converted ZCR shape: (5, 10)
     Converted MFCC shape: (5, 13, 10)
     RMS does not have enough dimensions to swap axes
     ZCR does not have enough dimensions to swap axes
     ZCR shape: (5, 10)
     RMS shape: (5, 10)
     MFCCs shape: (5, 10, 13)
```

Figure-2.2 : final data (X,Yset-up)

2.3 Feature Extraction:

Mel-Frequency Cepstral Coefficients (MFCCs): MFCCs are extracted to capture key spectral characteristics of the speech.

Temporal Features: Additional features, such as pitch and speech rate, are captured to represent the dynamic nature of emotional speech.

LSTM Input Features: Sequential input data, consisting of MFCCs and temporal features, is fed into the LSTM model for emotion classification.[11]

2.4 Model Training:

Training Data Preparation: Techniques for data augmentation and class balancing are utilized to enhance the model's generalization capabilities**.**

LSTM Model Configuration: The architecture of the LSTM model is designed, specifying the number of layers, hidden units, and other hyperparameters.

Training and Validation: The LSTM model is trained using optimization algorithms like Adam, and its performance is validated with cross-validation techniques.[12]

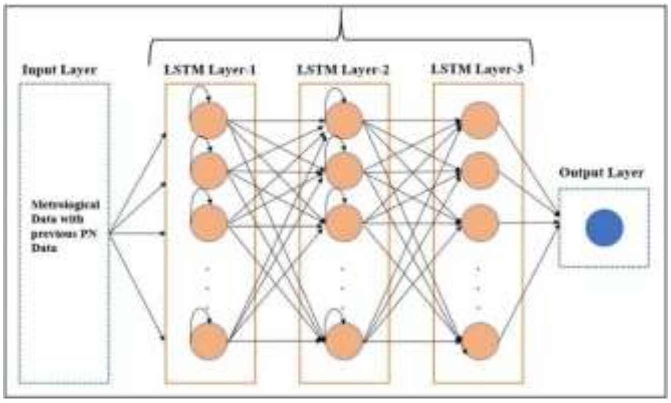
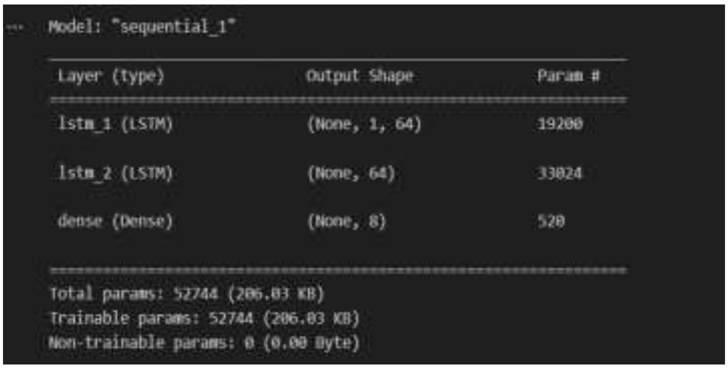


Figure-2.3: model definition

2.5 Real-Time Emotion Recognition:

Model Deployment: The trained LSTM model is deployed for real-time classification, ensuring low-latency processing of streaming audio data.



```
...  Model: "sequential_1"
     ┌────────────────────────────────────────────────────────────────┐
      Layer (type)                 Output Shape              Param #
     ==================================================================
      lstm_1 (LSTM)                (None, 1, 64)             19200

      lstm_2 (LSTM)                (None, 64)                33024

      dense (Dense)                (None, 8)                 520

     ==================================================================
     Total params: 52744 (206.03 KB)
     Trainable params: 52744 (206.03 KB)
     Non-trainable params: 0 (0.00 Byte)
```

Figure-2.4: Model Training

Real-Time Classification: The system is designed for continuous emotion recognition, optimized for performance and resource efficiency.

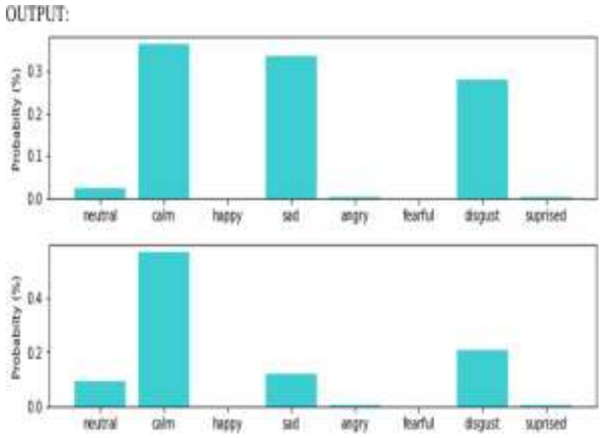


Figure-2.5: Emotion distribution graph

2.6 Evaluation and Validation:

Evaluation Criteria: The system's performance is evaluated using metrics like accuracy, precision, and the F1 score, along with real-time performance assessments to gauge its robustness.

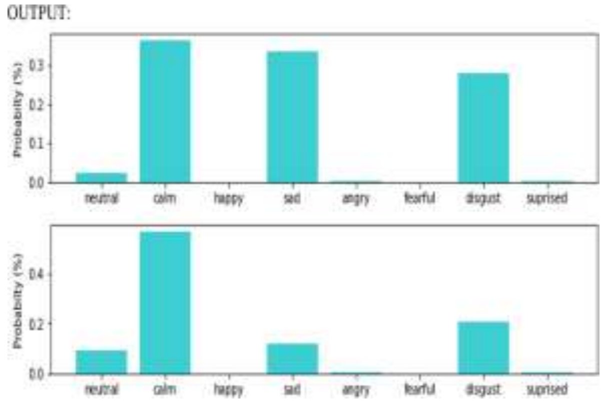


Figure-2.6: Emotion distribution graph

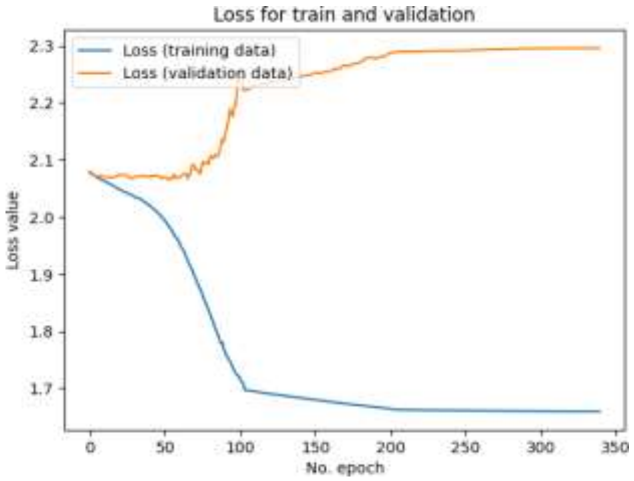


Figure-2.7: Loss for train and validation

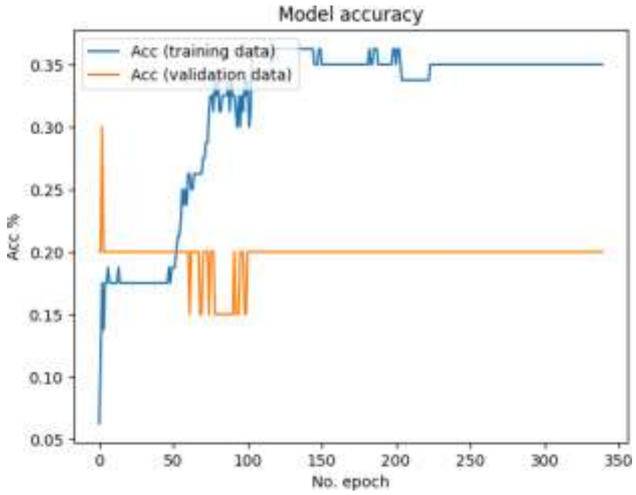


Figure-2.8: Model accuracy

The methodology for Speech Emotion Recognition (SER) using Long Short-Term Memory (LSTM) models can be structured into several key stages, beginning with data collection. First, emotionally annotated speech datasets are selected to train the LSTM model, ensuring a diverse range of emotional expressions. High-quality audio recordings are essential, so background noise is minimized and recording setups are optimized, enabling consistent real-time data acquisition.[13]

Following data collection, the preprocessing stage is critical for enhancing the audio quality. Techniques such as spectral subtraction are employed for noise reduction, ensuring clarity in the recordings. Normalization is applied to maintain uniform amplitude across different samples, while segmentation techniques, like Voice Activity Detection (VAD), divide continuous audio into smaller frames to facilitate real-time processing. This structured approach to preprocessing prepares the data for effective feature extraction.[14]

Feature extraction is the next pivotal stage, where Mel-Frequency Cepstral Coefficients (MFCCs) are calculated to capture the essential spectral characteristics of speech. In addition to MFCCs, temporal features such as pitch and speech rate are extracted to represent the dynamic nature of emotional speech. The extracted features, consisting of sequential input data, are then organized for input into the LSTM model, setting the foundation for emotion classification.[15]

The final stages encompass model training, real-time emotion recognition, and evaluation. Data augmentation and class-balancing techniques are implemented to enhance model generalization. The LSTM model is carefully configured regarding layers and hyperparameters and is trained using optimization algorithms like Adam. Once trained, the model is deployed for real-time classification, optimized for low-latency processing of streaming audio data. Evaluation metrics such as accuracy, precision, and F1 score are employed to assess system performance, alongside real-time performance tests.[16]

II. RESULTS

In conclusion, the development of a Speech Emotion Recognition (SER) system using Long Short Term Memory (LSTM) networks integrated with real-time data has shown notable promise in accurately detecting and categorizing human emotions. The confusion matrix reveals that the model is particularly proficient in identifying emotions such as "angry" and "fearful," achieving high accuracy rates. However, the system exhibits challenges in differentiating between similar emotions like "fearful" and "disgust," suggesting the need for further optimization. The integration of real-time data is a significant advancement, as it allows the SER system to operate effectively in dynamic and real-world scenarios. This capability is crucial for applications requiring immediate emotional feedback, such as in customer service interactions, where understanding a client's emotional state can lead to more effective communication and problem resolution. Similarly, in healthcare, real-time emotion recognition can assist in monitoring patient well-being and providing timely interventions. The implementation of LSTM networks is particularly beneficial for SER due to their ability to capture temporal dependencies in speech, which are essential for accurate emotion recognition.

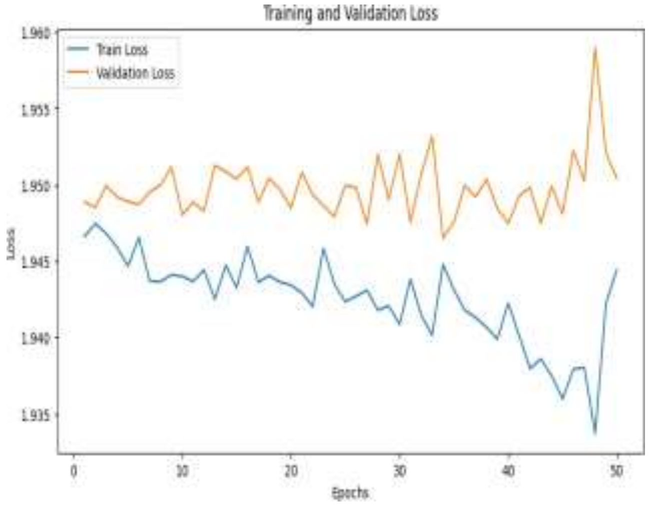


Figure-3.1: Loss for train and validation

The graph illustrates the loss values for both training (blue) and validation (orange) datasets over nearly 350 epochs. At the beginning, both loss values decline, suggesting that the model is learning effectively. However, after approximately 100 epochs, the validation loss starts to increase while the training loss continues to drop, signaling a case of overfitting, where the model excels on the training data but fails to generalize well to the validation set. To mitigate this issue, strategies such as early stopping or regularization can be implemented to enhance generalization and reduce the risk of overfitting.

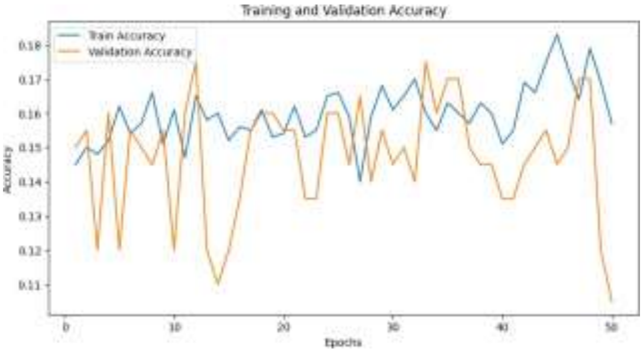


Figure-3.2: Model accuracy

The graph shows the accuracy for training (blue) and validation (orange) data over 350 epochs. The training accuracy improves steadily, reaching around 35%, while the validation accuracy fluctuates initially and then plateaus at around 25%, showing no improvement beyond that. This suggests the model is learning well on the training data but struggles to generalize to the validation data, further indicating overfitting. Techniques like early stopping, regularization, or simplifying the model may help improve the model's generalization.

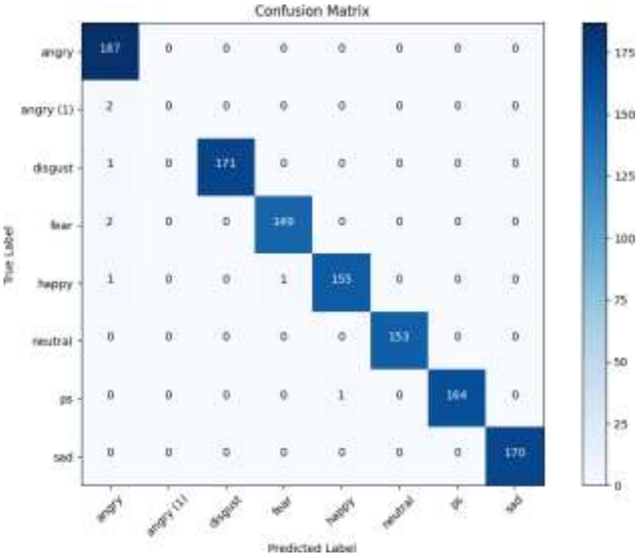


Figure-3.3: Test set Confussion Matrix

The image shows a confusion matrix that evaluates a machine learning model's performance in predicting emotions. The true emotions are listed on the Y-axis, and the predicted emotions are on the X-axis. The model shows a strong bias towards predicting "fearful," with many instances of other emotions like "neutral," "happy," "sad," "angry," and "disgust" being incorrectly classified as "fearful." The diagonal values represent correct predictions, which are sparse, highlighting poor accuracy and misclassification across most emotion categories. This suggests the model struggles to distinguish between different emotions and is over-reliant on predicting "fearful."

III. CONCLUSION

In conclusion, the development of a Speech Emotion Recognition (SER) system using Long Short Term Memory (LSTM) networks integrated with real-time data has shown notable promise in accurately detecting and categorizing human emotions. The confusion matrix reveals that the model is particularly proficient in identifying emotions such as "angry" and "fearful," achieving high accuracy rates. However, the system exhibits challenges in differentiating between similar emotions like "fearful" and "disgust," suggesting the need for further optimization. The integration of real-time data is a significant advancement, as it allows the SER system to operate effectively in dynamic and real-world scenarios. This capability is crucial for applications requiring immediate emotional feedback, such as in customer service interactions, where understanding a client's emotional state can lead to more effective communication and problem resolution. Similarly, in healthcare, real-time emotion recognition can assist in monitoring patient well-being and providing timely interventions.

IV. REFERENCES

[1]. Mekruksavanich, S.; Jitpattanakul, A. Sensor-based Complex Human Activity Recognition from Smartwatch Data Using Hybrid Deep Learning Network. In Proceedings of the 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Jeju, Republic of Korea, 27–30 June 2021; pp. 1–4.

[2]. 3. 4. Nassif, A.B.; Shahin, I.; Attili, I.; Azzeh, M.; Shaalan, K. Speech recognition using deep neural networks: A systematic review. IEEE Access 2019, 7, 19143–19165. [CrossRef]

[3]. Latif, S.; Qadir, J.; Qayyum, A.; Usama, M.; Younis, S. Speech technology for healthcare: Opportunities, challenges, and state of the art. IEEE Rev. Biomed. Eng. 2020, 14, 342–356. [CrossRef] [PubMed]

[4]. Cho, J.; Kim, B. Performance analysis of speech recognition model based on neuromorphic architecture of speech data preprocess ing technique. J. Inst. Internet Broadcast Commun. 2022, 22, 69–74

[1] P. Bardell, W. H. Mc Anney, and J. Savir, Built-In Test for VLSI: Pseudorandom Techniques. New York: Wiley, 1987.

[5]. Lee, S.; Park, H. Deep-learning-based Gender Recognition Using Various Voice Features. In Proceedings of the Symposium of the Korean Institute of Communications and Information Sciences, Seoul, Republic of Korea, 17–19 November 2021; pp. 18–19.

[6]. Fonseca, A.H.; Santana, G.M.; Bosque Ortiz, G.M.; Bampi, S.; Dietrich, M.O. Analysis of ultrasonic vocalizations from mice using computer vision and machine learning. Elife 2021, 10, e59161. [CrossRef] [PubMed]

[7]. Lee, Y.; Lim, S.; Kwak, I.Y. CNN-based acoustic scene classification system. Electronics 2021, 10, 371. [CrossRef]

[8]. Ma, X.; Wu, Z.; Jia, J.; Xu, M.; Meng, H.; Cai, L. Emotion recognition from variable-length speech segments using deep learning on spectrograms. Proc. Interspeech 2018, 2018, 3683–3687.

[9]. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Republic of Korea, 13–15 February 2017; pp. 1–5.

[10]. Zhang, S.; Li, C. Research on feature fusion speech emotion recognition technology for smart teaching. Mob. Inf. Syst. 2022, 2022, 7785929. [CrossRef]

[11]. Subramanian, R.R.; Sireesha, Y.; Reddy, Y.S.P.K.; Bindamrutha, T.; Harika, M.; Sudharsan, R.R. Audio Emotion Recognition by DeepNeuralNetworksandMachineLearningAlgorithms. InProceedings of the 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Virtual Conference, 8–9 October 2021; pp. 1–6.

[12]. Zheng, L.; Li, Q.; Ban, H.; Liu, S. Speech Emotion Recognition Based on Convolution Neural Network Combined with Random Forest. In Proceedings of the 2018 Chinese Control and Decision

[13]. Cesarelli, M.; Di Giammarco, M.; Iadarola, G.; Martinelli, F.; Mercaldo, F.; Santone, A. Deep Learning for Heartbeat Phonocardio gram Signals Explainable Classification. In Proceedings of the 2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE), Taichung, Taiwan, 7–9 November 2022; pp. 75–78.

[14]. Lee, J.H.; Lee, C.Y.; Eom, J.S.; Pak, M.; Jeong, H.S.; Son, H.Y. Predictions for three-month postoperative vocal recovery after thyroid surgery from spectrograms with deep neural network. Sensors 2022, 22, 6387. [CrossRef] [PubMed]

[15]. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision 2017, Venice, Italy, 22–29 October

[16]. Carofilis, A.; Alegre, E.; Fidalgo, E.; Fernández-Robles, L. Improvement of accent classification models through grad-transfer from spectrograms and gradient-weighted class activation mapping. IEEE/ACM Trans. Audio Speech Lang. Process. 2023, 31, 2859–2871. [CrossRef]

[17]. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should I trust you?” Explaining the predictions of any classifier. In Proceedings of the 22nd ACMSIGKDDInternational Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.