Speech Emotion Recognition with Hybrid CNN-LSTM and Transformers Models: Evaluating the Hybrid Model Using Grad-CAM

HMLS Kumari[1#], HMNS Kumari[2], and UMMPK Nawarathne[3]
[1]Computing Centre, Faculty of Engineering, University of Peradeniya, Sri Lanka
[2]Faculty of Information Technology and Communication Sciences, Tampere University, Finland
[3]Faculty of Computing, Sri Lanka Institute of Information Technology, Sri Lanka

[#]lihinisangeetha99@gmail.com

ABSTRACT Emotional recognition and classification using artificial intelligence (AI) techniques play a crucial role in human-computer interaction (HCI). It enables the prediction of human emotions from audio signals with broad applications in psychology, medicine, education, entertainment, etc. This research focused on speech-emotion recognition (SER) by employing classification methods and transformer models using the Toronto Emotional Speech Set (TESS). Initially, acoustic features were extracted using different feature extraction techniques, including chroma, Mel-scaled spectrogram, contrast features, and Mel Frequency Cepstral Coefficients (MFCCs) from the audio dataset. Then, this study employed a Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and a hybrid CNN-LSTM model to classify emotions. To compare the performance of these models, classical image transformer models such as ViT (Visual Image Transformer) and BEiT (Bidirectional Encoder Representation of Images) were employed on the Mel-spectograms derived from the same dataset. Evaluation metrics such as accuracy, precision, recall, and F1-score were calculated for each of these models to ensure a comprehensive performance comparison. According to the results, the hybrid model performed better than other models by achieving an accuracy of 99.01%, while the CNN, LSTM, ViT, and BEiT models demonstrated accuracies of 95.37%, 98.57%, 98%, and 98.3%, respectively. To interpret the output of this hybrid model and to provide visual explanations of its predictions, the Grad-CAM (Gradient-weighted Class Activation Mappings) was obtained. This technique reduced the black-box character of deep models, making them more reliable to use in clinical and other delicate contexts. In conclusion, the hybrid CNN-LSTM model showed strong performance in audio-based emotion classification.

INDEX TERMS Convolutional neural network, Grad-CAM, Hybrid model, Image transformers, Long Short-Term Memory, Speech emotion recognition.

I. INTRODUCTION

The most natural way for people to communicate is through speech, yet it can be difficult to infer emotions from speech, as the context is important, particularly in lengthy discussions. Emotion recognition is the first significant advancement in speech-driven computing systems, which are essential for improving human-computer interaction. As a result, speech- emotion recognition has grown in importance in human life and has a wide range of uses in areas such as automatic translation systems, call centers, health care, and human-computer interaction [1]-[6].

Additionally, over time, the study of speech emotion recognition has grown in popularity [7], [8]. The theory of emotion representation has laid the foundation for this emotion recognition research. It offers methods to acquire various emotional details using labelling data with the right targets. This helps machines to learn and predict emotions more effectively [9]. In previous studies, researchers have primarily focused on discovering the most effective features to represent emotions in machines. However, through developments intraditional machine learning and signal processing techniques, it has been able to better understand which features in voice signals are most useful for identifying emotions [10]. Emotion detection has recently moved toward this new deep learning- based methodology as deep learning has become more and more popular in fields like computer vision and speech recognition [11]. MFCC features, chroma features, Mel-scaled spectrogram features, and contrast features are types of audio features that are used in the field of audio signal processing, particularly in the analysis of music and speech, and these features help to represent different aspects of the audio signal effectively [12],[13]. Therefore, this study used MFCC features, chroma features, Mel-scaled spectrogram features, and contrast features to train three different models, such as CNN, LSTM, and a hybrid of CNN and LSTM.

An alternative method for classifying emotions using audio files involves converting the audio files into their corresponding Mel-spectrograms and training those images using image transformers. This proposed study utilized several image transformers, including ViT (Vision Transformer) and BEiT (Bidirectional Encoder Representation of Images), to train and classify emotions from audio files. While recent work has applied CNN-LSTM models to SER, most efforts confine themselves to accuracy improvements, ignoring the interpretability of predictions despite its importance in sensitive domains like healthcare and education. This study represents one of the first attempts to integrate Grad-CAM with a hybrid CNN-LSTM SER model, providing visual justifications of model predictions across Mel-spectrogram features. This integration enhances model interpretability, thereby promoting transparency and increasing confidence in its applicability for real-world deployment. Furthermore, the proposed hybrid approach is designed to effectively leverage the spatial feature extraction of CNN and the temporal sensitivity of LSTM on an optimized balance. In contrast to existing methods that weakly couple CNN and LSTM layers, we propose a well-optimized form that attains strong accuracy and interpretability. The proposed model performs more accurately on the TESS dataset than traditional architectures and demonstrates its strength with visual explanation techniques, a technique that has been relatively underexplored in SER literature. Finally, the proposed highest accuracy model is evaluated using the Grad-CAM Explainable AI technique to demonstrate how the model made predictions. This helps to reduce the black box nature of the deep learning model used in this study.

This paper is organized as follows. Section II provides a brief literature review of the related studies on speech-emotion recognition (SER). Section III describes the materials used and methodologies followed during the study. Section IV presents a comprehensive discussion of the results obtained, and Section V concludes this paper.

II. LITERATURE REVIEW

CNN and LSTM are one of the most popular deep-learning techniques. Researchers have recently used CNN and LSTM techniques with MFFCs to improve speech emotion recognition systems. Using the well-known Surrey Audio- Visual Expressed Emotion (SAVEE), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and Toronto Emotional Speech Set (TESS) datasets, N. P. Tigga and S. Garg [14] have employed a CNN and LSTM hybrid model to identify gender-biased emotions. Following the feature extraction using the MFCC approach, the hybrid network was applied to each dataset. The model detected seven distinct emotions: happy, sadness, anger, fear, neutral, disgust, and surprise. For the SAVEE, RAVDESS, and TESS datasets, they obtained accuracy rates of 91.66%, 85.89%, and 93.80%, respectively. Furthermore, H. Qazi and B. N. Kaushik [15] trained a CNN and LSTM hybrid model using spectrograms and the SAVEE dataset as inputs. They were able to recognize speech and emotions with an accuracy of 94.26% using this model.

Additionally, a comparative investigation of a voice emotion recognition system was carried out by L. Kerken et al. [16]. They used the Spanish and Berlin databases to extract the voice signal's modulation spectrum (MS) and MFCC. Their findings showed that all classifiers, using speaker normalization (SN) and feature selection, were able to reach an accuracy of 83% for the Berlin database. On the other hand, the RNN classifier with feature selection and without SN achieved the best accuracy of 94% for the Spanish database.

H. S. Kumbhar and S. U. Bhandari [17] proposed a SER model incorporating a component that blends IS09, a widely used feature for SER, with a Mel spectrogram. In this study, they created a more dependable dataset using the labelling results from the interactive emotional dyadic motion capture database (IEMOCAP). The model's experimental outcomes on this enhanced dataset verified a weighted accuracy (WA) of 73.3%.

In another study, Y. Yu and Y.J. Kim [18] reported a notable improvement in accuracy, achieving 98%, 91%, and 93% for speech emotion recognition with the TESS dataset. By utilizing the Vision Transformer (ViT), a lightweight model, they effectively demonstrated its potential in enhancing speech emotion recognition systems.

C.S.A. Kumar et al. [19] proposed a study titled "Speech emotion recognition using CNN-LSTM and Vision Transformer," which compared and evaluated CNN-LSTM and ViT for speech emotion recognition systems. For this instance, they used the EMO-DB dataset, which is an assortment of poignant voice recordings from Berlin's Technical University, containing seven different emotions and ten people. However, the authors obtained an accuracy of 88.05% and 85.36% for the suggested CNN-LSTM and ViT models, respectively.

Many affiliated scholars and institutions have underscored the need to address the research gap in evaluating and analyzing the existing knowledge of SER systems. To address this lack of interpretability in speech emotion recognition (SER) models, we developed a model that combines CNN, LSTM, and hybrid CNN-LSTM and compared their performance to traditional Transformer

models, such as ViT (Vision Transformers) and BEiT (Bidirectional Encoder Representation from Images), using the TESS dataset. While many dominant SER models focus on achieving higher accuracy, they often function as black boxes, thereby compromising reliability and trustworthiness in sensitive or high-stakes applications. Our approach directly fills this gap by integrating Grad-CAM visualizations with the top-accuracy hybrid model to provide clear explanations of how the model is making its predictions. This enhances model transparency and enables more trust in SER applications, particularly in clinical and educational settings.

III. METHODOLOGY

A. Data

This study used the Toronto Emotional Speech Set (TESS), which was extracted from an online data repository [20]. The dataset was comprised of 2800 audio recordings, where each record consisted of various words and emotion combinations. Two actresses, aged 26 and 64, were enlisted to create these voices using the carrier phrase "Say the word _," s. A set of 200 target words representing seven distinct emotions: disgust, wrath, fear, happiness, pleasant surprise, sadness, and neutrality was spoken by them.

The steps of the methods carried out during this study are depicted in Figure 1. Firstly, acoustic features were extracted from the dataset.

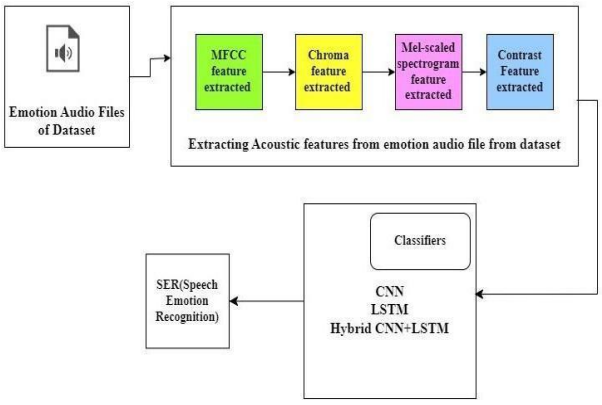


Figure 1. The steps of the proposed model for CNN, LSTM, and hybrid CNN and LSTM

B. Acoustic Feature Extraction

Speech signal requires preprocessing to remove background noise before identifying key elements in the speech. This can be done by dividing speech into manageable sections, and it helps to deal with the challenges of working with sound characteristics. Extracting features from speech makes it easier to work with, providing a concise and reliable representation of the original speech [14]. This study employed different feature extraction methods, including MFCC, chroma, Mel-scaled spectrogram, and contrast, to extract the acoustic features from the audio dataset.

1) Mel-Frequency Cepstral Coefficients Feature Extraction:

MFCC is the most commonly used method in most recent works. In speech, the vocal tract's impact is evident in a brief look at the power spectrum of sound. The mel unit is used to measure the pitch or frequency of a signal. The formula to convert speech from frequency (f) to Mel is given by (1).

$$Mel\ (f) = 2595 * \log_{10}(1 + f/100) \quad (1)$$

MFCC converts unclear speech signals with poor frequency into understandable signals with better resolution frequencies. This process involves seven basic steps, such as MFCC pre- emphasis, framing, windowing, Fast Fourier Transform (FFT), Mel filter bank, computing discrete cosine transform (DCT), and delta energy [21].

2) Chroma and Mel-Scaled Spectrogram Feature Extraction:

Chromogram helps to understand the musical tones in an audio signal, focusing on the 12 pitch classes, which are beneficial for identifying both the harmony and melody of the audio. This will lead to fine pitches that present different emotions in audio. To get Chroma features, Short-Time Fourier Transforms (STFTs) can be applied on the audio data utilizing the Librosa library [22]. Moreover, a spectrogram uses FFT analysis to show how a sound's pitch changes over time. It creates a Mel spectrogram for each part by dividing the pitch range into Mel scale frequencies and then separating the primary frequencies [23].

3) Contrast Feature Extraction: In speech emotion processing,

contrast feature extraction involves identifying and quantifying differences or variations in specific aspects of the audio signal [23]. These aspects could include characteristics such as pitch, intensity, spectral content, or timing. For example, in the context of emotion speech, contrast feature extraction might involve detecting differences in pitch between different segments of speech or variations in intensity levels within a sentence.

After acoustic feature extraction, three different classification methods, such as CNN, LSTM, and a hybrid of CNN and LSTM, were employed on the data.

C. Classification Methods

1) Convolutional Neural Network: A convolutional neural network is an artificial neural network that is important for

understanding the patterns in speech emotion audio [23], [24]. The key aspect of CNN models is the layers. In this study, the convolution layer, max pooling layer, flatten layer, dense layers, and dropout layers were used to train the model. The CNN model included an additional Conv1D layer with 32 filters and a kernel size of three, a max pooling layer with a pool size of two, a flatten layer, a dense layer with 128 neurons and a "relu" activation function, a dropout layer with a rate of 0.2, a dense layer with 64 neurons and a dropout rate, and a Conv1D layer with 32 filters and a kernel size of three.

2) Long Short-Term Memory: Recurrent Neural Networks (RNNs) are like repetition in neural networks, where information from previous steps is used in the current step. However, they face difficulties with remembering information that occurred too long ago. For tasks like speech recognition, where the context is important, there is a need for a solution that can retain and use context information effectively. Long Short-Term Memory Networks are a type of RNN designed to address this issue [23]. In speech recognition, where the signal is continuous over time, LSTM enhances the connection between adjacent time frames, capturing the emotional characteristics more effectively and improving recognition performance [23]. A LSTM layer of 128 neurons and 'return_sequences=True' to return the full sequence of outputs, rather than indicating only the output at the last time step, was used in this study. This was followed sequentially by another LSTM layer of 64 neurons, a dense layer with 64 neurons or units with 'relu' as an activation function, and a dropout layer with a rate of 0.3. Finally, a dense output layer was applied with a softmax activation function.

3) Hybrid of CNN and LSTM: While recurrent neural networks, such as LSTM, retain data from previous steps, making them well-suited for sequential data, convolutional neural networks process spatial data. In other words, CNNs identify patterns in space, whereas LSTMs identify patterns that develop over time. LSTMs are the preferred method for speech processing since speech signals develop sequentially. The proposed model's hybrid CNN+LSTM structure included a Conv1D layer with 64 filters, a kernel size of 3, the activation function "relu", a 128-neuron LSTM layer, a Max Pooling layer with pooling size 2, a dropout layer with a rate of 0.2, and sequentially return sequences as "False". The final addition was a dense output layer with a softmax activation function. Figure 2 depicts the design of the CNN+LSTM hybrid model. However, the CNN model, LSTM model, and hybrid of CNN and LSTM model were trained for 50 epochs, 30 epochs, and 50 epochs, respectively.

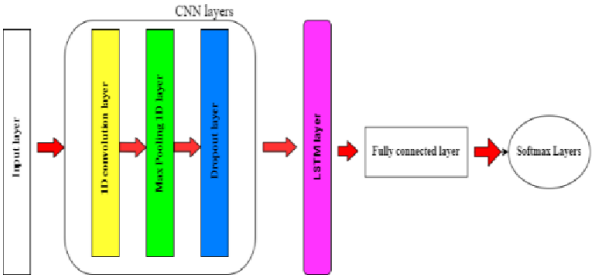


Figure 2. CNN and LSTM hybrid model architecture

To compare the performance of these classification models, two widely used classical image transformer models were employed on the Mel spectrograms obtained from the same dataset. Using image transformer models, the emotion of audio can be classified after being transformed into spectrograms. TESS audio recordings were first transformed into Mel spectrograms, and then image transformer models were applied. The Mel spectrogram is a crucial tool for providing transformer models with sound information in a way that mimics human auditory perception. To create a Mel spectrogram, raw audio waveforms are processed through a series of filter banks. The result is a 128 x 128 matrix for each sample, representing 128 filter banks and 128-time steps, encapsulating both the frequency content and the temporal dynamics of the audio clip [25].

D. Transformer-Based Vision Models

1) Vision Transformer (ViT): The work of vision transformers in computer vision originated due to the success of transformers in Natural Language Processing (NLP). Unlike other methods in computer vision, the image is split into a sequence of patches in the initial stage. In Vision Transformers, an image is split into small patches, and each small patch is considered a 'word' in a sentence. These patches are processed using a standard transformer model, which is similar to the way that the text is handled in Natural Language Processing (NLP) tasks. As a result, ViT performs better for many image classification tasks.

2) Bidirectional Encoder Representation from Images (BEiT): The BEiT is a widely used method of applying transformers to computer vision tasks. BEiT adapts the principle of Bidirectional Encoder Representation of Transformers (BERT) models, originally used for natural language processing, and applied to image processing. Before pre-training, BEiT initially creates an 'Image tokenizer' that breaks an original image into small visual pieces based on its learned set of patterns. Both picture patches and visual tokens are used to view each image during pre-training. After that, some of these picture patches are randomly masked with a unique mask embedding.

Before applying these transformer models to classify emotions, data preprocessing techniques such as random-sized crop, normalization, and resizing were applied to the Mel- spectrogram. These models were then trained with a batch size of 32 and 20 epochs.

To evaluate the performance of the models discussed in this study, accuracy, precision, recall, and F1-score were calculated with 5-fold stratified cross-validation in order to ensure balanced and reliable assessment across the distributions of the classes.

E. Classification Metrics

Classification metrics play a major role in this study. To evaluate the performance of CNN, LSTM, hybrid CNN and LSTM, ViT, and BEiT, accuracy, precision, recall, and F1- score were calculated using Equations (2), (3), (4), and (5), respectively.

$$Accuracy = (True\ Positive + True\ Negative) / (True\ Positive + True\ Negative + False\ Positive + False\ Negative) \quad (2)$$

$$Precision = (True\ Positive) / (True\ Positive + False\ Positive) \quad (3)$$

$$Recall = (True\ Positive) / (True\ Positive + False\ Negative) \quad (4)$$

$$F1\text{-}score = (2 \times Precision \times Recall) / (Precision + Recall) \quad (5)$$

After calculating the evaluation metrics, the model that achieved the highest accuracy, precision, recall, and F1-score was selected as the best-performing model, and to explain the output of this model, Grad-CAM, which is an explainable artificial intelligence (XAI) technique, was used.

F. The Explainable AI Technique - Grad-CAM

The complex artificial intelligence (AI) models are being applied in many industries. As a result, XAI is crucial to evaluate the predictions of those AI models, especially in fields like healthcare and finance. The main aim of these XAI models is to make the model's decision-making process transparent, which helps build trust in the model's predictions. In healthcare, explainability will increase trust of clinicians towards the predictions made by the AI model and can improve the security of patients. CAM (Class Activation Mapping) is a one XAI technique [27]. This used a global average pooling layer to replace a fully connected layer. This method results in a heatmap of an image for a specific class. This heatmap could explain how the CNN categorized the image as a particular class. CAM cannot generate a heatmap using intermediate layers and is not compatible with transfer learning models. To overcome these limitations, Grad-CAM was introduced. Unlike CAM, Grad-CAM does not require a global average pooling layer but operates on the gradients of the target class score concerning the feature maps. These gradients are averaged globally to obtain importance weights, which are used similarly to CAM to compute a weighted sum over the feature maps. Lastly, a ReLU activation is applied to focus the visualization on the most effective positive regions, and thus Grad-CAM becomes a more generalizable and applicable tool for model interpretability [28].

IV. RESULTS AND DISCUSSION

This study used the TESS, which contained 28000 audio recordings created by two actresses. Firstly, acoustic features were extracted from the dataset, and three different classification methods, including CNN, LSTM, and a hybrid of CNN and LSTM, were applied. After that, evaluation metrics were calculated for each of these models, and Table 1 depicts the accuracies achieved by these three models.

Table 1. The accuracies of the CNN, LSTM, and the hybrid of CNN and LSTM models

Model	Layers used	Learning rate	Accuracy
CNN	Conv1D layer Max pooling layer Flatten layer Dropout layer Dense layer	0.001	95.37%
LSTM	LSTM layer Dropout layer Dense layer	0.001	98.57%
A hybrid of CNN and LSTM	Conv1D layer Max pooling layer Dropout layer LSTM layer	0.001	99.01%

According to Table 1, it is clear that the hybrid of the CNN and LSTM model performs better than the CNN model and the LSTM model, with an accuracy of 99.01%. However, the LSTM model achieved an accuracy of 98.57%, which is slightly lower than that of the hybrid model. In addition, to evaluate the performance of these three models in detail, precision, recall, and F1-score were calculated and are presented in Table 2.

When considering Table 2, it is observed that the hybrid model achieved the highest performance across three evaluation metrics, with a precision of 99.30%, a recall, and an F1-score of 99.29%. This highlights the hybrid model's accuracy and robustness in classification when compared to two individual CNN and LSTM models.

Table 2. The precision, recall, and F1-score of the CNN, LSTM, and the hybrid of CNN and LSTM models

Model	Precision	Recall	F1-score
CNN model	95.30%	95.30%	95.37%
LSTM model	98.55%	98.56%	98.57%
CNN+LSTM model	99.30%	99.29%	99.29%

However, to compare the performance of these classification models, two major classical image transformer models were applied to the Mel spectrograms obtained from the TESS dataset, and the accuracies calculated for these models are demonstrated in Table 3.

Table 3. The accuracies of the ViT and BEiT models

Model	Learning rate	Accuracy
ViT	0.0001	98%
BEiT	0.001	98.3%

When considering Table 3, it is clear that the BEiT model achieved the highest accuracy of 98.3% when compared to that of the ViT model. In addition to accuracy, precision, recall, and F1-score were calculated to provide a more comprehensive model evaluation. The calculated metrics are depicted in Table 4.

Table 4. The precision, recall, and F1-score of the ViT and BEiT models

Model	Precision	Recall	F1-score
ViT Model	98.00%	98.00%	98.00%
BEiT model	98.31%	98.30%	98.30%

As demonstrated in Table 4, the BEiT model outperforms the ViT model across all metrics by achieving a precision of 98.31%, a recall of 98.30%, and a F1-score of 98.30%. This indicates that the BEiT model performs well compared the ViT model in classification tasks. However, it is evident from Tables 1, 2, 3 and 4 that the CNN and LSTM hybrid model performs better with the dataset by obtaining higher accuracy as well as notable precision, recall, F1-score than the other suggested models. This indicates that this proposed hybrid model has a higher classification power when compared to both the novel and classical models. In order to provide additional evidence for hybrid model's robustness in effectively classifying emotion categories, classification report as well as the confusion matrix were obtained as in Figures 3, and 4 respectively.

According to the classification report and confusion matrix shown in Figures 3, and 4 respectively, it is clear that the hybrid model performs well with minimal misclassifications as across all classes demonstrating its strong and balanced performance in emotion recognitions. Despite the model's higher values in evaluation metrics, it is essential to examine the accuracy and loss curves of this hybrid model. Therefore, the accuracy curve and the loss curve obtained for the hybrid model are depicted in Figures 5 and 6, respectively.

```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        82
           1       0.99      1.00      0.99        80
           2       0.99      1.00      0.99        80
           3       1.00      0.99      0.99        80
           4       0.99      1.00      0.99        81
           5       1.00      0.99      0.99        80
           6       0.99      1.00      0.99        80

    accuracy                           0.99       563
   macro avg       0.99      0.99      0.99       563
weighted avg       0.99      0.99      0.99       563


Weighted Precision: 0.9930
Weighted Recall: 0.9929
Weighted F1-Score: 0.9929
```

Figure 3. Classification report of the hybrid of the CNN and LSTM model

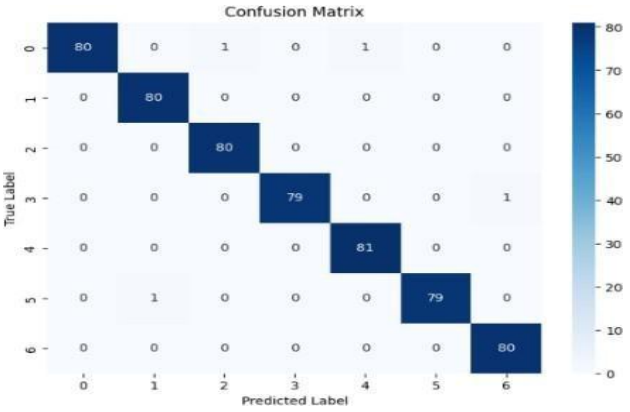


Figure 4. Confusion matrix of the hybrid of the CNN and LSTM model

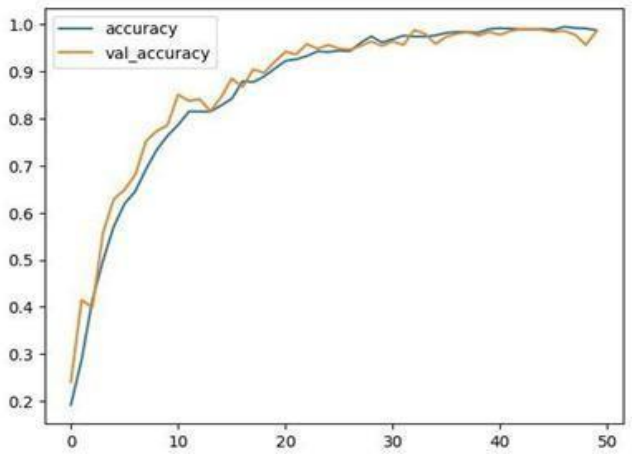


Figure 5. Accuracy curve of the hybrid of the CNN and LSTM model

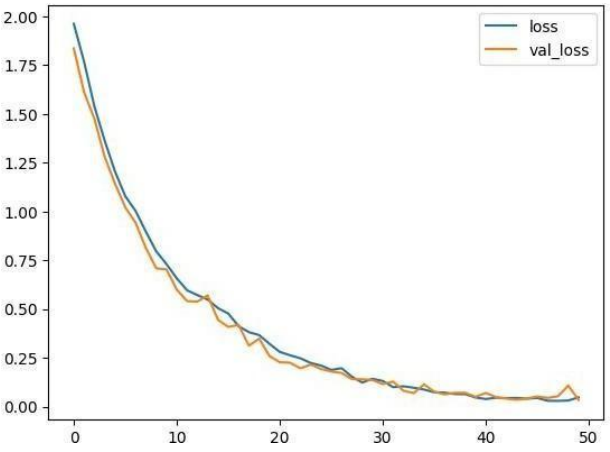
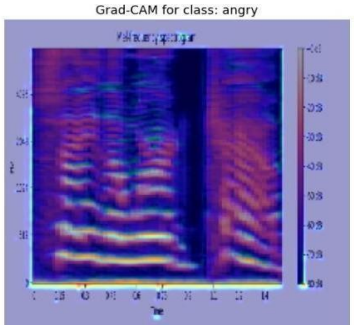


Figure 6. Loss curve of the hybrid of the CNN and LSTM model

According to Figures 5 and 6, it is clear that the suggested model learns effectively over time. Therefore, this study concludes that the hybrid CNN and LSTM model, under the given conditions, is more suitable for speech emotion recognition. However, to observe how this proposed model made its final predictions, Figures 7,8, 9, 10,11,12, and 13 were generated to show how Grad CAM evaluation was performed for classes angry, fear, disgust, happy, sad, surprise, and neutral predictions in the test dataset, respectively. To explain briefly, additional metrics are shown in each Grad-CAM figure.

The Grad-CAM visualization for the "angry" emotion class, which is represented in Figure 7, highlights significant time- frequency regions in the Mel-spectrogram that contributed mainly to the classification decision of the model. The red and yellow high-activation regions indicate where the model was focusing, which likely corresponds with speech features typical in anger, such as high energy and pitch. The model labelled "angry" with 77.49% confidence, concentrating on a small but crucial region (1.74% of the spectrogram) between 0.22–2.59 seconds and 367.0–6899.1 Hz. Masking this region resulted in a very slight decrease in confidence (0.54%), confirming its importance. The model's attention to these specific features allows for the interpretation and verification of its emotion classification decision.



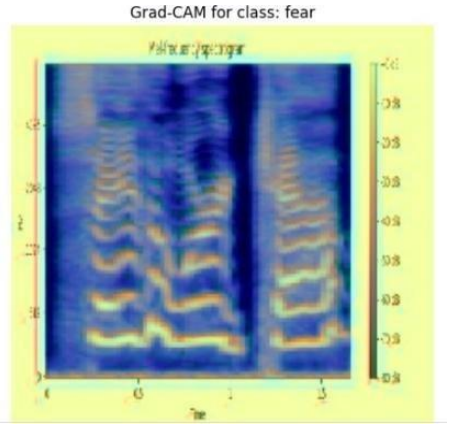

```
Grad-CAM Analysis for class: angry
Metric                Value
------------------------------
Model Confidence      77.49% (angry)
 Grad-CAM Max Activation  = 0.96
Activated Area          = 1.74%,
 Focus Region (time: 0.22-2.59s, freq: 367.0-6899.1Hz)
Confidence Drop (after masking focus region) = -0.54%
mean_activation       0.07643380
max_activation        0.95776683
activated_area_pct    1.74
entropy               8.311
time_focus_range      (0.22, 2.59)
freq_focus_range      (367.0, 6899.1)

Class probabilities:
angry: 0.7749
disgust: 0.0581
fear: 0.0383
happy: 0.0150
neutral: 0.0417
sad: 0.0208
surprise: 0.0511
```

Figure 7. Grad-CAM output for angry class test images

Figure 8 shows the Grad-CAM representation for the fear class. The Grad-CAM visualization of the fear class highlights significant time-frequency regions in the Mel-spectrogram that contributed significantly to the classification decision of the model. The amount of yellow and red colour can be seen throughout the graph, and this is as a result of high-activation regions indicating where the model was focusing, which likely corresponds with speech features typical in fear. The model labelled "fear" with 54.78% confidence, concentrating on a small but crucial region (42.20% of the spectrogram) between 0.0–2.97 seconds and 0.0–7926.6 Hz. Masking this region resulted in a drop in confidence (19.66%), confirming its importance.

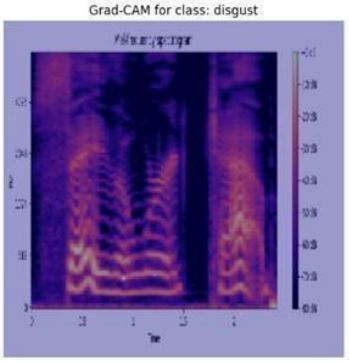
In addition, Figures 8, 12, and 13 show more yellow in colour since the model is more confident that those regions contributed to its final decision. The model confidence levels are very high in these Grad-CAM figures.



```
Grad-CAM Analysis for class: fear
Metric               Value
------------------------------
Model Confidence      54.73% (fear)
 Grad-CAM Max Activation  = 0.98
Activated Area          = 44.20%,
 Focus Region (time: 0.0–2.97s, freq: 0.0–7926.6Hz)
Confidence Drop (after masking focus region) = 19.66%
mean_activation      0.44908139
max_activation       0.98126942
activated_area_pct   44.20
entropy              9.310
time_focus_range     (0.0, 2.97)
freq_focus_range     (0.0, 7926.6)

Class probabilities:
angry: 0.0625
disgust: 0.0671
fear: 0.5473
happy: 0.1223
neutral: 0.0554
sad: 0.0178
surprise: 0.1277
```

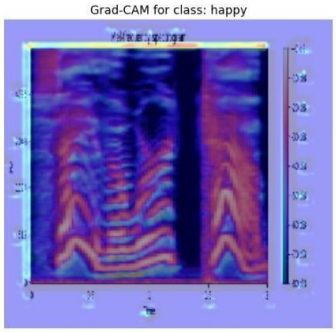
Figure 8. Grad-CAM output for fear class test images




```
Grad-CAM Analysis for class: disgust
Metric               Value
------------------------------
Model Confidence      52.23% (disgust)
 Grad-CAM Max Activation  = 0.53
Activated Area          = 0.01%,
 Focus Region (time: 1.05–1.05s, freq: 6091.7–6091.7Hz)
Confidence Drop (after masking focus region) = 0.00%
mean_activation      0.00008228
max_activation       0.52723080
activated_area_pct   0.01
entropy              1.018
time_focus_range     (1.05, 1.05)
freq_focus_range     (6091.7, 6091.7)

Class probabilities:
angry: 0.0208
disgust: 0.5223
fear: 0.0502
happy: 0.1162
neutral: 0.0984
sad: 0.0924
surprise: 0.0996
```

Figure 9. Grad-CAM output for disgust class test images




```
Grad-CAM Analysis for class: happy
Metric               Value
------------------------------
Model Confidence      71.81% (happy)
 Grad-CAM Max Activation  = 0.94
Activated Area          = 1.62%,
 Focus Region (time: 0.17–2.72s, freq: 293.6–7559.6Hz)
Confidence Drop (after masking focus region) = -0.99%
mean_activation      0.12068452
max_activation       0.93679237
activated_area_pct   1.62
entropy              8.858
time_focus_range     (0.17, 2.72)
freq_focus_range     (293.6, 7559.6)

Class probabilities:
angry: 0.0179
disgust: 0.0326
fear: 0.0432
happy: 0.7181
neutral: 0.0663
sad: 0.0266
surprise: 0.0953
```

Figure 10. Grad-CAM output for happy class test images

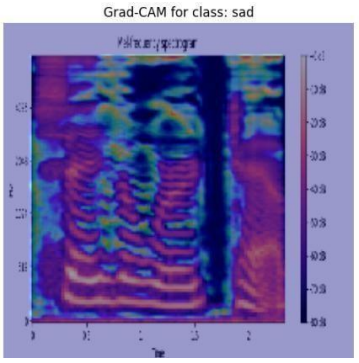
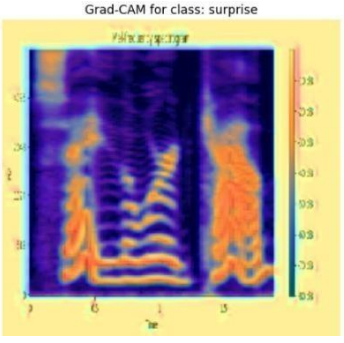


Figure 11. Grad-CAM output for sad class test images

```
Grad-CAM Analysis for class: sad
Metric               Value
------------------------------
Model Confidence      71.87% (sad)
 Grad-CAM Max Activation  = 0.92
Activated Area          = 3.64%,
 Focus Region (time: 0.25–2.37s, freq: 733.9–6678.9Hz)
Confidence Drop (after masking focus region) = 3.09%
mean_activation      0.08298303
max_activation       0.91536552
activated_area_pct   3.64
entropy              7.960
time_focus_range     (0.25, 2.37)
freq_focus_range     (733.9, 6678.9)

Class probabilities:
angry: 0.0119
disgust: 0.0536
fear: 0.0152
happy: 0.0407
neutral: 0.1308
sad: 0.7187
surprise: 0.0291
```

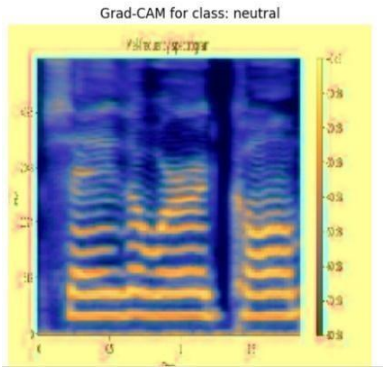
Figure 11. Grad-CAM output for sad class test images



```
Grad-CAM Analysis for class: surprise
Metric               Value
------------------------------
Model Confidence      79.13% (surprise)
 Grad-CAM Max Activation  = 0.98
Activated Area          = 51.18%,
 Focus Region (time: 0.0-2.97s, freq: 0.0-7926.6Hz)
Confidence Drop (after masking focus region) = 55.17%
mean_activation      0.41372573
max_activation       0.98044980
activated_area_pct   51.18
entropy              9.080
time_focus_range     (0.0, 2.97)
freq_focus_range     (0.0, 7926.6)

Class probabilities:
angry: 0.0091
disgust: 0.0315
fear: 0.0158
happy: 0.0541
neutral: 0.0767
sad: 0.0214
surprise: 0.7913
```

Figure 12. Grad-CAM output for surprise class test images




```
Grad-CAM Analysis for class: neutral
Metric               Value
------------------------------
Model Confidence      91.82% (neutral)
 Grad-CAM Max Activation  = 0.93
Activated Area          = 51.21%,
 Focus Region (time: 0.0-2.97s, freq: 0.0-7926.6Hz)
Confidence Drop (after masking focus region) = 74.96%
mean_activation      0.46167940
max_activation       0.93491471
activated_area_pct   51.21
entropy              9.281
time_focus_range     (0.0, 2.97)
freq_focus_range     (0.0, 7926.6)

Class probabilities:
angry: 0.0064
disgust: 0.0086
fear: 0.0047
happy: 0.0113
neutral: 0.9182
sad: 0.0213
surprise: 0.0296
```

Figure 13. Grad-CAM output for neutral class test images

Therefore, we can conclude that the model predictions are correct and that we can use the Grad-CAM to evaluate the proposed model. These methods can reduce the black box nature of AI model predictions and can increase trust among AI models, clinicians, and researchers.

However, despite the promising results, this study faces several limitations. The dataset, which was used in this study, contained audio files generated by two actresses. As a result, these models fail to capture the gender variations in the speech emotion recognition process. Furthermore, by limiting it to two actresses of two different ages, it only identifies speech emotions at these ages. Therefore, hidden patterns at different age limits may not be included in the trained model. In addition, this study is restricted to seven emotions, which may not fully capture the complexity and subtlety of human emotional expression, potentially restricting the model's ability to generalize to real-world emotional variability. Furthermore, this study compared the performance of the proposed model with two other classification methods and classical image transformers, limiting the scope of evaluation. However, this study could be used as a benchmark for future research, and these limitations could be addressed in further studies using not only human audio but also sounds related to birds, animals, and environmental contexts, etc. Moreover, enhancing this study by including audio from different age groups, genders, and cultural backgrounds would provide more robust results. Furthermore, to improve the generalizability as well as interpretability, future research could improve this study using diverse datasets and evaluating the model's performance across various datasets. In addition, this study opens the pathway for implementing and evaluating other neural network models to recognize speech emotions. This would lead to more generalized results and conclusions, thereby contributing to speech emotion recognition systems.

V. CONCLUSION

This study implemented CNN, LSTM, and a CNN-LSTM hybrid model using the TESS dataset. MFCC, chroma, Mel- scaled spectrogram, and contrast feature extraction techniques were used to extract the acoustic features prior to implementing the above-mentioned models. The CNN and LSTM hybrid model achieved a remarkable accuracy of 99.01%, while the CNN and LSTM models separately achieved a better performance of 95.37% and 98.57%, respectively. To assess the performance of these classification models, two classical image transformer models, including ViT and BEiT, were employed on the Mel-spectrogram of audio files from the TESS dataset, and it was identified that the ViT model acquired an accuracy of 98%, while the BEiT model acquired a better accuracy of 98.3%. However, these two transformer models were unable to outperform the hybrid model in terms of accuracy. Moreover, to evaluate the output of this model thoroughly, the Grad-CAM, which is a novel explainable AI technique, was used in this study. While the TESS dataset is a widely used, well-annotated benchmark for SER tasks, it was collected from only two female speakers aged 26 and 64. Thus, our findings might not generalize well across genders, ages, and speaking styles. This reduced demographic coverage might bring bias or reduce robustness when applied to large populations. Therefore, in future work, we plan to include other datasets such as Ryerson Audio-Visual Database of Emotional Speech and

Song (RAVDESS) or Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) that offer higher speaker variability. Overall, despite these limitations, the current work is a strong baseline and demonstrates the potential of the proposed hybrid CNN-LSTM model when applied to well-annotated emotional speech data.

REFERENCES

[1] W. Zheng, W. Zheng, and Y. Zong, "Multi-scale discrepancy adversarial network for crosscorpus speech emotion recognition," Virtual Reality & Intelligent Hardware, vol. 3, no. 1, pp. 65–75, Feb. 2021, doi: https://doi.org/10.1016/j.vrih.2020.11.006.

[2] G. A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," International Journal of Speech Technology, vol. 23, no. 1, pp. 45–55, Jan. 2020, doi: https://doi.org/10.1007/s10772-020-09672-4.

[3] J. H. L. Hansen and D. A. Cairns, "ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect environments ☆," Speech Communication, vol. 16, no. 4, pp. 391–422, Jun. 1995, doi: https://doi.org/10.1016/0167-6393(95)00007-b.

[4] C. Spencer et al., "A Comparison of Unimodal and Multimodal Measurements of Driver Stress in Real-World Driving Conditions," PsyArXiv (OSF Preprints), Jun. 2020, doi: https://doi.org/10.31234/osf.io/en5r3.

[5] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," IEEE Xplore, May 01, 2004. https://ieeexplore.ieee.org/document/1326051 (accessed Feb. 26, 2021).

[6] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and D. M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on bio-medical engineering*, vol. 47, no. 7, pp. 829–837, Jul. 2000, doi: https://doi.org/10.1109/10.846676.

[7] "M. Young, 'The Technical Writers Handbook,' Mill Valley, CA University Science, 1989. - References - Scientific Research Publishing," Scirp.org, 2021. https://www.scirp.org/reference/referencespapers?referenceid=99878 6 (accessed Sep. 01, 2024).

[8] "Speech and Multimedia Transmission Quality (STQ); Requirements for Emotion Detectors used for Telecommunication Measurement Applications; Detectors for written text and spoken speech TECHNICAL SPECIFICATION." Accessed: Sep. 01, 2024. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/103200_103299/103296/01.01.0 1_60/ts_103296v010101p.pdf

[9] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," International Journal of Speech Technology, vol. 15, no. 2, pp. 99–117, Jan. 2012, doi: https://doi.org/10.1007/s10772-011-9125-1.

[10] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, pp. 572–587, Mar. 2011, doi: https://doi.org/10.1016/j.patcog.2010.09.020.

[11] M. Ren, W. Nie, A. Liu, and Y. Su, "Multi-modal Correlated Network for emotion recognition in speech," Visual Informatics, vol. 3, no. 3, pp. 150–155, Sep. 2019, doi :https://doi.org/10.1016/j.visinf.2019.10.003.

[12] "Indexof/class/archive/cs/cs224n/cs224n.1214/reports," Stanford.edu, 2021. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/reports/ (accessed Sep. 01, 2024).

[13] M. M. Rezapour Mashhadi and K. Osei-Bonsu, "Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest," PloS one, vol. 18, no. 11, p. e0291500, 2023, doi: https://doi.org/10.1371/journal.pone.0291500.

[14] N. P. Tigga and S. Garg, "Speech Emotion Recognition for multiclass classification using Hybrid CNN-LSTM," International Journal of Microsystems and Iot, vol. 1, pp. 9–17, 2023.

[15] H. Qazi and B. N. Kaushik, "A hybrid technique using CNN+ LSTM for speech emotion recognition," International Journal of Engineering and Advanced Technology (IJEAT), vol. 9, no. 5, pp. 1126–1130, 2020.

[16] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cleder, "Automatic Speech Emotion Recognition Using Machine Learning," social media and Machine Learning, Mar. 2019, doi: https://doi.org/10.5772/intechopen.84856.

[17] H. S. Kumbhar and S. U. Bhandari, "Speech Emotion Recognition using MFCC features and LSTM network," Sep. 2019, doi: https://doi.org/10.1109/iccubea47591.2019.9129067

[18] Y. Yu and Y.-J. Kim, "Attention-LSTM-Attention Model for Speech Emotion Recognition and Analysis of IEMOCAP Database," Electronics, vol. 9, no. 5, p. 713, Apr. 2020, doi: https://doi.org/10.3390/electronics9050713.

[19] "Speech Emotion Recognition Using CNN-LSTM and Vision Transformer," Dntb.gov.ua, 2023. https://ouci.dntb.gov.ua/en/works/7WQ2BrPl/ (accessed Sep. 01, 2024).

[20] "Toronto emotional speech set (TESS)," www.kaggle.com. https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech- set-tess

[21] L. Toledo, A. Luiz, and J. Fiais, "A Deep Learning Approach for Speech Emotion Recognition Optimization Using Meta- Learning," Electronics, vol. 12, no. 23, pp.

4859–4859, Dec. 2023, doi: https://doi.org/10.3390/electronics12234859.

[22] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, pp. 569–576.

[23] S. Waldekar and G. Saha, "Wavelet Transform Based Mel- scaled Features for Acoustic Scene Classification.," in INTERSPEECH, 2018, vol. 2018, pp. 3323–3327.

[24] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[25] "Baeldung on CS," www.baeldung.com, Mar. 19, 2021. https://www.baeldung.com/cs/.

[26] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," arXiv preprint arXiv:2106.08254, 2021.

[27] Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[28] Jeong, Seung-Min, et al. "Exploring Spectrogram-Based Audio Classification for Parkinson's Disease: A Study on Speech Classification and Qualitative Reliability Verification." Sensors 24.14 (2024): 4625.

AUTHOR BIOGRAPHY

H.M.L.S. Kumari

H.M.L.S. Kumari, an Instructor at the Computer Center, Faculty of Engineering, University of Peradeniya, completed a Bachelor of Science Honours degree in Computer Science at the Faculty of Applied Sciences, Vavuniya Campus, University of Jaffna. Her research interests include Deep Learning, Computer Vision, Artificial Intelligence, and Explainable AI, particularly in the health sector.

H.M.N.S. Kumari

H.M.N.S. Kumari is a volunteer researcher currently engaged in numerous projects. She completed her Master of Science degree in Computing Sciences, specializing in Statistical Data Analytics at Tampere University, Finland, and her Bachelor of Science degree in Statistics and Operations Research at the University of Peradeniya, Sri Lanka. Her research interests include Statistical Data Analytics, Bayesian Statistics, Machine Learning, Deep learning, Computer Vision, Explainable Artificial Intelligence, Signal Processing, and Public Health.

U.M.M.P.K. Nawarathne

U.M.M.P.K. Nawarathne, an Assistant Lecturer at the Faculty of Computing, Sri Lanka Institute of Information Technology, completed her Bachelor of Science Honors Degree in Information Technology, specializing in Data Science, in 2021. Her main research interests lie in the areas of Generative Artificial Intelligence and applied Data Science.