A Deep Neural Network Approach Integrating CNN and BiLSTM-Transformer Architectures for Emotion Recognition from Speech

MSc Research Project
Data Analytics

Divyansh Anand
Student ID: 22240217

School of Computing
National College of Ireland

Supervisor: Abdul Qayum

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Divyansh Anand
Student ID:	22240217
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Abdul Qayum
Submission Due Date:	12/08/2024
Project Title:	A Deep Neural Network Approach Integrating CNN and BiLSTM-Transformer Architectures for Emotion Recognition from Speech
Word Count:	7065
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	15th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Deep Neural Network Approach Integrating CNN and BiLSTM-Transformer Architectures for Emotion Recognition from Speech

Divyansh Anand
22240217

Abstract

The goal of speech emotion recognition is to make human-computer interaction more efficient in several areas such as customer service, entertainment industry, human-computer interaction, healthcare, and education. Previous work in speech emotion analysis presented some issues like limited choice of features, model complexity, noise variability, and insufficient data samples, which negatively affected the prediction of emotions. This paper provides an in-depth study of speech emotion recognition using a hybrid deep neural network architecture that combines 1-D Convolutional Neural Network (CNN) and BiLSTM-Transformer models to analyze data from the Ravdess and Crema-D datasets. To make the datasets appropriate for emotion detection, all were preprocessed by means of librosa library to get rid of non-speech segments. Important sound characteristics such as Mel-Frequency Cepstral Coefficients(MFCC), Root Mean Square Energy (RMSE), and Zero Crossing Rate (ZCR) were extracted to get the spectral characteristics, intensity of feelings, and dynamic features present in the emotions. In order to improve model's generalization and robustness noise injection, time stretching, time shifting as well as pitch shifting have been applied during data augmentation. The proposed model leverages the strengths of both CNN and BiLSTM-Transformer components. The proposed model's 1-D CNN captures local patterns in sound whereas the BiLSTM-Transformer handles sequence data and complex hierarchical structures of audio. Various datasets such as Ravdess and Crema-D were used to train and test the performance of the model in the emotion classification task. To evaluate model's performance, training-validation accuracy graph, confusion matrix and overall metrics which include precision, recall, and F1-score are used. Ravdess dataset achieved a high accuracy of 83.3%, surprise, angry, disgust and sad were among those emotions which this model identified with great accuracy. Crema-D dataset achieved 82.7% accuracy, and showed solid performance in detecting neutral, fear, and happy emotions. Accuracy plots between training and validation demonstrated good generalization for the unseen data and confusion matrices highlighted the emotion categories where improvement could be made.

Keywords: Speech Emotion Detection, Convolutional Neural Network, Bidirectional Long Short-Term Memory, Transformer, Ravdess, Crema-D

1 Introduction

It is easy for humans to determine the emotions behind words, but to build a system that recognizes such emotions with high accuracy is a game changer for entire sectors. Speech Emotion Recognition is a technology where we train machines over speech signals to predict human emotions which involves understanding of various acoustic characteristics of voice signal. Voice has different features like pitch, frequency and intensity which all together contribute in analysis of voice audio signals. Machines with better speech emotion recognition systems can enhance human computer interactions, improve customer service and can contribute to education and healthcare.(Cowen et al. (2019))

Earlier, there were two-step speech emotion recognition systems. First, they would decide which of the basic features to extract from the input – pitch or pace, for example. In the second step, one would use machine learning algorithms such as, Support Vector Machines classifiers which can be linear types but a complex linear type, Bayesian Network which is also a linear classifier or even use Gaussian Mixture Models which are nonlinear classifiers to analyze the feature selected to recognize the emotion. (Pulatov et al. (2023)) While these systems were serving reasonable purposes well enough at the time, they had some problems. One problem was that choosing the right feature from the speech was a troublesome task and it was dependent on the consultants as they had to have knowledge of the acoustic structure of the speech and the meaning of the emotions and so on. Another problem was that these algorithms could not study dependencies in speech and temporal characteristics easily. (Li et al. (2024))

Subsequently, deep learning algorithms emerged as a rather strong contender for SER. They allow learning of patterns from basic speech data by the use of machines. Different methods are, for example, Convolutional Neural Networks (CNNs), which are exceptionally good at the detection of patterns in audio signals, whereas Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) are predominantly good at capturing temporal dependencies. (Issa et al. (2020))(Parry et al. (2019))

Recently, hybrid model architectures have been introduced to perform speech emotion recognition tasks efficiently. (Kim and Lee (2023)) A hybrid model architecture combines different machine learning models to leverage their respective strengths for improved performance. By integrating more than one technology together, data can be processed more efficiently and hidden patterns can be extracted and used for the analysis. For instance, a model combined with CNNs, LSTMs and Transformer can make use of all the powers associated with the individual models. CNNs are highly effective in extracting hierarchical data and can capture all the low level and high level features from the audio input, whereas LSTM models perform well in analysing long term dependency in sequential data and can capture temporal dependency in the data. Lastly, Transformers with their self-attention mechanisms, efficiently capture dependencies between input and output sequences. Together, the hybrid model can combine the strengths of all the individual models.

This research proposes the new approach which is the combination of One-Dimensional Convolutional Neural Network (CNN), Bi-Directional Long Short-Term Memory Network (Bi-LSTM), and the Transformer Model. The input audio data which is used here is first fed into the CNN to extract low level and high level features from the audio. Using two LSTMs to take data forwards and backwards improves the overall context understanding of the model. These can effectively cooperate with the CNN layers to analyze temporal factors and sequential structure of the speech. Lastly, Transformer mechanism is used to encode the language which gives opportunity to enhance understanding of the language and possibly increase the exactness of the model when recognizing emotions.

Research Question: To find the performance of a deep neural network architecture that integrates one-dimensional Convolutional Neural Network, Bidirectional Long Short Term-Memory (LSTM), and Transformer for robust speech emotion detection from features of speech.

The objective of this research is to evaluate the performance of the proposed integrated deep neural model combining Convolutional Neural Network, Bidirectional Long Short-Term Memory, and Transformer for audio data based on accuracy. By Building a generalized model that will be utilized for two different datasets namely Ravdess and Crema-D will bring diversity to the model. This research will extend the knowledge arising from past research work in the area of integrated deep neural networks.

The research is meant to connect a computer with machine learning systems that can identify emotions from sound. It does this through the use of two main datasets for emotion recognition; CREMA-D and RAVDESS. These datasets are publicly available for research and study purposes. (Livingstone and Russo (2018))(Cao et al. (2014)) The research is meant to connect a computer with machine learning systems that can identify emotions from sound. It does this through the use of two main datasets for emotion recognition; CREMA-D and RAVDESS. The same system of emotion recognition can be used in various fields, as it includes audio samples drawn from different genders, ages, and races. This system is helpful in monitoring mental health and improving therapy because it offers an accurate assessment of emotions. In education, however, it may enable teachers to adjust their teaching methods based on emotional responses so as to enhance student engagement. Moreover, this technology is also capable of personalizing feedback in entertainment or customer service thereby leading to greater customer satisfaction.

In this study, Section 2 will discuss the literature review of individual and integrated machine learning algorithms. Section 3 will comprehensively detail the proposed methodology to process speech data. The results of this case study are discussed in Section 4. A detailed discussion about the challenges faced and what could have been better is highlighted in Section 5. Section 6 will discuss the conclusion and highlight all the important details discussed in the case study.

2 Related Work

In the field of speech emotion recognition, many researchers have explored different kinds of individual and integrated models with multiple engines to analyze audio data have been introduced. This section reviews the effectiveness of three models namely Convolutional Neural Networks, Bidirectional Long Short-Term Memory Networks, Transformer models as well as hybrid approaches that integrate these architectures in detecting emotions from speech data.

2.1 Individual Machine Learning Architectures

One-dimensional CNN works by performing functions to overlapping parts of the input sequence and thus enables the learning of hierarchical features. This makes it possible for the CNN to capture both the low-level features including the pitch and tone and the high-level features like the intonation patterns which help in the determination of the emotional state of the subject. Also, CNNs are known to be parameter-efficient models where weights are shared while having fewer hyperparameters than other deep learning models which helps them learn faster and are better generically trained engines when analyzing signals with high dimensions such as audio signals.

Several studies have highlighted the success of the CNN model when used for speech emotion recognition. In research by Issa et al. (2020), a one-dimensional CNN model using MFCC features achieved an accuracy of 71.61% for the Ravdess dataset. Similarly, Mocanu and Tapu (2022) proposed a deep neural network with CNN which achieved 82.91% accuracy on the same Ravdess data. Additionally, Mountzouris et al. (2023) proposed a model integrating a convolutional neural network with an attention mechanism which gave prediction results of 74% for the Savee database and 77% for the Ravdess. They used MFCCs from the audio input and highlighted as an important feature to be considered while analyzing audio data, it works similarly to how humans interpret sound. The addition of the attention model helped the algorithm focus more on emotional information segments of the audio as compared to non-speech parts of the audio. Another notable study by Ayadi and Lachiri (2023) combined conv1D with the LSTM model which significantly improved the algorithm in terms of limiting over-fitting and enhancing the stability of integrated model performance compared to individual models. The proposed model achieved training accuracies of 87.97% and 66.51% for audio-song and audio-speech data.

Long Short-Term Memory Networks are the improved versions of standard Recurrent Neural Networks that minimize the vanishing and exploding gradient problem and are very efficient in processing sequential data with long-term dependency. This capability is particularly relevant for SER tasks since identifying emotion-bearing features that can occur in a time frame of any duration is critical. The Bidirectional-LSTM variant further enhances the capability by processing data in both forward and backward directions, capturing more information for processing.

For detecting emotions from speech, LSTM models are effectively used to capture temporal dependencies in the audio data. Parry et al. (2019) utilized LSTM networks with MFCC features and achieved an accuracy of 59.67% on the Emo-DB dataset and 53.97% on the Ravdess dataset. Another interesting research by Senthilkumar et al. (2022) showcased an approach with CNN and BiLSTM to process the speech signals to identify the emotional state of the speaker. The system achieved an accuracy of 77.02% for the Ravdess dataset. Another hybrid approach by Andayani et al. (2022) which merged LSTM and Transformer architecture. Multi-head Attention mechanism used in the transformer encoder layer helps the model capture features from different sequence positions, thus learning long-term dependencies. MFCC is applied to extract the features of the speech input and the LSTM-Transformer classifier is applied for classification tasks to enhance the recognition performance on different datasets. The proposed hybrid model reached 75.62% recognition success with the Ravdess dataset.

The Transformer is a neural network architecture proposed in the work of Vaswani et al. (2017) ‘Attention is all you need’. It is most suitable for sequence-to-sequence transformation tasks like machine translation and does not have any recurrent layers in its architecture but has used attention mechanisms as in the original model. The Transformer design records links between input and output patterns by employing self-attention mechanism. In tasks related to machine translation, Transformer has provided significant results because of the use of self-attention mechanism and feed-forward neural networks.

Jing et al. (2021) applied the transformer to Log-Mel Filterbank Energies features and achieved a performance of 74.9% on the Emo-DB dataset. The Transformers have a parallelized architecture that reduces the training and inference time during processing. This makes the Transformer architecture more powerful than traditional Recurrent Neural Networks and Long Short-Term Networks. However, transformers are still computationally expensive due to the need to recalculate the history of the Transformer at each step in time.

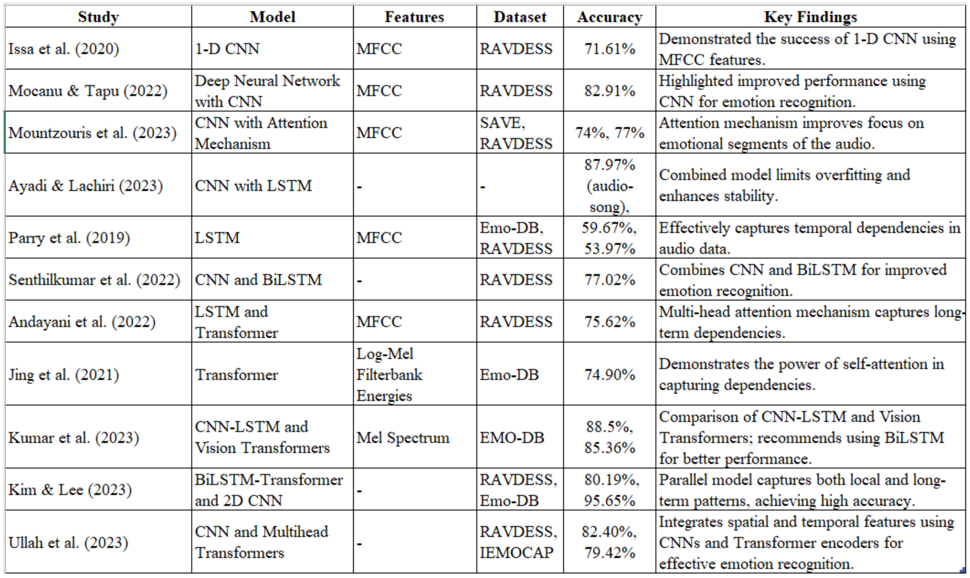
2.2 Hybrid Machine Learning Architectures

Some hybrid models that integrate the power of two or more models have surpassed in terms of performance. In a study by Kumar et al. (2023) two hybrid models were compared: Mel spectrum techniques used in CNN-LSTM and Vision Transformers. The method employed in the study was a unidirectional LSTM and the study only trained on one dataset emanating from EMO-DB and with only four emotions. The performance of both models was good and the best of the two models is the CNN-LSTM which had an accuracy of 88.5%, and for Mel spectrum Vision Transformer the score has been improved up to 85.36%. However, they are many suggestions through which the given approach may be enhanced. For instance, replacing Bi-LSTM by a unidirectional LSTM may help in increasing the performance. Moreover, using more than one dataset can increase the variability and check the performance of the model with the various speakers and voice inputs. Further work could include the integration of these models and the use of datasets which are larger so as to investigate emotions from speech.

Similarly, a study by Kim and Lee (2023) presents a powerful approach for speech emotion recognition. It introduces a novel approach of using BiLSTM-Transformer and 2D CNN models in parallel for the same input sequence. This study leverages the strengths of all the algorithms to capture long-term patterns and local patterns of speech effectively. The 2D CNN model excels at identifying specific parts of speech data while BiLSTM-Transformer identifies the long-term structure of speech data ensuring deep analysis of the input data. This integrated model achieved a high accuracy of 95.65% for Ravdess and 80.19% for the Emo-DB dataset. To further enhance the model, potential improvements include improvising data flow, detailed visualization of the pre-processing part of raw input data, and testing the models against different datasets can provide valuable insights for refining and can make the model generalized.

According to a study conducted by Ullah et al. (2023), a new approach of SER was introduced by integrating Convolutional Neural Networks and multi-head transformers. The research explained how to integrate both the spatial and temporal features through the process of convolutional neural networks and transformer encoders. Spatial features were considered using one or more parallel CNNs while temporal features were considered through a transformer encoder and the model gave an accuracy of 82.40% for eight emotions on the RAVDESS dataset and 79.42% on the IEMOCAP dataset.

Table 1 given below, concludes all the research papers discussed above.



Study	Model	Features	Dataset	Accuracy	Key Findings
Issa et al. (2020)	1-D CNN	MFCC	RAVDESS	71.61%	Demonstrated the success of 1-D CNN using MFCC features.
Mocanu & Tapu (2022)	Deep Neural Network with CNN	MFCC	RAVDESS	82.91%	Highlighted improved performance using CNN for emotion recognition.
Mountzouris et al. (2023)	CNN with Attention Mechanism	MFCC	SAVE, RAVDESS	74%, 77%	Attention mechanism improves focus on emotional segments of the audio.
Ayadi & Lachiri (2023)	CNN with LSTM	-	-	87.97% (audio-song),	Combined model limits overfitting and enhances stability.
Parry et al. (2019)	LSTM	MFCC	Emo-DB, RAVDESS	59.67%, 53.97%	Effectively captures temporal dependencies in audio data.
Senthilkumar et al. (2022)	CNN and BiLSTM	-	RAVDESS	77.02%	Combines CNN and BiLSTM for improved emotion recognition.
Andayani et al. (2022)	LSTM and Transformer	MFCC	RAVDESS	75.62%	Multi-head attention mechanism captures long-term dependencies.
Jing et al. (2021)	Transformer	Log-Mel Filterbank Energies	Emo-DB	74.90%	Demonstrates the power of self-attention in capturing dependencies.
Kumar et al. (2023)	CNN-LSTM and Vision Transformers	Mel Spectrum	EMO-DB	88.5%, 85.36%	Comparison of CNN-LSTM and Vision Transformers; recommends using BiLSTM for better performance.
Kim & Lee (2023)	BiLSTM-Transformer and 2D CNN	-	RAVDESS, Emo-DB	80.19%, 95.65%	Parallel model captures both local and long-term patterns, achieving high accuracy.
Ullah et al. (2023)	CNN and Multihead Transformers	-	RAVDESS, IEMOCAP	82.40%, 79.42%	Integrates spatial and temporal features using CNNs and Transformer encoders for effective emotion recognition.

Table 1: Literature Review Summary

This study introduces a new deep neural network architecture by integrating convolutional neural networks, bi-directional long short-term memory, and transformer models. While previous research either relied on extracting temporal dynamics from LSTM architectures or focused on localized features using CNNs for sequential data analysis, this research will utilize the power of all the integrated technologies. This study incorporates BiLSTMs property of long term sequence analysis with the 1D CNN pattern recognition feature for local patterns required to create a better understanding of the emotional content in the speech. This approach does not only enhance the strength of each model but also address the issues of each model; therefore, enhancing the effectiveness of the over all model in identifying emotions from the speech data.

3 Methodology

This research followed the Knowledge Discovery in Databases or KDD approach which includes stages from dataset selection, processing, and transformation to finding useful insights from the data. All the stages of KDD followed in this research have been explained in the below subsections.

3.1 Data Selection and Understanding

After going through several research papers mentioned in the literature review, several datasets have been discovered. The most commonly used datasets in the field of speech analysis are RAVDESS and CREMA. These datasets are publicly accessible through kaggle website. This research compares the performance of the hybrid deep neural network model against the given datasets which will help in drawing interesting conclusions.

- **Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS):**

  The RAVDESS dataset by Livingstone and Russo (2018) contains data that aims to facilitate studies on emotional recognition and communication in both speech and song contexts. There are 24 English-speaking actors behind the recordings, out of which 12 are females and 12 are males. It includes speech expression categories such as 'surprise', 'happy', 'angry', 'disgust', 'fear', 'sad', and 'neutral'. The original dataset contains three types of recordings namely audio-visual (AV), audio-only (AO) and video-only (VO), for this research only audio speech files are considered. Altogether, the RAVDESS dataset serves as a valuable dataset for researchers in psychology, neuroscience and clinical therapy by enhancing the understanding of human expressions in different fields.

- **Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D):**

  CREMA-D introduced by Cao et al. (2014) includes 7442 original clips from 91 actors. The dataset comprises facial and vocal emotional expressions that are spoken for different emotional states. It has 48 male and 43 female voices between the ages from 20 to 74 from a different variety of races and ethnicities. There are 6 emotion categories, namely angry, disgust, fear, happy, sad, and neutral in the dataset. Overall, the CREMA-D dataset is a dependable and detailed resource that records various emotions through multi-modal approach which is significant for the study of speech emotion analysis.

3.2 Data Preparation

Data Preparation phase helps organize the data collected from different data sources Ravdess and Crema-D. During this phase, raw audio files from different datasets are read, processed, and organized into a structured format. By extracting emotion labels from filenames, consolidating the data into data frames, and checking for null values in data frames, this research created a generalized structure for the data to be processed further. Figure 1 and Figure 2 below, give a sample of the raw data after the data preparation phase. It highlights the emotion counts for every category in the Ravdess and Crema datasets.

3.3 Data Preprocessing

Preprocessing is a crucial phase for audio analysis, this study removed non-speech segments from the original audio during this phase. Non-speech segments are important to remove as they lack speech-related information and can introduce noise into emotion classification tasks. As a result, their removal can enhance the efficiency of emotion recognition systems. In this research, effects.trim function from librosa library (McFee et al. (2020))was used as part of the preprocessing process to remove unwanted portions of

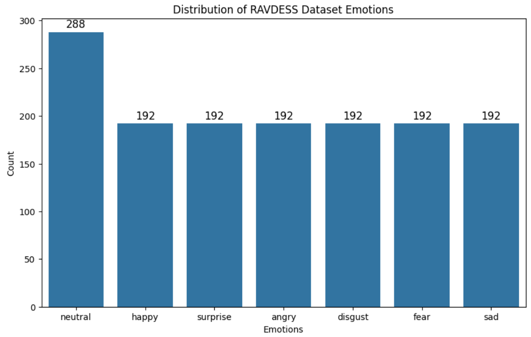


Figure 1: Count of emotions for Ravdess Dataset

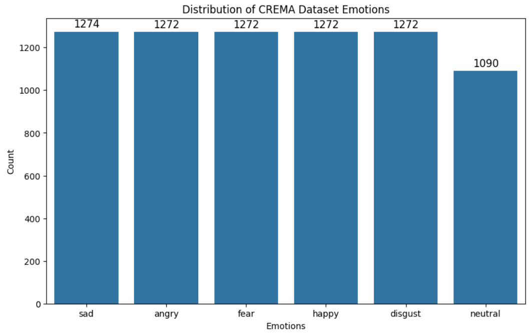


Figure 2: Count of emotions for Crema Dataset

speech segments. This function works on the principle that before detecting the onsets, it is imperative to determine the overall amplitude envelope of the signal, which represents the temporal fluctuations in the intensity level of the signal. The adaptive threshold is then conducted based on this envelope; segments having an amplitude less than a fixed threshold power of the signal are recognized as non-speech and excluded. It further makes the audio data set suitable for emotion recognition by handling the temporal dimension and shrinking the data size. Figure 3 (a) represents the original waveform and spectrogram of speech data, whereas Figure 3 (b) represents the change in the waveform and spectrogram of speech data after preprocessing. This has been performed for both the datasets used in this research.

3.4 Audio Feature Extraction and Data Augmentation

- **Audio Features**
  Audio feature selection and extraction is a crucial step in every speech emotion recognition research. Human emotions are communicated through various aspects of the speaker's voice quality such as pitch, articulation, intensity, and other vocal features. The appropriate selection of audio features can play a major role in the performance of emotion recognition systems. In this study, the following audio features are considered.

  - Mel-Frequency Cepstral Coefficients (MFCC): MFCCs are widely used in the field of speech recognition and were introduced by Davis and Mermelstein

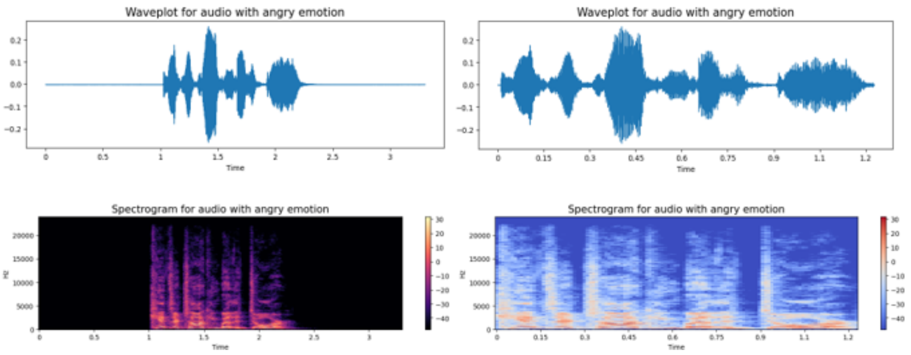


Figure 3: (a) Before Preprocessing (b) After Preprocessing

(1980). MFCCs capture the essential characteristics of speech signals by simulating the human ear's perception of sound frequencies. This feature is particularly effective for emotion classification, as it captures the unique aspects of speech that vary with emotional states. In this study, 30-dimensional MFCCs are extracted.

  - Root Mean Square Energy (RMSE): RMSE measures intensity or the power of the speech signal. This feature is particularly valuable for 'toning' of the emotional intensity of the speaker or the activity level. High energy can be associated with feelings such as anger or excitement.
  - Zero Crossing Rate (ZCR): This feature measures the amount of local variations and complexity in the signal this is rather proportional to the loudness of the signal or signal noise, ZCR values are high during anger that is normally accompanied by high pitch variations in speech and other noises.

- **Data Augmentation Techniques**
  Data augmentation techniques are used to enhance the robustness of the model and improve its generalization. The study by Jahangir et al. (2022) demonstrates the effectiveness of using data augmentation techniques in speech emotion recognition models. In data augmentation, new syntactic data is created by adding small changes to our initial training data. The below techniques are used to create syntactic data for this study.

  - Noise Injection: This technique involves an element of randomness that will enable the model to cope with the other background noise that appears when the model is in the real world. It is performed by training the model on the noisy data and it becomes immune to the noise and hence has a good performance on the noisy test data. An example of noise injection is depicted in Figure 4 (a).
  - Time Stretching: In order to reproduce variations in speaking rate time compression and expansion alter the rate of the signal yet retain the pitch. This technique is beneficial to the model in the sense that it aids in learning and generalizing where there are speakers who speak at different paces. This is demonstrated in Figure 4 (b).

- Time Shifting: Time shifting changes when an audio signal starts basically messing with the speech's beginning time. This method helps the model handle small misalignments in audio data better. There is an example of time shifting inFigure 4 (c).

- Pitch Shifting: Pitch shifting alters a signal's pitch and frequency without changing how long it is. This approach takes into account how voice pitch differs from person to person and changes with emotions. Figure 4 (d) shows you what data looks like after a pitch shift.

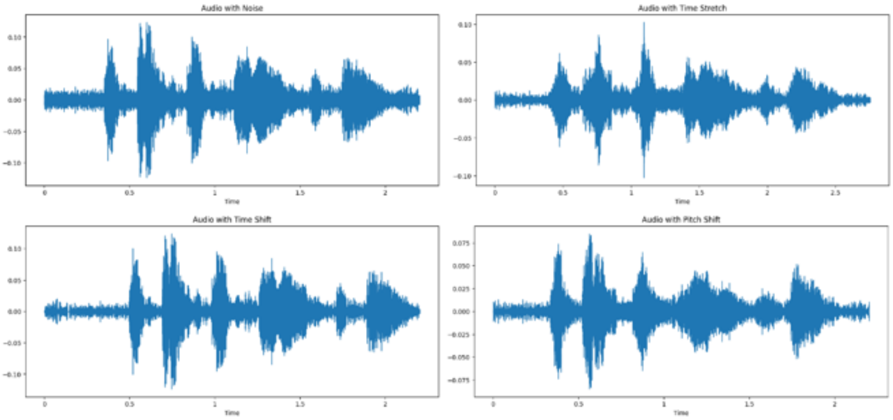


Figure 4: (a) Audio with Noise (b) Audio with Time Stretch (c) Audio with Time Shift (d) Audio with Pitch Shift

- **Combined Features**
  These features are then compiled together to create a feature set that includes a wide range of expressions and characteristics of emotions. This set of features consists of MFCCs, ZCR, RMSE, along with data augmentation techniques: noise injection, time stretching, time shifting, and pitch shifting. All these functions are performed with the help of the librosa library. (McFee et al. (2020) By aggregating all these various features, the model gains an improved understanding to classify different emotions reflected in the speech.

3.5 Model

To optimize the analysis of the speech emotion data, this research has developed a hybrid model which is a combination of CNN, Bidirectional LSTM, and Transformer. Every one of them has its own advantages and can be used in the process of improving the general performance of the model. The first is the Convolutional Neural Network(CNN) that performs convolution on the layers to some extent in order to extract features from the audio signal and hence perform the identification of local features and aspects within the audio signals. CNN is designed specifically for the capturing of the variations of tones and rhythms used to differentiate one emotional state from another.

Apart from CNN, there are Bidirectional Long Short-Term Memory (BiLSTM) networks incorporated in the architecture because of their capability to work with sequences and capture long dependencies in speech. BiLSTM layers serve as profound ways of optimizing the model with the effects that the model can capture the temporal relationships between the emotions captured over time. Moreover, Transformer modules are incorporated into layout to make use of the self-attention mechanisms when processing nested relations of intricate complexity within the audio data. The use of CNN, BiLSTM, and Transformer makes a robust model combination that utilizes each model's advantage and has the capacity to increase the results of detecting emotion from speech.

3.6 Model Evaluation Techniques

The evaluation techniques collectively provide a comprehensive understanding of the model's performance. Using the study by Saidani et al. (2023) as reference, this research has used following evaluation techniques to measure the performance of the hybrid model.

- Plotting Training and Validation Accuracy: The graph represents the training as well as the validation accuracy against the respective epochs. It utilizes the history object which usually is derived from the fit method of a Keras Model. This plot is useful for determining how well the model is learning over time as well as whether it is overfitting or underfitting.

- Confusion Matrix: Confusion matrix is used to give the idea about how effectively the model has classified the items. It shows the actual and predicted classes and can in turn be used as a method of identifying the classes that were given poor prediction accuracy.

- Metric Calculation: The main objective of this given function is to calculate and display the total amount of precision, total amount of recall, and total F1-score of the given model.

  - Precision: The proportion of accurately predicted actual positives to the number of all the positives that have been predicted.
    Precision = TP/(TP+FP).

  - Recall: The proportion of positive instances classified correctly to all the instances that actually belong to the class.
    Recall = TP/(TP+FN).

  - F1-Score: The weighted average of Precision and Recall.
    F1 Score = 2*(Recall * Precision) / (Recall + Precision).

  - Support: The number of true instances for each class.

4 Design Specification

A combination of BiLSTM-Transformer and a 1—D CNN model is explained in this study. By utilising their individual advantages, the 1-D CNN and BiLSTM-Transformer models improve the accuracy of emotion recognition by using the same audio characteristics as the first input.

- **One-Dimensional Convolution Neural Network:** The 1-D CNN used here is particularly suitable for dealing with raw audio features. It is designed to extract local features of the signal which are valuable for recognition of emotional states at short time intervals. The first branch of the architecture is 1-D CNN in this study, it has used five convolutional layers for learning local features where ReLU activation has been applied. The dropout layers are integrated after convolution layers to avoid overfitting the model. Finally, the flattening layer exists in the architecture in order to flatten the 3D output of convolutional layers to a 1D tensor.

- **BiLSTM-Transformer**: LSTM-based models are specifically developed for handling sequence data that are of variable length and hence can be effectively used for the analysis of time series data sets like audio data. Here, forward and backward information are processed equally due to the BiLSTM which is efficient in processing sequence data. The Transformer model that employs the attention mechanism, assigns different weights to various parts of the input data to highlight essential portions. Moreover, incorporating both BiLSTM and Transformer enables the model to model the temporal dependencies in the audio data as well as capture the complex hierarchical features in the audio data. In this research, the input layer takes the preprocessed audio feature input. Further, bidirectional LSTM captures sequential dependency of audio in both backward and forward direction. There are three transformer blocks with four attention heads used to capture complex sequential structure from the input.

4.1 Preprocessing for Model Input

Before feeding the data to the model, preprocessing of the model input is performed. This step ensures the input to be in the correct format and scaled appropriate to the model. In this research following steps are performed on the audio features before feeding them to the model.

- Data Preparation: This step is performed to separate feature set and labels from the data frame.

- Encoding Labels: The emotion labels are in text format which are encoded using one-hot encoding. One hot encoding coverts categorical emotion values into binary matrix representation.

- Splitting Data: This step is performed to split the data into training and testing dataset. This research has used 80-20 split to evaluate the performance of the model on the unseen data.

- Standardization: This step ensures the features have a mean of zero and standard deviation of one which helps neural network training process by increasing the convergence rate. This process helps to stabilise the optimisation, address the features of equal importance, eliminate cases of large numbers multiplication and generally improve the model performance.

4.2 Model Architecture

The combined deep neural network architecture which combines all the three models CNN and BiLST-Transformer is represented in the Figure 5. The 1-D CNN processes raw audio features to capture local patterns, while the BiLSTM–Transformer branch handles sequence data, leveraging temporal dependencies and complex hierarchical structures of audio. The outputs from these branches are integrated and passed through additional layers for final classification. The details of the additional layers are given below.

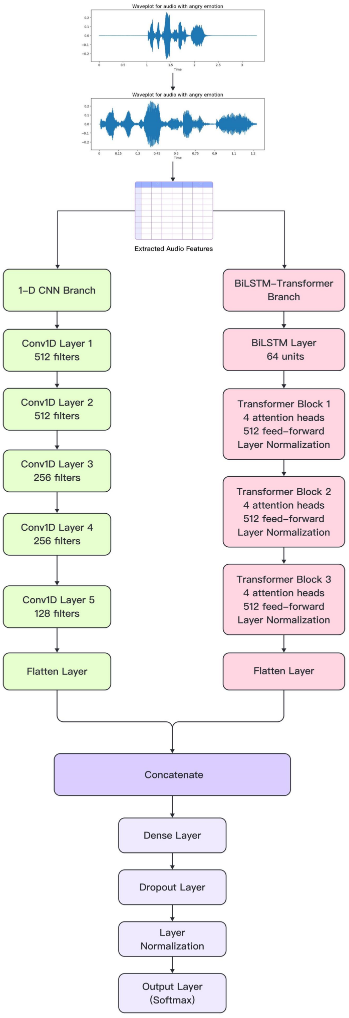


Figure 5: Integrated Deep Learning Architecture for Speech Emotion Recognition

- Concatenate: The flattened outputs of the 1-D CNN and BiLSTM-Transformer branches are combined. This step merges the rich feature representations learned from both branches, allowing the model to utilize diverse information from the audio data.

- Dense Layer: The concatenated output is passed through a dense layer with ReLU activation. This layer helps in learning complex representations from the combined features, enhancing the model's capability to distinguish between different emotions.

- Dropout Layer: A dropout layer is included immediately after the dense layer to reduce the chances of overfitting the model. This layer is useful in randomizing part of the training data to avoid overfitting by increasing generalization performance of the model to the unseen data.

- Layer Normalization: Layer normalization is used in order to normalize the activations within each layer of the neural network. It brings the majority of inputs of each layer to a normal distribution so that the model is trained much faster and converges earlier.

- Output Layer:At last, the data obtained are taken to an output layer for emotion classification after applying softmax activation function. This layer gives a probability distribution over all possible emotion classes to the model and makes an accurate prediction possible.

5 Implementation

The implementation of speech emotion recognition using a hybrid deep neural network combining CNN, BiLSTM, and Transformer models follows three stages:

First, the audio data from the two datasets, namely RAVDESS, and CRMA-D are obtained. The effects in the library are used in pre-processing of audio files to eliminate the non-speech portion of it. The trim function assists in the reduction of non-useful audio data which are not relevant. The characteristics of the speech as detected by MFCC, RMSE as well as ZCR, are underlined in this speech. To enhance the capability of the temporal model, and apply high-level features, data augmentation techniques were used to generate the feature data set, of signals with noise, stretched, shifted and pitch change. The data set is made into training and testing sets with equal ratio of 80:20. Features are then reshaped into a (number of samples, time step, 1) from which their mean and standard deviation are then removed. The emotion labels are encoded by One-Hot encoding which transforms nominal variables into a vector suitable for use in a classifier.

In the second stage, different architectures are incorporated into the hybrid model to optimize its performance. The CNN component begins with the input layer receiving audio features of dimensionality: (time steps, 1). In the proposed model, there is a single path of CNN with different convolutional layers, The first layer has 512 filters with the size of the kernel being 5 x 1. ReLU activation function, batch normalization, and max pooling are thereafter applied with a pool size of 5 and the size of the stride being 2. The same structure is continued, though the next layers decrease the filter count to 256 and 128 respectively, retaining the same basic structure and including the dropout layers rate 0.2 to avoid overfitting. In parallel, the BiLSTM layer with the 64 units finds temporal dependencies in the given audio data and generates its output which is expanded to feed a Transformer model. The transformer model used in this research has the following architecture blocks: self-attention with 4 headings, feed-forward network with 512 units, layer normalization with epsilon 10, and dropout with 0.1 rates. The outputs of CNN and BiLSTM-Transformer branches are concatenated and passed through a 64-unit dense layer, ReLU activation function, and a dropout layer with a dropout rate of 0.5. At last, a softmax activation function is used to predict the emotion classes.

The third and the last stage focuses on the analysis of the sustainability of the hybrid model. The results of the output model will be discussed using a training and testing accuracy graph, and confusion matrix of the identified subset of emotions with the overall evaluation metrics comprising of precision, recall, and F1 measure. Classification metric is also computed to evaluate the performance of data for all the emotion categories to understand which emotion categories performed well.

6 Evaluation

This section presents all the results for the proposed deep neural network with CNN and BiLSTM-Transformers models in the construction of the hybrid deep neural network architecture. To check the performance of the proposed model, this research has used RAVDESS and CREMA datasets.

6.1 Ravdess Dataset

- Training and Validation Accuracy: Figure 6 represents the training and validation accuracy plot, illustrating the model's learning process over 20 epochs. The training accuracy steadily increases, eventually reaching over 90%, indicating that the model is effectively learning the training data. The validation accuracy also improves significantly, peaking around 80%, which suggests good generalization to unseen data. Overall, the model performs well on the validation set.

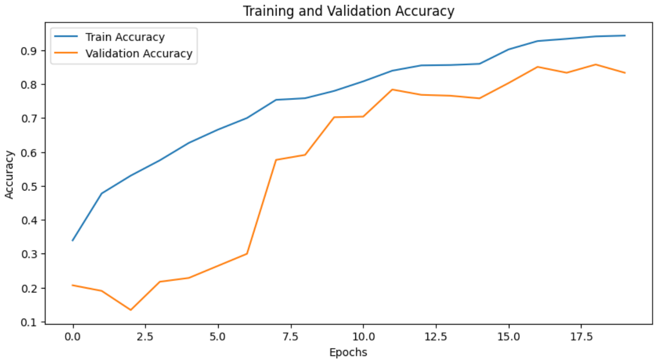


Figure 6: Training Vs Validation Accuracy Graph for Ravdess

- Confusion Matrix: Figure 7 represents a confusion matrix illustrating the model's performance in classifying emotions into seven categories: surprise, happy, angry, neutral, disgust, fear and sad. Each row represents the true emotion, while each column represents the predicted emotion. The numerical values within the matrix indicate the number of instances where a true emotion was classified as a particular predicted emotion. The diagonal values represent correct classifications. The model performs well in classifying disgust, surprise, neutral, sad, and angry, with high diagonal values and darker blue squares. Whereas, fear and happy emotion has the lowest overall performance, with lower diagonal values and more spread-out distribution across predicted emotions.

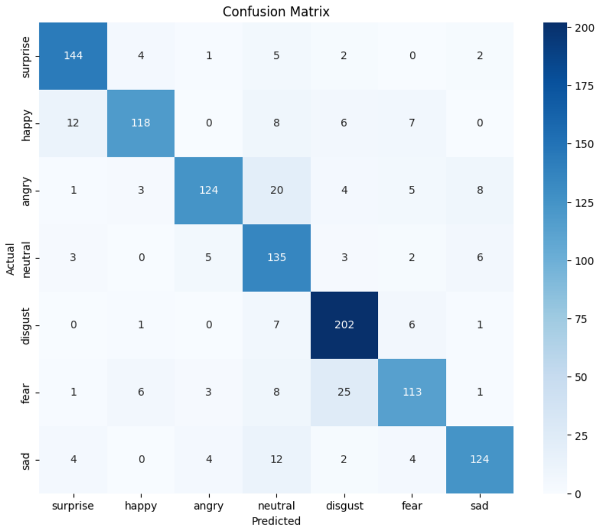
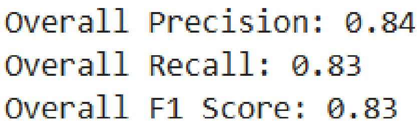


Figure 7: Confusion Matrix for Ravdess

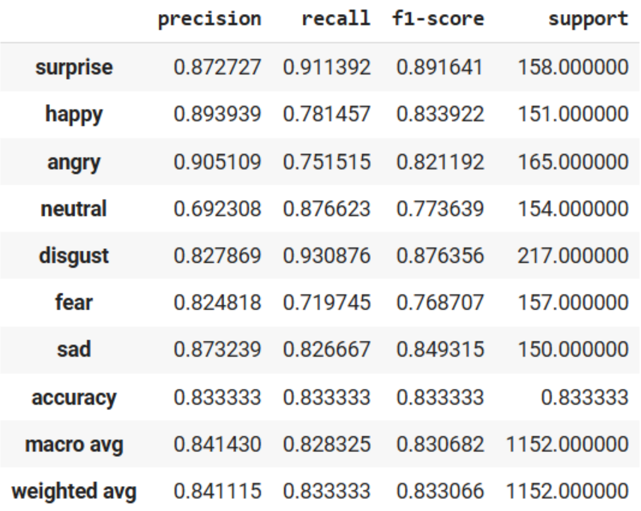
- Overall Metrics: The Figure 8 demonstrates a strong ability to accurately identify positive cases, evidenced by its precision of 0.84, meaning that 84% of the positive predictions made by the model were correct. This indicates the model is quite effective at identifying true positives. Additionally, the model has a recall of 0.83, showing that it correctly identified 83% of all actual positive cases, reflecting its good sensitivity in detecting positive instances. The F1-score of 0.83 balances precision and recall, suggesting that the model maintains a good equilibrium between correctly identifying positive cases and avoiding false positives. Overall, the high precision and recall confirm the model's robust performance, while the F1-score indicates a balanced and effective model.



```
Overall Precision: 0.84
Overall Recall: 0.83
Overall F1 Score: 0.83
```

Figure 8: Overall Metrics for Ravdess

- Classification Report with Metrics: Figure 9 presents the evaluation metrics for the emotion classification model. The model demonstrates strong overall performance, achieving an accuracy of 83.33%. It excels in recognizing surprise, with high precision (0.87) and recall (0.91), indicating accurate identification of positive cases. The model also shows proficiency in detecting happy, angry, disgust, and sad emotions, maintaining f1-score values above 0.80. However, there's room for improvement in classifying neutral and fear expressions, as indicated by lower precision and recall scores. Despite these areas, the model generally exhibits a balanced performance across different emotion categories.



	precision	recall	f1-score	support
surprise	0.872727	0.911392	0.891641	158.000000
happy	0.893939	0.781457	0.833922	151.000000
angry	0.905109	0.751515	0.821192	165.000000
neutral	0.692308	0.876623	0.773639	154.000000
disgust	0.827869	0.930876	0.876356	217.000000
fear	0.824818	0.719745	0.768707	157.000000
sad	0.873239	0.826667	0.849315	150.000000
accuracy	0.833333	0.833333	0.833333	0.833333
macro avg	0.841430	0.828325	0.830682	1152.000000
weighted avg	0.841115	0.833333	0.833066	1152.000000

Figure 9: Classification Report with Metrics

6.2 Crema-D Dataset

- Training and Validation Accuracy: Figure 10 represents the training and validation accuracy plot, illustrating the model's learning process over 50 epochs. The training accuracy steadily increases, eventually reaching over 80%, indicating that the model is effectively learning the training data. The curve is relatively smooth, suggesting a consistent learning process. Whereas, the validation accuracy also improves significantly, peaking around 80%, which shows a good generalization to unseen data.

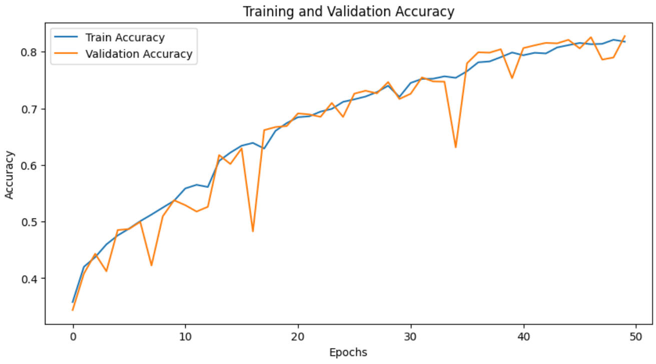


Figure 10: Training Vs Validation Accuracy

- Confusion Matrix: The Figure 11 illustrates the confusion matrix which presents a visual representation of the model's performance in classifying emotions into six categories: neutral, sad, angry, happy, fear, and disgust. Each row represents the true emotion, while each column represents the predicted emotion. The values within the matrix indicate the number of instances where a true emotion was classified as a particular predicted emotion. The diagonal values (from top-left to bottom-right) represent correct classifications. For example, 890 neutral instances were correctly classified as neutral. These values contribute to the overall accuracy of the model.

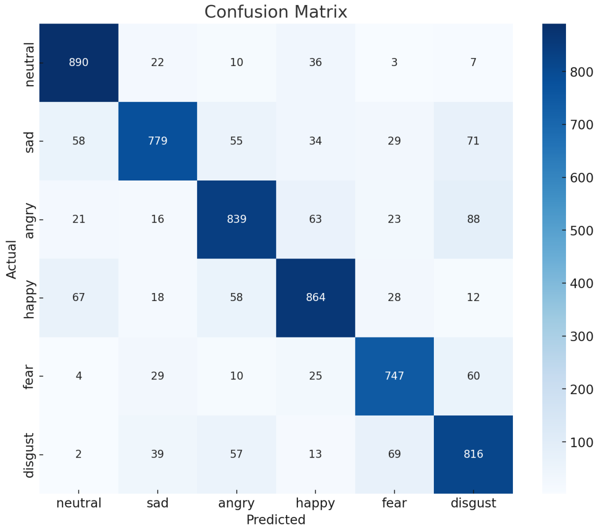
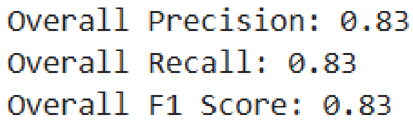


Figure 11: Confusion Matrix for Crema

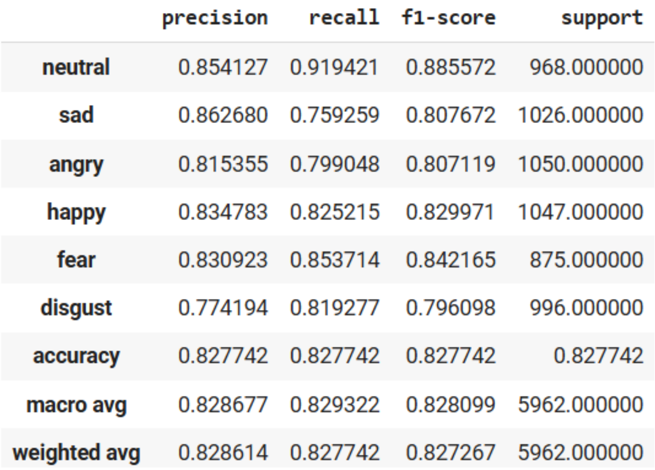
- Overall Metrics The Figure 12 demonstrates a strong ability to accurately identify positive cases, evidenced by its precision of 0.83, meaning that 83% of the positive predictions made by the model were correct. This indicates the model is quite effective at identifying true positives. Additionally, the model has a recall of 0.83, showing that it correctly identified 83% of all actual positive cases, reflecting its good sensitivity in detecting positive instances. The F1-score of 0.83 balances precision and recall, suggesting that the model maintains a good equilibrium between correctly identifying positive cases and avoiding false positives. Overall, the high precision and recall confirm the model's robust performance, while the F1-score indicates a balanced and effective model.



```
Overall Precision: 0.83
Overall Recall: 0.83
Overall F1 Score: 0.83
```

Figure 12: Overall Metrics

- Classification Report with Metrics The Figure 13 provides evaluation results and demonstrates that the model achieves a high level of accuracy (82.77%) in classifying emotions. The model exhibits consistent performance across multiple emotion classes, as evidenced by the balanced macro-average F1-score of 0.828099. The model identifies neutral and fear emotions with high precision and recall values. While demonstrating better performance in detecting neutral, fear, and happy emotions, there's slight room for improvement in detecting sad emotion instances. Additionally, the model's precision in recognizing disgust could be enhanced, despite achieving a decent recall rate.



	precision	recall	f1-score	support
neutral	0.854127	0.919421	0.885572	968.000000
sad	0.862680	0.759259	0.807672	1026.000000
angry	0.815355	0.799048	0.807119	1050.000000
happy	0.834783	0.825215	0.829971	1047.000000
fear	0.830923	0.853714	0.842165	875.000000
disgust	0.774194	0.819277	0.796098	996.000000
accuracy	0.827742	0.827742	0.827742	0.827742
macro avg	0.828677	0.829322	0.828099	5962.000000
weighted avg	0.828614	0.827742	0.827267	5962.000000

Figure 13: Classification Report with Metrics

6.3 Discussion

This study focused on the idea of combining 1D CNN and Bidirectional Long Short Term Memory - Transformer (BiLSTM-Transformer) together to use the strengths of both approaches to extract emotions from the audio speech data. The 1D CNN algorithm performs well in generating temporal information as it is able to extract features within limited time intervals. This allows 1D CNN to identify special features from speech data needed to recognize emotions. Speech on the other hand is a flow of utterance in a language, and in the context of emotion identification, one has to look for the sequence and order of the sounds that make up the language. This is where the BiLSTM–Transformer comes into play, which learns the representation of input sequences in an efficient way, preserves sequential dependencies, and generalizes well. BiLSTM processes the speech data in parallel from the initial beginning to the end and from the end of the data set to the initial beginning, thereby giving a 2-way approach. Subsequently, the transformer part of the model helps the model learn features of the speech selectively by providing the requisite weight to various parameters. By using the strengths of both BiLSTM and the transformer, the model ensures a complete understanding of the speech context. Overall, by combining 1D CNN with BiLSTM–Transformer, a model is obtained that can look at both specific patterns and the overall context in speech, leading to better emotion recognition.

When comparing the outcomes of our proposed model with the one in “A BiLSTM-Transformer and 2D CNN Architecture for Emotion Recognition” proposed by Kim and Lee (2023), the proposed speech emotion recognition model in this research performed well with Ravdess dataset with overall accuracy 83.3% in detecting all the 7 emotions as compared to the research by Kim and Lee (2023) which achieved accuracy of 80.19% for Ravdess dataset. One of the differences in approach was the use of 1D CNN model instead of 2D CNN. Another, proposed research used data augmentation techniques to make the model more robust and generalized with different audio data scenarios. Our research with 83.3% accuracy with Ravdess performed slightly better than the model in the study Ullah et al. (2023) (82.4%) which used CNN and multihead transformer.

Overall, the improvement of the accuracy proves that our proposed method is effective for the emotion recognition tasks, which means that the model's architecture, and selected data preprocessing methods for generating an audio feature set are suitable for handling the dataset. These results are also significant not only for this study's overall methodological framework but also provides a foundation for further enhancement of the emotion recognition systems, which will continue to evolve.

7 Conclusion and Future Work

In order to understand the speech emotion recognition system, this study proposed and assessed a hybrid deep neural network that combines transformers, bidirectional long short-term memory networks (BiLSTM), and 1D convolutional neural networks (CNN). The goal of this was to observe the growth of the speech-emotion detection engine's accuracy by means of utilising the power of each model. The model proposed in this work was able to record high accuracy rates in Ravdess as well as Crema datasets. Ravdess dataset was able to record an accuracy of 83.33% in detecting 7 emotion categories. Similarly, Crema-D dataset obtained 82.77% accuracy for 6 emotion categories. Additionally, our experimental results show that the new method outperformed previous models when tested on a similar dataset, RAVDESS. Specifically, our model achieved a higher accuracy rate compared to the 80.19% reported by Kim and Lee (2023) and the 82.4% reported by Ullah et al. (2023). The selected data preprocessing methods and audio feature extraction techniques proved suitable for handling the datasets, contributing to the model's overall effectiveness. The data augmentation techniques used for feature extraction from audio data made the model more robust and generalized. At last, this study successfully integrated CNN, BiLSTM, and Transformer models to classify emotions in speech with high efficiency and revealed good results with different datasets.

For future scope, this research can be extended with various modifications. One such modification can be change in dataflow among the models, either 1D CNN could feed the output to BiLSTM-Transformer model or vice versa. Another approach can be integrating the models in hierarchical order where lower-level models process initial categories of emotions and pass the results to higher-level models for further refinement which could improve the model's accuracy and efficiency. The ability to give 1D CNN and BiLSTM-Transformer models an opportunity to cross-validate their predictions with each other might add to their decision-making process and the objectiveness of the predictions as a whole. Managing computation resources is important, during the research the researcher switched from Jupyter Notebook to Google Colab due to computational constraints. Further, the testing of the model to incorporate more new test datasets for different speakers, accents, or expressive tones would also strengthen the generalization and reliability of the model suggesting its applicability in different scenarios.

References

Andayani, F., Theng, L. B., Tsun, M. T. and Chua, C. (2022). Hybrid lstm-transformer model for emotion recognition from speech audio files, *IEEE Access* **10**: 36018–36027.

Ayadi, S. and Lachiri, Z. (2023). Deep neural network architectures for audio emotion recognition performed on song and speech modalities, *International Journal of Speech Technology* **26**(4): 1165–1181.

Cao, H., Cooper, D., Keutmann, M., Gur, R., Nenkova, A. and Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset, *IEEE transactions on affective computing* **5**: 377–390.

Cowen, A., Sauter, D., Tracy, J. L. and Keltner, D. (2019). Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression, *Psychological Science in the Public Interest* **20**(1): 69–90.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE transactions on acoustics, speech, and signal processing* **28**(4): 357–366.

Issa, D., Demirci, M. F. and Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks, *Biomedical Signal Processing and Control* **59**: 101894.

Jahangir, R., Teh, Y. W., Mujtaba, G., Alroobaea, R., Shaikh, Z. H. and Ali, I. (2022). Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and features fusion, *Machine Vision and Applications* **33**(3): 41.

Jing, D., Manting, T. and Li, Z. (2021). Transformer-like model with linear attention for speech emotion recognition., *Journal of Southeast University (English Edition)* **37**(2).

Kim, S. and Lee, S.-P. (2023). A bilstm–transformer and 2d cnn architecture for emotion recognition from speech, *Electronics* **12**(19).
URL: *https://www.mdpi.com/2079-9292/12/19/4034*

Kumar, C., Maharana, A., Krishnan, S., Hanuma, S., Lal G, J. and Ravi, V. (2023). *Speech Emotion Recognition Using CNN-LSTM and Vision Transformer*, pp. 86–97.

Li, M., Zheng, Y., Li, D., Wu, Y., Wang, Y. and Fei, H. (2024). Ms-senet: Enhancing speech emotion recognition through multi-scale feature fusion with squeeze-and-excitation blocks, *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 12271–12275.

Livingstone, S. and Russo, F. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, *PLOS ONE* **13**: e0196391.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E. and Nieto, O. (2020). librosa/librosa: 0.8. 0 (version 0.8. 0), *Zenodo, Jul* **22**.

Mocanu, B. and Tapu, R. (2022). Emotion recognition from raw speech signals using 2d cnn with deep metric learning, *2022 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–5.

Mountzouris, K., Perikos, I. and Hatzilygeroudis, I. (2023). Speech emotion recognition using convolutional neural networks with attention mechanism, *Electronics* **12**(20).
URL: *https://www.mdpi.com/2079-9292/12/20/4376*

Parry, J., Palaz, D., Clarke, G., Lecomte, P., Mead, R., Berger, M. and Hofer, G. (2019). Analysis of deep learning architectures for cross-corpus speech emotion recognition., *Interspeech*, pp. 1656–1660.

Pulatov, I., Oteniyazov, R., Makhmudov, F. and Cho, Y.-I. (2023). Enhancing speech emotion recognition using dual feature extraction encoders, *Sensors* **23**(14).
URL: *https://www.mdpi.com/1424-8220/23/14/6640*

Saidani, O., Aljrees, T., Umer, M., Alturki, N., Alshardan, A., Khan, S. W., Alsubai, S. and Ashraf, I. (2023). Enhancing prediction of brain tumor classification using images and numerical data features, *Sensors and Diagnostics* **13**: 2544.

Senthilkumar, N., Karpakam, S., Devi, M. G., Balakumaresan, R. and Dhilipkumar, P. (2022). Speech emotion recognition based on bi-directional lstm architecture and deep belief networks, *Materials Today: Proceedings* **57**: 2180–2184.

Ullah, R., Asif, M., Shah, W. A., Anjam, F., Ullah, I., Khurshaid, T., Wuttisittikulkij, L., Shah, S., Ali, S. M. and Alibakhshikenari, M. (2023). Speech emotion recognition using convolution neural networks and multi-head convolutional transformer, *Sensors* **23**(13): 6212.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need, *Advances in neural information processing systems* **30**.