A Comprehensive Study on Enhancing Emotion Recognition in Speech Using Deep Neural Networks

Debmalya Deb
Data Mining and Machine Learning 2
National College of Ireland
x21242101@student.ncirl.ie

Giorgia Luzia Pscheidt
Data Mining and Machine Learning 2
National College of Ireland
x22184261@student.ncirl.ie

Saheli Dutta
Data Mining and Machine Learning 2
National College of Ireland
x21246513@student.ncirl.ie

***Abstract*—Understanding people's emotions in speech is a core aspect of automated phone system technology which allows us to comprehend the caller's current moment's behaviour and mood which is the core objective of this study. The main motive of this project is to collect a bunch of audio files(7442 in count) that represent a dataset and split them concerning their emotions, for instance, sad, happy, angry, fearful, disgusted, and neutral and then feed those emotions through a Convolution Neural Network(CNN) and Long Short Term Memory(LSTM) model after Mel-Frequency Cepstral Coefficient(MFCC) and raw feature extraction that can capture emotional state respectively from the raw audio(.wav format) files. To measure the performance of the models this paper breaks the dataset into validation and testing and compares it based on model accuracy and loss. This study employs Keras Tuner Optimiser to find the best parameters for both of the deep neural network models before the final model fit. While doing this analysis the authors check all the parameters of the two models and award the CNN model as the best-fit model for the Emotion in Speech Recognition study. Finally, this research concludes its outcome by including a confusion matrix and a classification report.**

***Index Terms*—Audio Analysis, Convolution Neural Network, Long Short Term Memory, Mel-Frequency Cepstral Coefficient, Keras Tuner.**

I. Introduction

The emotion expressed through speech provides worthwhile information for understanding the intention of the user, emotional state, and other humane aspects. Speech Emotion Recognition(SER) always plays a vital part to support Human-Computer Interaction (HCI) systems, allowing better natural and useful communication between humans and machines [1]. This is why SER become an integral element in various applications, especially in Interactive Voice Response(IVR) systems where the caller's present mood and emotions can be identified before proceeding into further conversation.

The motivation for this study derives from the need to manage the current challenges and investigate unexplored strategies for enhancing the implementation of SER systems. In recent years, deep-learning techniques such as convolutional neural network models(CNN) and recurrent neural network(RNN) models have been implemented in this particular domain where these techniques have shown advantageous developments. This study seeks to deliver a complete analysis of emotion recognition from various speeches utilizing CNN and LSTM models. Moreover, Datasets like RAVDESS and TESS are widely explored in many types of research to recognize emotion in speech but the CREMA-D dataset is still less investigated. Hence, in particular, this dataset is utilized for a speech-emotion recognition system.

The entire process consists of three phases, In the first phase, wave plot, and spectrogram are used to represent and understand the amplitude of an audio signal over time, and the presence of spectral frequencies respectively where wavelet analysis is done to capture frequency, temporal details of the whole audio signal. In the second step, feature extraction is done through MFCC for the CNN model to analyze pivotal spectral characteristics of audio, and for the LSTM model raw features from audio are extracted. Finally, with the help of the Keras tuner, optimal parameters are found and later these parameters are fed into both models.

- **Research Question** To what extent does Convolutional Neural Network(CNN) model with Mel Frequency Cepstral Coefficients feature extraction improve Speech Emotion Detection over Long Short Term Memory(LSTM) model?

The main objective is to experiment with the capabilities of the CNN model with MFCC and how this can work better than the heavier model like LSTM to improve emotion recognition in speech.

This paper is organized as follows. Review of related works where the background of speech emotion recognition using traditional and deep neural network models is discussed in Section II. In Section III, the method followed in this paper to understand emotion is highlighted with all important aspects. Section IV discusses the outcome and justifies the outcome with proper evaluation metrics, also helping to reach a conclusion about the better model. Limitations and future work will be discussed in the following section.

II. Related works

Detailed deep learning techniques [1] for SER just like DBM, RNN, DBN, CNN, and AE review have been provided with the two-phase feature extraction [2] and classification process using source-based excitation features, prosodic features, vocal traction factors and other hybrid features [3] with a nonlinear classifier(speech signal is no stationary) which include g Gaussian Mixture Model (GMM) and Hidden

Markov Model (HMM) [4]. The benefits are the comfort of model training and how the weights have been distributed and limitations observed in significant layer-wise architecture and the risk of over-learning when memorizing the layer data. Based on this paper, this study utilized feature extraction in LSTM and CNN, which also justifies the research question.

In this context one more study [5] has been conducted using MFCC extraction and the LSTM-Transformer model by evaluating The Ryerson AudioVisual Database of Emotional Speech and Song (RAVDESS) [6]. 75.33% weighted accuracy and 73.12% unweighted accuracy have been achieved which has outperformed the previous models according to the authors. Only a few limitations are being found just like adding more features while audio extraction such as prosody features and data augmentation to get a few more sample data for training the model. This research links the above paper by implementing the LSTM model, MFCC feature extraction, and data augmentation to get more audio features.

Feedforward Neural Network [7] using has been developed in which information moves only in the forward direction starting from the input node through the hidden layers up to the output node [8] where the accuracy was not satisfactory due to less training data. In the data prepossessing stage, they use the spectrogram waveform like this study. 93% test accuracy has come out. And to mitigate the over-fitting problem multiple hidden layers, dense layers, activation layers, flatten layers and dropout layers. According to the authors in the future scope of improvements, noise filtering could be used which is considered a limitation of this project.

One more interesting paper [9] directly connects this study by Implementing a 2D CNN-LSTM model using MFCC for multimodal music emotion detection which emphasizes the significance of audio and lyrics as essential components for organizing music based on its emotional content. Surrey Audio-Visual Expressed Emotion (SAVEE) [10], Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [11], Toronto Emotional Speech Set(TESS) and Crowd Sourced Emotional Multimodal Actors Dataset(CREMA-D) datasets have been used in this study. It is been found that the model outperforms significantly with female audio than male. To mitigate this problem this training model has been fed through with both male and female audio. As a future work data augmentation techniques can be implemented to get more samples for better training.

Another study utilises [12] the CNN model for automatic speech recognition(ASR) using open-sourced Kaggle datasets. Converting the speeches into texts and with the help of text transcription they train and test the data.

In this context, a paper [13] uses MFCC to classify the audio data and groups them before passing it through CNN hidden layers. Like this paper, this paper also breaks the audio into spectrogram and waveplot for visual analysis and measures the frequency and pitch. 78% test accuracy has been achieved and utterance-level characteristics have been obtained in this study. Data augmentation and noise filtering could have been implemented which would be the only limitation of this paper.

A time series waveform as input in an end-to-end CNN deep learning model to detect emotion in speech [14]. This paper demonstrated how deep CNN varies using raw waveforms to detect environmental sounds [15]. The model has been set up using Keras, Tensorflow, and deep learning neural networks API to provide high-level methods and lastly, Adam optimizer has been used with an initial learning rate of 0.0001. This study finds low-class accuracy for “happiness emotion”. Two datasets have been taken under consideration to perform this study while in future work the author asks to investigate some more datasets for better training.

A new vector feature which is a combination of speech characteristics, zero crossing rate, MFCC, pitch voice, and probability of voice has been proposed in a study [18] with various kinds of machine learning algorithms such as Support Vector Machine (SVM), Logistic Regression not only that deep learning classifiers and architectures like Convolutional Neural Network (CNN) are used in this analysis. In the feature extraction, the first 13 features are taken as the first derivatives in MFCC and the rest comes under the second derivative. It is been analyzed that the general feature vector outperformed the MFCC on speech emotion classification using the CNN model.

A fantastic research [19] paper has been found on Covid-19 detection for SER which first starts with extraction of cepstral, spectral, and periodicity features for high dimensional audio to efficiently detect Covid-19 and asthma using gradient boosting machine learning algorithm. The performance of the audio quality sometimes depends upon the background noise level [20] which affects the audio quality during extraction that’s why the study analyses the change in noise level before and after data preprocessing using low pass filters. In this comparative study among LightGBM, SVM, Random Forest and KNN the LightGBM classifier wins with respect to the performance. The improvement of the preprocessing is the major contribution on this paper although to improve it further as in the future work combination of audio can be taken.

III. Speech Emotion Recognition Framework

This study utilizes the Cross-industry standard process for data mining(CRISP-DM) as a structured framework where MFCC feature extraction is done in the data pre-processing phase to understand the data which helps to confirm reliable inputs. Model creation, parameter tuning, model validation, and model testing reflect the model building, evaluation, and deployment stage which help to improve the accuracy and practical usefulness of sentiment recognition from speech.

The entire process can be divided into three parts.

- Firstly, audio files are collected from Kaggle, after this visual presentation of speeches, feature extraction is conducted.
- In the second phase, data is divided into training, testing, and validation set. Keras tuner is utilized to get optimum parameters for model building.

- Finally, after building a model with optimum parameters, models are tested using a classification report and confusion matrix, loss, and accuracy.

A. Details of Dataset

In this study, Crowd Sourced Emotional Multimodal Actors Dataset(CREMA-D)dataset is used from Kaggle[1]. This dataset contains 7442 original audio clips from 91 professional actors(48 male and 43 female), actors are from a variety of races(Asian, Caucasian, African American, and Hispanic) between the ages of 20 and 74. Fig. 1 shows the count of emotions in the CREMA-D dataset.

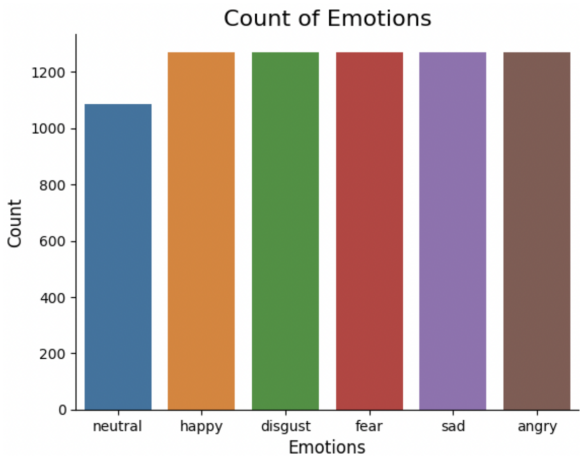

Fig. 1. Emotions Present in CREMA-D Dataset

There are six target classes and they are anger, disgust, happy, sad, fear, and neutral in Fig. 2.

	Emotions	Path
0	neutral	/content/drive/MyDrive/Emotion in Speech/Audio...
1	happy	/content/drive/MyDrive/Emotion in Speech/Audio...
2	disgust	/content/drive/MyDrive/Emotion in Speech/Audio...
3	fear	/content/drive/MyDrive/Emotion in Speech/Audio...
4	disgust	/content/drive/MyDrive/Emotion in Speech/Audio...

Fig. 2. Dataframe of CREMA-D Dataset

B. Data Preprocessing

Spectrogram and Waveplot are used to understand the frequency spectrum and amplitude of audio files over time. Especially, using morlet wavelet transformation, the spectrogram is employed to understand the audio file's frequency change over time, with different scales capturing several frequency ranges. Although, Fourier transform is very commonly used in audio signals but doesn't provide good resolution while simultaneously allowing both time and frequency domain [16]. Also, Fourier transform concentrates on the individual frames of audio files while the wavelet analysis takes into account the entire audio signal to simultaneously analyze the time and frequency domain with good resolution.
Fig. 3 displays waveplot of angry emotion where the x-axis represents time and the y-axis represents amplitude.

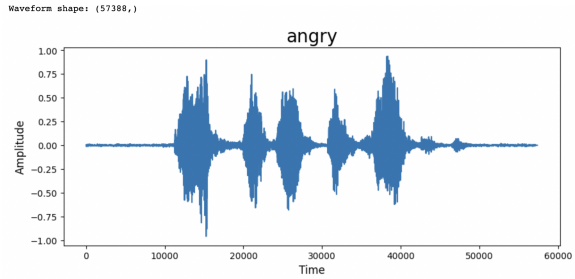

Fig. 3. Waveplot of Angry Emotion

Fig. 4 displays a a wavelet spectrogram of angry emotion where the the x-axis represents time and the the y-axis represents frequency.

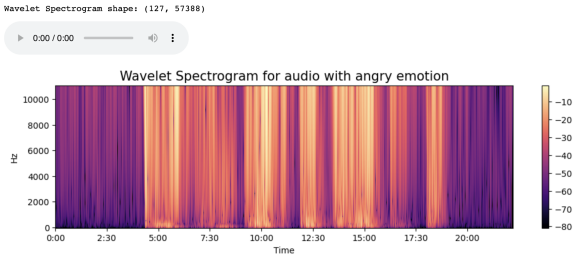

Fig. 4. Wavelet Spectrogram of Angry Emotion

Data Augmentation such as stretching, shifting, pitching, and amplitude scaling is done to generalize the dataset, but it leads to overfitting the model due to excessive augmentation which further led to some patterns in the training dataset. Hence, data augmentation is discontinued in this study.

C. Feature Extraction

To make the data more comprehensible, MFCC feature extraction for the CNN model is employed. MFCC is commonly used because of its ability to imitate the human auditory system's sensitivity to the distinct frequency spectrum. This particular technique uses triangular filters which sit on the mel frequency scale to identify suitable spectral features, further converting the energy of these filters into de-correlated coefficients. Hence, this method is very efficacious because it syncs with human perception which helps to understand emotion in speech recognition tasks [17]. Fig. 5 shows MFCC feature extraction is done using forty coefficients.
However, for the LSTM model, MFCC feature extraction is not utilized because the LSTM model needs data in sequence for input. At the same time, the MFCC method considers full-length audio without considering sequencing format. Hence, the raw audio signal is used to pre-process the data. Moreover, after extracting the features from audio signals, Padding is done to to guarantee that all audio files have the exact length for compatibility because the LSTM demand fixed-size inputs.

[1]https://www.kaggle.com/datasets/ejlok1/cremad

```
array([-2.8555411e+02,  1.2049206e+02, -1.6643663e+01,  4.1877354e+01,
       -6.4962044e+00,  2.1643887e+00, -2.8035318e+01,  5.2555642e+00,
       -7.7527280e+00,  3.4441559e+00, -4.4781303e+00, -6.7129841e+00,
        9.2220306e+00, -7.2924409e+00,  3.2732396e+00, -8.3401680e+00,
       -3.5440199e+00, -6.4309120e+00, -1.2588555e+00, -6.9931746e+00,
       -3.7238328e+00, -6.8922567e-01, -6.5523257e+00, -6.6113341e-01,
       -1.6452968e+00,  4.3221679e+00, -4.1887197e-01,  2.1999202e+00,
       -4.0927143e+00,  1.1927364e+00, -1.7052670e+00, -4.5846705e+00,
       -5.2702701e-01, -6.2238688e+00, -7.4646676e-01, -2.8239086e+00,
        7.9684444e-02, -1.5343150e-02,  3.9648964e+00,  4.1059055e+00],
      dtype=float32)
```

Fig. 5. Feature Extraction of Audio using MFCC

D. Data Encoding

This SER study consists of 6 different emotions and to convert those emotions into numerical representation this study utilizes one hot encoding method to the ”Emotions” section in the data frame because any machine learning models can not comprehend the emotions unless they are numerically represented. This encoding technique does not prefer any specific emotions, it just helps to represent the categorical labels.

E. Data Splitting

In the data splitting phase, the study breaks the complete dataset into training, validation, and testing after preprocessing the data. Firstly, 75% of the entire dataset takes the training category and the rest has been split into validation(25% of testing data) and test(75% of testing data) sets.

After splitting the dataset, the input data has been scaled using the standard scaler technique because the prepossessed data is having a mixed combination of negative and positive values. And specifically, the standard scaling has been used to maintain a mean of 0 and a standard deviation of 1.

Lastly, the dimension of the input data has been expanded to 3 channels (RGB) for model compatibility.

F. Model Building

This study analyses the performance of the speech emotion recognition(SER) model by employing a CNN and LSTM model.

1) ***CNN Framework****:* A sequential CNN model has been developed with having Leaky Rectified Linear Unit(Leaky ReLu) activation function in its hidden layers. Initially, this research has been started with the Rectified Linear Unit(Relu) activation function but down the line, the researchers faced the dead neuron problem along with that they were unable to handle the negative value and hence the model was unable to provide efficient accuracy. The padding technique has been used in the deep layers so that the actual information does not get lost along with that 5 filters with stride 1 are added to catch distinct features in the input data while preserving resolution. Moreover, the Maxpooling technique has been added in the sequential neural network's layer so that the downsized audio file will be having significant features by designating the highest value within each pooling window with less computation time to recognize the exact emotion. The dropout regularization method is also used by deactivating some random neurons in the training phase to prevent overfitting problems and faster convergence. Fig. 6 shows structure of CNN model.

Layer (type)	Output Shape	Param #
conv1d	(None, 40, 160)	960
leaky_re_lu	(None, 40, 160)	0
max_pooling1d	(None, 20, 160)	0
dropout	(None, 20, 160)	0
conv1d_1	(None, 20, 448)	3,58,848
leaky_re_lu_1	(None, 20, 448)	0
max_pooling1d_1	(None, 10, 448)	0
dropout_1	(None, 10, 448)	0
conv1d_2	(None, 10, 256)	5,73,696
leaky_re_lu_2	(None, 10, 256)	0
max_pooling1d_2	(None, 5, 256)	0
dropout_2	(None, 5, 256)	0
conv1d_3	(None, 5, 480)	6,14,880
leaky_re_lu_3	(None, 5, 480)	0
max_pooling1d_3	(None, 3, 480)	0
dropout_3	(None, 3, 480)	0
flatten	(None, 1440)	0
dense	(None, 64)	92,224
leaky_re_lu_4	(None, 64)	0
dropout_4	(None, 64)	0
dense_1	**(None, 6)**	**390**
Total params:		16,40,998
Trainable params:		16,40,998
Non-trainable params:		0

Fig. 6. Diagram of Convolutional Neural Network

The last layer has been flattened just to make it fully connected with each and every neuron and to convert the 2-dimensional feature map into a 1-dimensional vector. Finally, to recognize 6 different emotions softmax activation function is utilized which helps to provide a useful probabilistic approach. Adaptive Moment Estimation(Adam) optimizer, which combines the benefits of Adaptive Gradient Algorithm(Adagrad) and Root Mean Squared Propagation(RMSprop), has been employed in this layer because it takes the best value learning rate while training to handle the noisy and sparse gradients. And categorical class entropy has been used under the loss function to handle multi-class classification.

Keras Tuner has been deployed for the above CNN model before the actual model building to get the optimum parameters that could lead to maximum efficiency. In this Keras tuner model, the study has performed 10 trials with 50 epochs by setting the callbacks such as early stopping and reduce lr

on plateau(rlrp). Rlrp adjusts the learning rate based on the improvement of the metrics on each epoch, it also helps to avoid the local minima with that the minimum learning rate has been set as 0.00000001. And early stopping helps to stop further training when it senses that further improvement does not happen or if it is falling down it also trains the model at the optimum point to diminish the over-fitting.

After the Keras tuner hyperparameter tuning, the model is fed through the tuner's searched best parameters with 1000 epochs. And again in the callback rlrp and early stopping have been used for better computation and a faster approach.

*2) **LSTM Framework**:* A sequential LSTM model has been developed with multiple dense layers. A Leaky ReLU activation function is employed in its dense layers with alpha 0.1. In this model building also dropout regularization technique has been implemented just like the above CNN model along with that in the compilation stage this model uses categorical_crossentropy as a loss function, Adam as an optimizer and softmax as an activation function.

Like the above CNN model here also Keras tuner has been deployed for hyperparameter tuning using 10 maximum training trials with 50 epochs. And same callback functions are also called while tuning. Fig. 7 shows structure of LSTM model.

Layer (type)	Output Shape	Param #
lstm	(None, 128)	66,560
dropout	(None, 128)	0
dense	(None, 160)	20,640
leaky_re_lu	(None, 160)	0
dropout_1	(None, 160)	0
dense_1	(None, 80)	12,880
leaky_re_lu_1	(None, 80)	0
dropout_2	(None, 80)	0
dense_2	(None, 48)	3,888
leaky_re_lu_2	(None, 48)	0
dropout_3	(None, 48)	0
dense_3	(None, 6)	294
Total params:		1,04,262
Trainable Params		1,04,262
Non-trainable params		0

Fig. 7. Diagram of Long Short Term Memory Network

IV. Result and Analysis

To evaluate the performance of the MFCC feature extracted CNN model and the raw feature extracted LSTM, this study focuses on accuracy and loss for both validation and test dataset to understand how models perform to recognize emotion in speech.

A. CNN Model Evaluation

- In the validation dataset using keras tuner hyperparameter tuning with 10 trials and 50 epochs, as per Fig. 8, this study gets the best value validation accuracy of 48%. The optimised parameters from this max trial will feed into the CNN model for the ultimate evaluation of speech recognition.

```
Trial 10 Complete [00h 01m 07s]
val_accuracy: 0.47526881098747253

Best val_accuracy So Far: 0.47526881098747253
Total elapsed time: 00h 09m 45s
```

Fig. 8. Validation Accuracy from Keras Tuner

- Now using the tuned parameters the CNN model gets trained with the validation dataset with 1000 epochs. The best value validation loss and accuracy are 1.40, and 47% respectively. As early stopping has been utilised while training that's why it stops the training at 15 epochs explained in Fig. 9. Noticeably the parameters referred from the Keras tuner which has been used to train this model lead to a similar kind of accuracy.

```
15/15 [==============================] - 0s 5ms/step - loss: 1.4013 - accuracy: 0.4688
Validation Loss: 1.401315689086914
Validation Accuracy: 0.46881720423698425
```

Fig. 9. Validation Accuracy and Loss

- The below visualisation, represents that the model is slightly over-fitting. From the accuracy graph, Fig. 10 shows training accuracy starts from around 52% which climbs up to 62% at 15 epochs out of 1000. Where the validation accuracy begins from 45% and reaches 47% on the 5th epoch but after that because of no improvement it stops at 15 epochs.

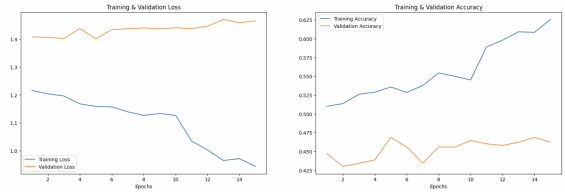

Fig. 10. Perfromance Graph of Training and Validation Datasets

- For the testing dataset, utilising Keras tuner hyperparameter tuning with 10 trials and 50 epochs, as per Fig. 11, this study brings the most suitable testing accuracy of 47%.

```
Trial 10 Complete [00h 01m 09s]
val_accuracy: 0.4563037157058716

Best val_accuracy So Far: 0.467048704624176
Total elapsed time: 00h 10m 51s
```

Fig. 11. Testing Accuracy from Keras Tuner

- The most satisfactory test loss and accuracy are observed at 1.36, and 46% respectively within 11 epochs out of 1000 shown in Fig. 12.

```
44/44 [==============================] - 0s 4ms/step - loss: 1.3671 - accuracy: 0.4556
Test Loss: 1.3671362400054932
Test Accuracy: 0.45558738708496094
```

Fig. 12. Testing Accuracy and Loss

- The visualisation illustrates in Fig. 13 shows that the model is a little over-fitting here as well, similar to the validation dataset. The accuracy and loss graph present that the training accuracy starts from around 48% which rises up to 62% at 11 epochs out of 1000. Whereas the testing accuracy starts from 46% and comes to 47% on the 4th epoch but after that, because of no progress it stops at the 11th epoch. The training loss starts from 1.3 and climbs down to 1.0 till the 11th epoch whereas the testing loss begins from 1.38 and reaches 1.4.

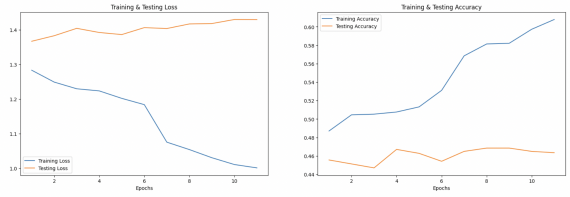

Fig. 13. Testing Accuracy and Loss

- Fig. 14 shows some comparison examples of the predicted and actual emotion for the CNN model. It is noted that for some emotions the model predicts correctly but for fear and neutral emotions, the model could not explicitly understand the emotions.

	Predicted Labels	Actual Labels
0	neutral	neutral
1	sad	fear
2	sad	neutral
3	sad	neutral
4	neutral	neutral
5	angry	angry
6	disgust	disgust
7	happy	disgust
8	sad	sad
9	sad	fear

Fig. 14. Predicted and Actual Labels

- The confusion matrix and the classification report present in Fig. 15, Fig. 16, which help to describe the overall performance of the CNN model to understand different classes. Such as for the angry class the model correctly predicts 64% instances under precision. For the true positive rate, the model predicts 41% neutral emotions correctly. And overall the model predicts 46% accuracy.

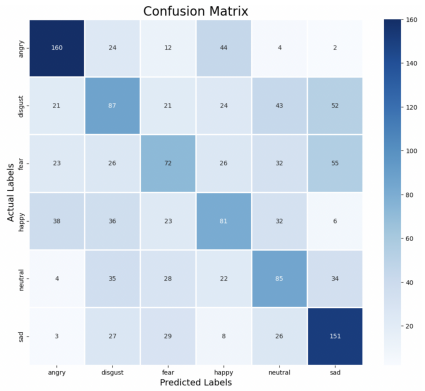

Fig. 15. Confusion Matrix

	precision	recall	f1-score	support
angry	0.64	0.65	0.65	246
disgust	0.37	0.35	0.36	248
fear	0.39	0.31	0.34	234
happy	0.40	0.38	0.38	216
neutral	0.38	0.41	0.40	208
sad	0.50	0.62	0.56	244
accuracy			0.46	1396
macro avg	0.45	0.45	0.45	1396
weighted avg	0.45	0.46	0.45	1396

Fig. 16. Classification Report

B. LSTM Model Evaluation

- In the validation dataset using keras tuner hyperparameter tuning with 10 trials and 50 epochs, according to Fig. 17, this analysis obtains the most satisfactory validation accuracy of around 20%.

```
Trial 10 Complete [00h 00m 23s]
val_accuracy: 0.18924731016159058

Best val_accuracy So Far: 0.20430107414722443
Total elapsed time: 00h 05m 18s
```

Fig. 17. Validation Accuracy from Keras Tuner for LSTM

- Now employing the best hyperparameters the LSTM model gets trained with the validation dataset with 1000 epochs. The most suitable validation loss and accuracy are 1.79, and 19% respectively have been achieved till the 29th epoch explained in Fig. 18.
- The graphical representation, shown in Fig. 19, illustrates that the model is under-fitting. The training accuracy starts from around 17% which rises up to 19% till the 29th epoch out of 1000. Where the validation accuracy

```
15/15 [==============================] - 0s 5ms/step - loss: 1.7913 - accuracy: 0.1871
Val Loss: 1.791272759437561
Val Accuracy: 0.18709677457809448
```

Fig. 18. Validation Accuracy and Loss for LSTM Model

begins from 17% and max up to 20% on the 19th epoch but after that because of no improvement it stops at the 29th epoch.

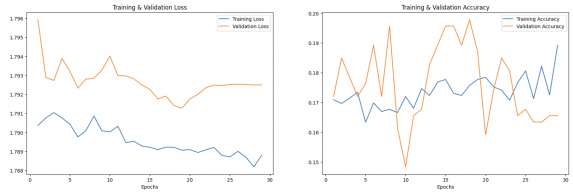

Fig. 19. Performance Graph of Training and Validation Dataset

- The LSTM model's most satisfactory test loss and accuracy are marked at 1.79, and around 18% respectively which is kind of similar to the result found in the validation set.
- Fig. 20 indicates a few comparison samples of the predicted and actual emotion for the LSTM model. It is recorded that for maximum emotions the model predicts incorrectly.

	Predicted Labels	Actual Labels
0	sad	neutral
1	happy	fear
2	fear	neutral
3	fear	neutral
4	happy	neutral
5	sad	angry
6	sad	disgust
7	disgust	disgust
8	sad	sad
9	fear	fear

Fig. 20. Predicted and Actual Labels

- The confusion matrix and the classification report present in Fig. 21, Fig. 22 represent the overall interpretation of the LSTM model to comprehend different classes. For example, in the disgust class, the model correctly predicts 31% instances under precision. And for the recall, the model predicts 66% sad emotions correctly. And overall the model predicts 18% accuracy.

	precision	recall	f1-score	support
angry	0.17	0.09	0.12	246
disgust	0.31	0.02	0.04	248
fear	0.19	0.14	0.16	234
happy	0.18	0.15	0.17	216
neutral	0.00	0.00	0.00	208
sad	0.18	0.66	0.28	244
accuracy			0.18	1396
macro avg	0.17	0.18	0.13	1396
weighted avg	0.18	0.18	0.13	1396

Fig. 21. Classification Report

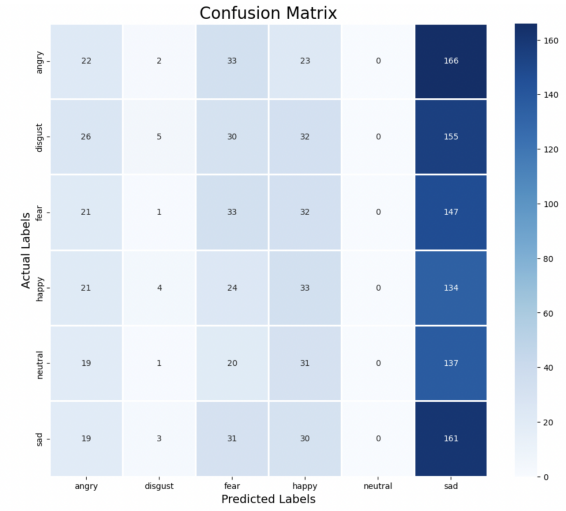

Fig. 22. Confusion Matrix

C. Conclusion

This research is predominantly based on MFCC feature extraction and the CNN model, where the CNN model performs well compared to the LSTM model which justifies the goal of this study as well. CNN model is able to identify emotions such as anger, sadness, and happiness with high precision whereas the LSTM model fails to identify most of the emotions except Sad. Although the CNN model works well, still further improvements are needed to achieve higher accuracy which will lead to identifying each and every emotion correctly. In the future, other feature extraction methods like zero crossing rate and croma can be implemented and more hyperparameter tuning can be done to achieve better results in the speech emotion recognition systems.

Acknowledgment

We would like to express our gratitude towards our professor Noel Cosgrave for his extreme contribution to helping this research in each stage.

References

[1] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A review," IEEE Access, vol. 7, pp. 117327–117345, Jan. 2019, doi: 10.1109/access.2019.2936124.

[2] 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: 10.1007/s10772-011-9125-1.

[3] M. E. 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: 10.1016/j.patcog.2010.09.020.

[4] A. D. Dileep and C. C. Sekhar, “GMM-Based Intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines,” IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 8, pp. 1421–1432, Aug. 2014, doi: 10.1109/tnnls.2013.2293512.

[5] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, “Recognition of Emotion in Speech-related Audio Files with LSTM-Transformer,” IEEE Conference, Mar. 2022, doi: 10.1109/icci54321.2022.9756100.

[6] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” PLoS ONE, vol. 13, no. 5: e0196391, pp. 1-35, 2018, doi: 10.1371/journal.pone.0196391

[7] K. Alam, N. Nigar, H. Erler, and A. Banerjee, “Speech Emotion Recognition from Audio Files Using Feedforward Neural Network,” IEEE, Feb. 2023, doi: 10.1109/ecce57851.2023.10101492.

[8] X. Wu et al., ”Speech Emotion Recognition Using Capsule Networks,” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6695-6699, doi: 10.1109/ICASSP.2019.8683163

[9] Priyanka Prashant Chimthankar.,(2021). Speech Emotion Recognition using Deep Learning ”norma.ncirl.ie/5142/1/priyankaprashantchimthankar.pdf”

[10] “Surrey Audio-Visual Expressed Emotion (SAVEE) database.” http://kahlan.eps.surrey.ac.uk/savee/

[11] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” PLOS ONE, vol. 13, no. 5, p. e0196391, May 2018, doi: 10.1371/journal.pone.0196391.

[12] K. S. Chintalapudi, I. A. K. Patan, H. V. Sontineni, V. S. K. Muvvala, S. V. Gangashetty and A. K. Dubey, ”Speech Emotion Recognition Using Deep Learning,” 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-5, doi: 10.1109/ICCCI56745.2023.10128612.

[13] S. Suganya and E. Y. A. Charles, ”Speech Emotion Recognition Using Deep Learning on audio recordings,” 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2019, pp. 1-6, doi: 10.1109/ICTer48817.2019.9023737.

[14] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, “Very deep convolutional neural networks for raw waveforms,” IEEE, Mar. 2017, doi: 10.1109/icassp.2017.7952190.

[15] A. Arul Edwin Raj, K. K. B, S. S and R. A, ”Speech Emotion Recognition using Deep Learning,” 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 505-509, doi: 10.1109/ICIDCA56705.2023.10100056.

[16] A. S. Dutt and P. D. Gader, “WAVELET Multiresolution analysis based Speech Emotion Recognition System using 1D CNN LSTM networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2043–2054, Jan. 2023, doi: 10.1109/taslp.2023.3277291.

[17] M. Hossan, S. Memon, and M. A. Gregory, “A novel approach for MFCC feature extraction,” Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference On, Dec. 2010, doi: 10.1109/icspcs.2010.5709752.

[18] A. Shah and T. Bhowmik, ”Speech Emotion Recognition using a Novel Feature Vector based on Voiced Probability and Speech Characteristics,” 2022 IEEE 6th Conference on Information and Communication Technology (CICT), Gwalior, India, 2022, pp. 1-5, doi: 10.1109/CICT56698.2022.9997929.

[19] T. K. Dash, C. Chakraborty, S. Mahapatra and G. Panda, ”Gradient Boosting Machine and Efficient Combination of Features for Speech-Based Detection of COVID-19,” in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 11, pp. 5364-5371, Nov. 2022, doi:10.1109/JBHI.2022.3197910.

[20] C. G. L. Prell and O. H. Clavier, “Effects of noise on speech recognition: Challenges for communication by service members,” Hear. Res., vol. 349, pp. 76–89, 2017.