Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 3 (2024)

Recognising Speech Based Emotion Using CNN in Deep Learning

[1]Dr.PVSSrinivas , [2]Arepalli Akshaya, [3]Bekkam Praneetha Singh ,[4]Dhandu Mounika

Department of CSE Vignana Bharathi Institute of Technology, Ghatkesar

***Abstract*:-** Building a powerful machine learning model for emotion recognition in speech is the project's primary objective. To do this, the model must be trained using a diverse dataset that includes a spectrum of emotions, including neutrality, sadness, anger, disgust, fear and happiness. With the use of cutting-edge methods like Convolutional Neural Networks and Deep Learning techniques, the model attempts to accurately anticipate the speaker's emotional state by analysing complex audio information. The strategy depends on utilising cutting-edge techniques, particularly Convolutional Neural Networks and Deep Learning. This strategic choice enables the model to grasp complex patterns in audio features, offering a nuanced understanding of the emotional content within speech. By adopting these advances techniques, the model aims to surpass traditional methods, enhancing its ability to recognize and classify emotions effectively. In initial stage, the model undergoes comprehensive training on a meticulously curated dataset. The model is trained using the tagged speech samples in this dataset, which span a variety of emotional states. The meticulous planning stage of the project is among its most crucial elements. Here, important features are extracted from the raw audio data by carefully processing it. This stage involves tasks such as audio segmentation, noise reduction, and feature extraction, ensuring that the model receives well-refined inputs. The subsequent stage involves the application of the trained model to real-world scenarios. Once equipped with the ability to recognize emotions in speech, the model can be deployed in practical setting, aiding professionals in psychology and speech therapy. In conclusion, the project presents a cutting-edge solution for emotion recognition in speech, combining advanced Machine Learning techniques with a meticulously curated dataset. The model's ability to accurately predict emotional state offers significant utility in psychology and speech therapy, providing professionals with a valuable tool for enhancing their understanding of emotional nuances.

Using the TESS dataset, our method creates a CNN model for audio classification. The code's objective is to categorise audio samples into six distinct emotional states, including fear, anger, disgust, happiness, neutrality, and sadness. As part of the process, MFCCs are retrieved as features from the audio data, and the dataset is then supplemented with several modifications. An output layer with softmax activation, a dense layer, and a convolutional layer make up the CNN model architecture. Using the TESS dataset for training and assessment, the model achieves a test accuracy of 93.33%. The main conclusions of this research shows that emotions may be accurately classified from audio samples using a CNN model. The model performs better when MFCC features and data augmentation methods like noise addition, temporal stretching, and pitch shifting are used. The suggested method for audio emotion classification is effective, as seen by the 93.99% accuracy that was attained. Applications include affective computing, human-computer interaction, and speech analysis that recognise emotions may be impacted by these findings.

Keywords:- SER, CNN, Mel-Frequency Cepstral Coefficient, HCI.

1.Introduction

The field of emotion recognition in speech falls under artificial intelligence and signal processing. Its objective is to use vocal signals to automatically detect and categorise the sentiment of a speaker. Speech can transmit a vast array of emotions, including happiness and sadness as well as fear, fury, disgust and neutrality. Speech emotion recognition is useful in many domains, such as virtual assistants, human-computer interaction, mental health monitoring, entertainment, and customer service. Voice emotion identification is one of the hardest problems in speech signal analysis. It takes into consideration as a research area problem that often tries to theorize the emotion from a speech signal. Challenges in speech emotion recognition include variability in emotional expression across speakers, cultural differences in emotional expression, context-dependent emotions, and the presence of noise in the speech signals. Scholars persistently investigate inventive methodologies to tackle these obstacles and enhance the precision and resilience of SER systems. In the end, speech emotion recognition technology is used in a wide range of industries, including customer service, recommended systems, and the medical industry. [3]. People express their sentiments, impressions, opinions, and emotions verbally through speaking. Voice, articulation, and fluency are the three basic components of speech production [11][12]. Adult speech reaches its peak when approximately 100 muscles work in unison to produce 14 distinct sounds per second [12].

Our primary goal is to develop a robust and long-lasting model for speech-based emotion recognition, which entails the automated identification of emotions conveyed through spoken language. The timing, pitch, and intensity of speech are among the acoustic characteristics that are mainly examined in this research. The objective is to use machine learning techniques to classify the speaker's emotional state. Precise identification of emotions in speech not only contributes to our understanding of human communication but also can facilitate the creation of emotionally intelligent systems in a variety of disciplines. Our research's well-defined goal is to outperform other established, traditional methodologies in terms of accuracy.

2. Objective

A conceptual model that explains a system's behaviour, composition, and other components is called an architecture. The TESS dataset, which we import from Kaggle, is the dataset that we are using in this model. Since features extracted from the audio stream and utilised as model input would produce noticeably improved results, we're utilising the MFCC in order in order to extract characteristics.

The training data is used to train the model, with the remaining data being divided 80:20 between training and testing. The prototype is then evaluated to determine whether or not it can predict the emotion accurately using the testing data. Next, we compute the model's accuracy to see whether it meets our desired threshold.

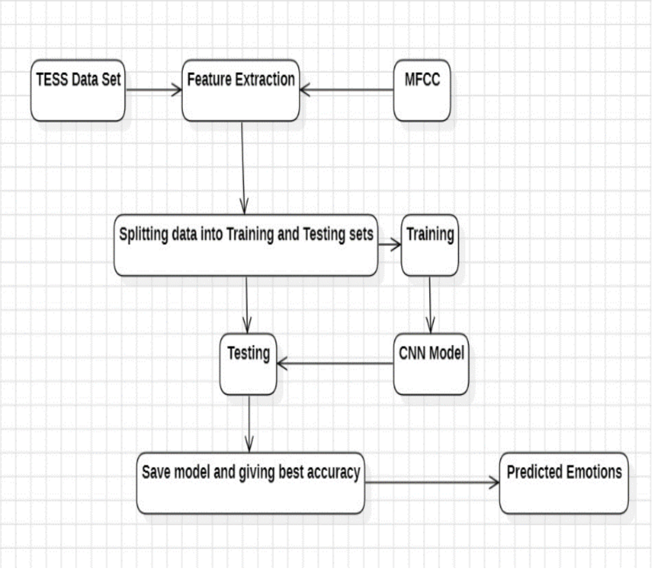

TESS Data Set
Feature Extraction
MFCC
Splitting data into Training and Testing sets
Training
Testing
CNN Model
Save model and giving best accuracy
Predicted Emotions

Fig1: System Architecture

3. Methodology

Advanced human-computer interaction (HCI) programmes that are more user-aware rely heavily on emotion recognition. Image, speech, face, and voice recognition problems are among the many that deep learning models are used to tackle [17], [18], [19], and [20]. Convolutional Neural Networks (CNN) appear to be the best option because our research involves a classification-based problem. If we would want, we can also employ Long-Short Term Memory models and Multilayer Perceptrons, however they don't work well enough to detect absolute emotions and have very low accuracy. So, for our project we are going to select Convolutional Neural Networks model.

I. About Dataset

We are using the TESS dataset in our investigation. The audio recordings in the TESS dataset, an acronym for Toronto Emotional Speech Set, were produced for scientific investigations in emotional computing, namely in the domain of speech emotion identification. It contains recordings of actors portraying several different emotions: anger, disgust, fear, happiness, sadness, and neutrality.

Two actors, ages 26 and 64, performed a set of 200 target words in the carrier phrase "Say the word _." Anger, contempt, fear, joy, grief, and neutral were among the numerous emotions that were recorded on tape for the compilation. In total, there are 2,800 data points (audio files). The two female performers and their emotions are categorised under separate folders in the dataset's organisational structure. Moreover, the audio file containing all 200 target words is included. The audio file is in the WAV format [16].

II. Feature Extraction

We are taking Mel-frequency Cepstral Coefficient (MFCCs) as feature extraction for our study. MFCC is the widely used technique for extracting the features from the audio signal. Tone variations are a common component of speech signals; each tone has a real frequency, f (Hz), and the Mel scale is used to calculate the subjective pitch. Below 1000 Hz, the mel-frequency scale has linear frequency spacing; above 1000 Hz, the spacing is logarithmic.

The formula used to calculate the mels for any frequency is:

$$mel(f)=2595x\log10(1+f/700)$$

Where, mel(f) is the frequency (mels) and f is the frequency (Hz).

A collection of coefficients known as MFCCs is used to represent the form of a sound signal's power spectrum. They are obtained by first utilizing a method such as the Discrete Fourier Transform (DFT) to convert the raw audio signal into a frequency domain, and then using the mel-scale to simulate how the human ear perceives sound frequency. Ultimately, the mel-scaled spectrum is used to construct cepstral coefficients.

III. Convolutional Neural Network

Although CNNs are more frequently linked with image processing jobs, they can also be employed for voice emotion recognition. They have, nevertheless, been effectively used to examine spectrograms and other audio data formats. In the end, our ultimate commitment to working with Speech-based Emotion detection is CNN.

CNN is mostly used to automatically identify pertinent characteristics in speech data so that it can detect emotions from audio inputs with robustness and efficiency, increase classification accuracy, and capture subtle patterns. CNN's ability to distinguish between two-dimensional information is notable. CNN can chameleonically extract feature to eliminate the dependence on human subjectivity or experience.

Convolutional, pooling, and dense layers—also known as fully connected neural network layers—are the three layers that normally comprise a CNN.

Convolutional Layer:

The core element of a convolutional neural network is a convolutional layer. The fundamental process of a convolutional layer is to apply a filter to an input in order to create an activation. A feature map, also known as an activation map, is created when the same filter is applied to an input multiple times. A feature map in 3x3 format will be the outcome of every 5x5 matrix.

The size of the feature map's matrix is determined by the following equation:

Size of map = $N – F + 1$

If the bias is 1, the filter size is F, and the picture size is N.

Pooling Layer:

In a convolutional neural network, the feature maps that are derived from convolutional layers are downsampled using a layer called the pooling layer. By working independently on each feature map, it decreases their spatial dimensions while maintaining important information. Average pooling and maximum pooling are the two types of pooling levels. The feature map's spatial dimensions are reduced by the pooling layer by reducing the number of parameters and computations in subsequent layers.

Fully Connected Layer:

CNNs use a completely integrated neural network stratum as the last stage of feature extraction and categorization. Every single neuron in this layer is linked to every single neuron in the layer above it. It transfers the spatial data that has been acquired by earlier layers to the intended output classes. Dropout regularisation can be employed to prevent overfitting, and Rectified Linear Unit (ReLU) activation functions are frequently employed in these layers. The mathematical definition of ReLU is:

$$f(x) = \max(0, x)$$

Fully connected layers facilitate complex pattern recognition and decision-making, enabling CNNs to perform tasks like image classification, object detection, and speech emotion recognition.

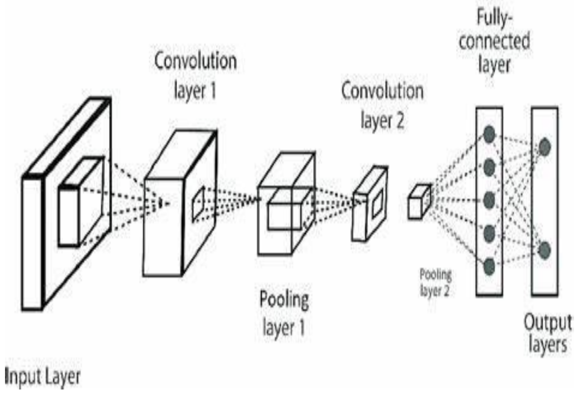


Fig2: Architecture of CNN

Softmax Activation Function:

To turn numbers or logits into probabilities, an activation function known as Softmax is employed. The outcome of a Softmax is a vector (let's call it v) that has the probabilities of each possible outcome. The sum of vector v's probability for each possible outcome or class is one.

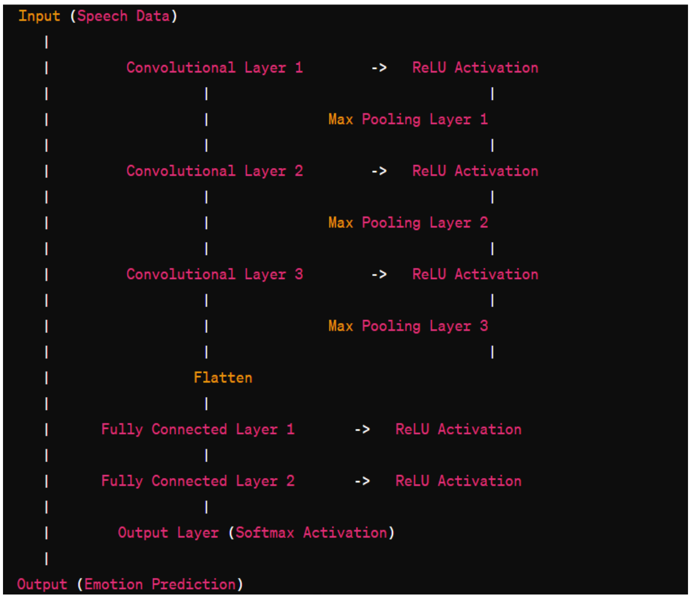
The Softmax activation function for a single class i in a vector of logits z has the following mathematical expression:

Softmax(z_i) = exp(z_i) / Σ(exp(z_j))

for j in all classes where exp (x) represents the exponential function.

It is frequently employed as a neural network's final activation function to normalise the network's output to a probability distribution across anticipated output classes.

Let us see how CNN actually works in recognizing the Speech-based Emotion:



```
Input (Speech Data)
   |
   |         Convolutional Layer 1        ->   ReLU Activation
   |                  |                                |
   |                  |              Max Pooling Layer 1
   |                  |                                |
   |         Convolutional Layer 2        ->   ReLU Activation
   |                  |                                |
   |                  |              Max Pooling Layer 2
   |                  |                                |
   |         Convolutional Layer 3        ->   ReLU Activation
   |                  |                                |
   |                  |              Max Pooling Layer 3
   |                  |                                |
   |                Flatten
   |                  |
   |      Fully Connected Layer 1       ->   ReLU Activation
   |                  |
   |      Fully Connected Layer 2       ->   ReLU Activation
   |                  |
   |        Output Layer (Softmax Activation)
   |
Output (Emotion Prediction)
```

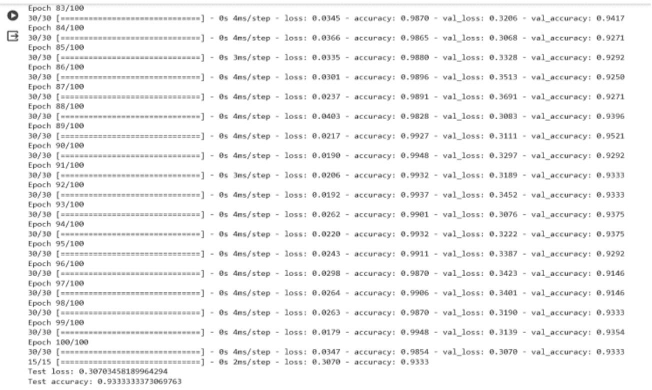
Fig3: Working of CNN for Predicting Emotion

The sequence that Convolutional layers, Pooling layers, Fully Connected layers, and activation functions like ReLU and Softmax go through to use CNN to identify the emotion in speech is shown in detail in the aforementioned graphic.

4. Results and Discussion

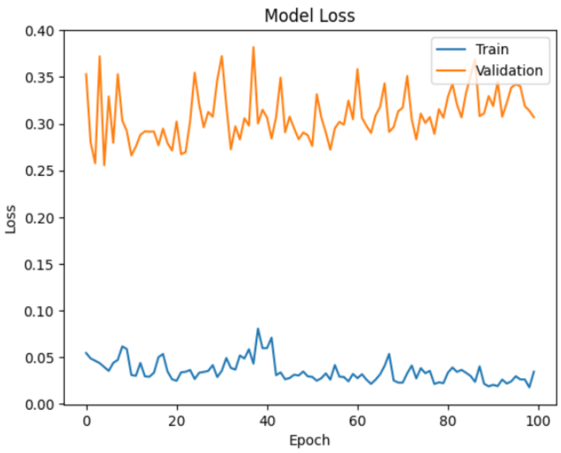
Speech signal is responsible for the communication between humans as it relays as interface to communicate with each other. We choose CNN model approach over other deep learning techniques for our study. Because, CNNs are well-suited for tasks where the spatial arrangement of features is important. In speech emotion recognition, the temporal nature of audio signals is crucial, and CNNs can effectively capture temporal dependencies by considering local patterns over time. This makes CNNs particularly suitable for processing sequential data like speech. After all, we implemented a CNN model for recognising the emotion in speech and gradually, it gives the greatest accuracy and predicted the absolute emotions.

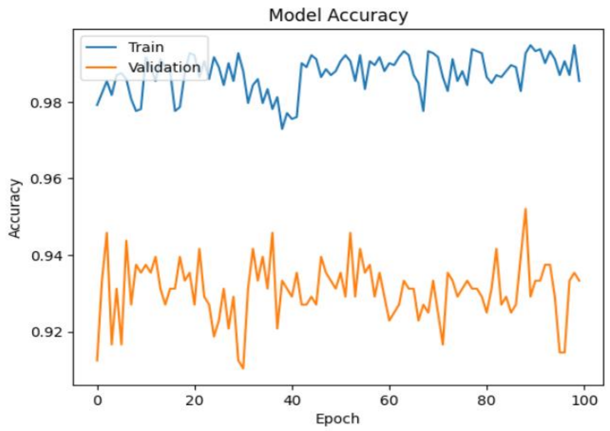
The accuracy and loss metrics are used to evaluate the performance of this model. Our objective is to reduce loss, which is the error between the anticipated and real emotion labels. The percentage of properly identified audio samples is known as accuracy, and it serves as a general indicator of the model's effectiveness. With the use of these measures, one may evaluate how well the model captures and predicts the emotional content of the audio samples.



```
Epoch 83/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0345 - accuracy: 0.9870 - val_loss: 0.3206 - val_accuracy: 0.9417
Epoch 84/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0366 - accuracy: 0.9865 - val_loss: 0.3068 - val_accuracy: 0.9271
Epoch 85/100
30/30 [==============================] - 0s 3ms/step - loss: 0.0335 - accuracy: 0.9880 - val_loss: 0.3328 - val_accuracy: 0.9292
Epoch 86/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0301 - accuracy: 0.9896 - val_loss: 0.3513 - val_accuracy: 0.9250
Epoch 87/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0237 - accuracy: 0.9891 - val_loss: 0.3691 - val_accuracy: 0.9271
Epoch 88/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0403 - accuracy: 0.9828 - val_loss: 0.3083 - val_accuracy: 0.9396
Epoch 89/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0217 - accuracy: 0.9927 - val_loss: 0.3111 - val_accuracy: 0.9521
Epoch 90/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0190 - accuracy: 0.9948 - val_loss: 0.3297 - val_accuracy: 0.9292
Epoch 91/100
30/30 [==============================] - 0s 3ms/step - loss: 0.0206 - accuracy: 0.9932 - val_loss: 0.3189 - val_accuracy: 0.9333
Epoch 92/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0192 - accuracy: 0.9937 - val_loss: 0.3452 - val_accuracy: 0.9333
Epoch 93/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0262 - accuracy: 0.9901 - val_loss: 0.3076 - val_accuracy: 0.9375
Epoch 94/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0220 - accuracy: 0.9932 - val_loss: 0.3222 - val_accuracy: 0.9375
Epoch 95/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0243 - accuracy: 0.9911 - val_loss: 0.3387 - val_accuracy: 0.9292
Epoch 96/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0298 - accuracy: 0.9870 - val_loss: 0.3423 - val_accuracy: 0.9146
Epoch 97/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0264 - accuracy: 0.9906 - val_loss: 0.3401 - val_accuracy: 0.9146
Epoch 98/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0263 - accuracy: 0.9870 - val_loss: 0.3190 - val_accuracy: 0.9333
Epoch 99/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0179 - accuracy: 0.9948 - val_loss: 0.3139 - val_accuracy: 0.9354
Epoch 100/100
30/30 [==============================] - 0s 4ms/step - loss: 0.0347 - accuracy: 0.9854 - val_loss: 0.3070 - val_accuracy: 0.9333
15/15 [==============================] - 0s 2ms/step - loss: 0.3070 - accuracy: 0.9333
Test loss: 0.30703458189964294
Test accuracy: 0.9333333373069763
```

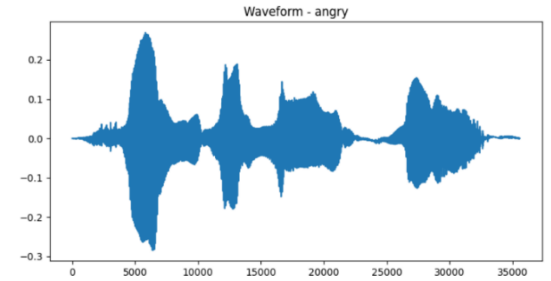
Based on auditory features, these metrics show how well the CNN model was deployed in reliably recognising the emotions.



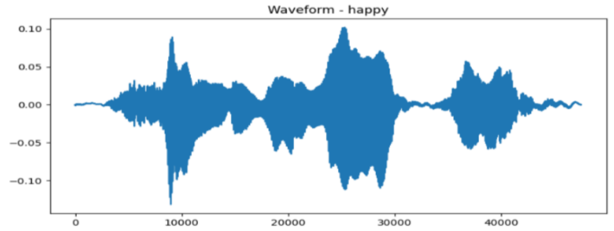


In order to investigate the connection between waveform features and emotional states of various audio signals, we conducted waveform analysis on an audio dataset, and the results are presented in this work.

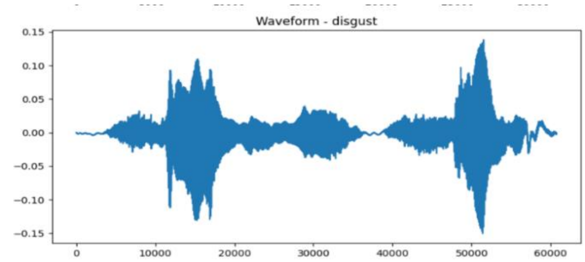
Angry:



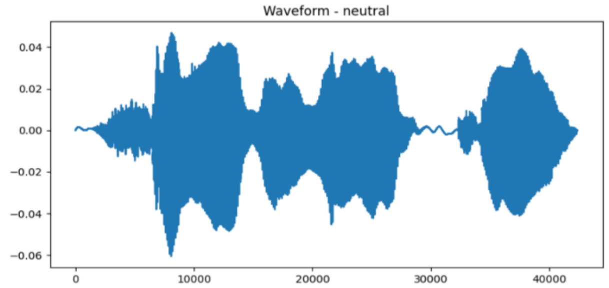
Happy:



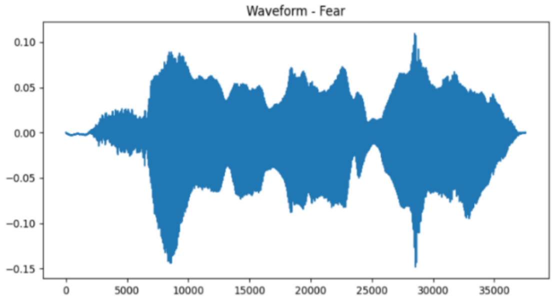
Disgust:



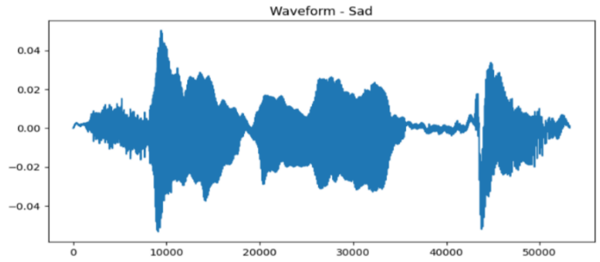
Neutral:



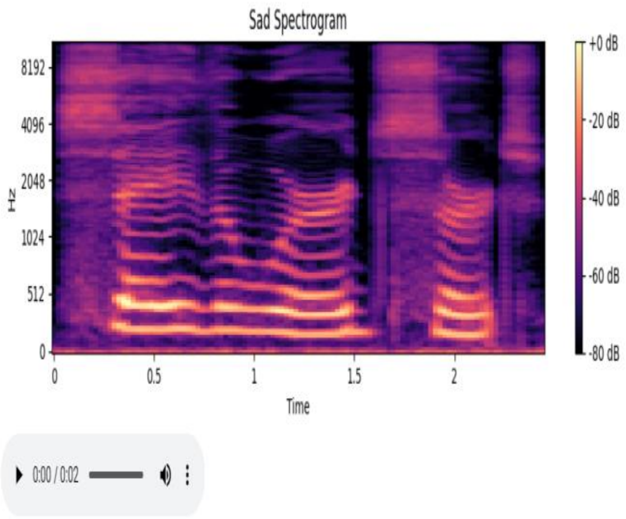
Fear:



Sad:

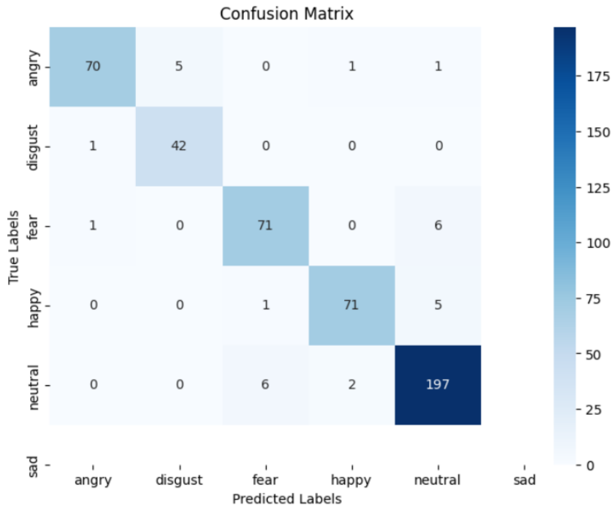


The frequency-time domain graphic representations of the audio signals are shown as follows.



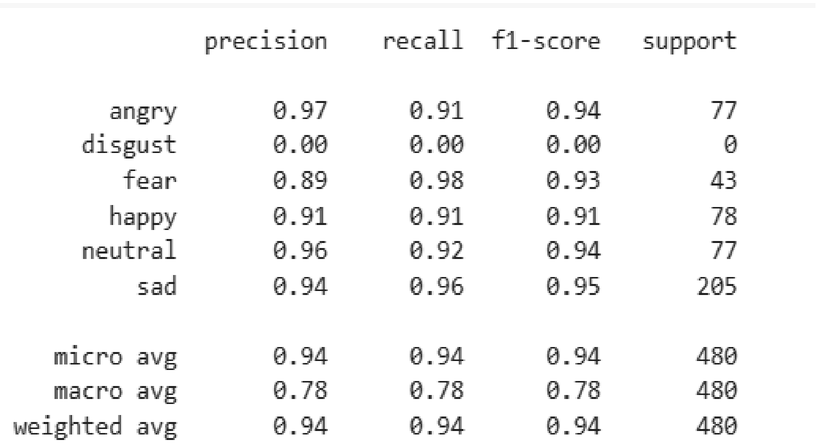
The obtained confusion matrix and classification report of our model is shown as below:

Confusion matrix:



	angry	disgust	fear	happy	neutral	sad
angry	70	5	0	1	1	
disgust	1	42	0	0	0	
fear	1	0	71	0	6	
happy	0	0	1	71	5	
neutral	0	0	6	2	197	
sad						

Classification report:



```
              precision    recall  f1-score   support

       angry       0.97      0.91      0.94        77
     disgust       0.00      0.00      0.00         0
        fear       0.89      0.98      0.93        43
       happy       0.91      0.91      0.91        78
     neutral       0.96      0.92      0.94        77
         sad       0.94      0.96      0.95       205

   micro avg       0.94      0.94      0.94       480
   macro avg       0.78      0.78      0.78       480
weighted avg       0.94      0.94      0.94       480
```

5. Discussion

The traditional methods in Machine learning are weak in practice. In our study, we used CNN model for determining the Emotions in speech. We have given several audio files as inputs to the model to improve the performance in recognising emotion in speech. We discussed about speech recognition methods for extracting audio features from speech sample. The kind of dataset, the classification technique we employed, and the feature extraction all affect how accurate the model is. By contrasting MFCC feature, surprisingly, we achieved a great accuracy of 93.33% which results in great performance.

References

[1] Stânea, A. B., Strilețchi, V., Strilețchi, C., & Stan, A. (2023, October 25). An analysis of large speech models-based representations for speech emotion recognition. 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). https://doi.org/10.1109/sped59241.2023.10314932

[2] Jie, C. (2021, December). Speech emotion recognition based on convolutional neural network. 2021 International Conference on Networking, Communications and Information Technology (NetCIT). https://doi.org/10.1109/netcit54147.2021.00028

[3] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access, 7, 117327–117345. https://doi.org/10.1109/access.2019.2936124

[4] Ismael, A. M., Alçin, M. F., Abdalla, K. H., & Şengür, A. (2020, September 17). Two-stepped majority voting for efficient EEG-based emotion classification. Brain Informatics, 7(1). https://doi.org/10.1186/s40708-020-00111-3

[5] Şengür, D., & Siuly, S. (2020, October 22). Efficient approach for EEG-based emotion recognition. Electronics Letters, 56(25), 1361–1364. https://doi.org/10.1049/el.2020.2685

[6] Alakus, T. B., Gonen, M., & Turkoglu, I. (2020, July). Database for an emotion recognition system based on EEG signals and various computer games – GAMEEMO. Biomedical Signal Processing and Control, 60, 101951. https://doi.org/10.1016/j.bspc.2020.101951

[7] Wei, C., Chen, L. L., Song, Z. Z., Lou, X. G., & Li, D. D. (2020, April). EEG-based emotion recognition using simple recurrent units network and ensemble learning. Biomedical Signal Processing and Control, 58, 101756. https://doi.org/10.1016/j.bspc.2019.101756

[8] Y. Hifny and A. Ali, "Efficient Arabic emotion recognition using deep neural networks", *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 6710-6714, May 2019.

[9] Zhang, L. M., Ng, G. W., Leau, Y. B., & Yan, H. (2023). A Parallel-Model Speech Emotion Recognition Network Based on Feature Clustering. IEEE Access, 11, 71224–71234. https://doi.org/10.1109/access.2023.3294274

[10] Ahn, Y., Lee, S. J., & Shin, J. W. (2021). Cross-Corpus Speech Emotion Recognition Based on Few-Shot Learning and Domain Adaptation. *IEEE Signal Processing Letters*, *28,* 1190–1194. https://doi.org/10.1109/lsp.2021.3086395

[11] Hariharan M, Vijean V, Fook CY, Yaacob S. “Speech stuttering assessment using sample entropy and Least Square Support Vector Machine.” In: 8th International Colloquium on Signal Processing and its Applications (CSPA). 2012. pp. 240-245.

[12] Sabur Ajibola Alim, Nahrul Khair Alang Rashid “Some Commonly Used Speech Feature Extraction Algorithms”, 2018. doi: 10.5772/intechopen.80419

[13] R. Lotfian and C. Busso, "Curriculum Learning for Speech Emotion Recognition From Crowdsourced Labels", *IEEE/ ACM Transactions on Audio Speech and Language Processing*, vol. 27, no.4, pp.815-826, 2019. doi:10.1109/ TASLP.2019.2898816.

[14] Z. Huang, M. Dong, Qirong Mao, Y. Zhan ”Speech Emotion Recognition using CNN”, pp. 801-804, 2014, doi: 10.1145/2647868.2654984

[15] Sun, L., Chen, J., Xie, K. *et al.* “Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition”. *Int J Speech Technol* **21**, 931–940, 2018. https://doi.org/10.1007/s10772-018-9551-4

[16] Pichora-Fuller, M. K., & Dupuis, K. (2020, February 13). Toronto emotional speech set (TESS). https://doi.org/10.5683/sp2/e8h2mf

[17] A. M. Badshah, J. Ahmad, N. Rahim, Sung W. B, “Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network”,2017, doi: 10.1109/PlatCon.2017.7883728.

[18] Hyan-Soo B, Ho-Jin L, Suk-Gyu L, “Voice recognition based on adaptive MFCC and deep learning”,2016, doi: 10.1109/ICIEA.2016.7603830

[19] S. Mittal, S. Agarwal, M. J. Nigam, “ Real Time Multiple Face Recognition: A Deep Learning Approach ”, 2018, doi:10.1145/3299852.3299853

[20] Kun Han, Dong Yu, Ivan Tashev, “Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine”, Interspeech 2014