Automatic Emotion Recognition using Deep Neural Network

R. Sujatha
School of Information Technology & Engineering, Vellore Institute of Technology, India;
r.sujatha@vit.ac.in
Jyotir Moy Chatterjee (Corresponding Author)
Department of Information Technology, Lord Budha Education Foundation, Kathmandu, Nepal
jyotirmoy.chatterjee.cse@gmail.com
Baibhav Pathy
School of Electrical Engineering, Vellore Institute of Technology, Vellore, India
baibhav.pathy2019@vitstudent.ac.in
Yu-Chen Hu
Department of Computer Science and Information Management, Providence University, Taipei, Taiwan, China
ychu@pu.edu.tw

Abstract

Emotions are a vital semantic part of the human correspondence. Emotions are significant for human correspondence as well as basic for human-computer cooperation. Viable correspondence between people is possibly achieved when both the importance and the emotion of the correspondence are perceived by all groups included. Understanding the significance of language has generally been concentrated on in natural language processing (NLP) as a semantic examination. In NLP, the text can be handled appropriately for classification. Emotion detection from facial emotion is the subfield of social signal processing applied in a wide assortment of regions, explicitly for human and PC collaboration. Many researchers have proposed various approaches, generally utilizing machine learning concepts. Automatic emotion recognition (AER) is significant for working with consistent intuitiveness between a person and a smart device toward fully acknowledging an intelligent society. Many researchers examined cross-lingual and multilingual speech emotion as a stage toward language-free emotion acknowledgment in natural speech. In the present work, we are proposing a deep learning-based AER system using four openly accessible datasets, namely Basic Arabic Vocal Emotions Dataset (BAVED), Acted Emotional Speech Dynamic Database (AESDD), Urdu written in Latin/Roman Script (URDU), and Toronto Emotional Speech Set (TESS), by utilizing the Jupyter notebook and a Pythonq library for music and audio synthesis named Librosa. The experimental results exhibited that the proposed approach achieves better than the existing approaches, i.e., the accuracy of the proposed system with the URDU dataset is 96.24%, the TESS dataset is 99.10%, the AESDD dataset is 65.97%, and the BAVED dataset is 73.12%.

Keywords: Automatic Emotion Recognition, Natural Language Processing, Python, Librosa Library, Machine Learning, Deep Learning.

1. Introduction

The perception, as well as characterization of emotions in speech, are a few of the highly conspicuous examination points that have recently acquired fame in the man-machine association. Having perceived the sentiments or emotions in human discussions could profoundly affect grasping a human's physical and mental circumstances [1]. The strategies for signal processing as well as artificial intelligence remain generally used to perceive human emotions because of features removed from pictures, videos, or speech signals. Be that as it may, these features could not perceive the apprehension feeling with similar accuracy as different emotions [2].

Notwithstanding the reasons, cross-lingual characterization might work with emotional acknowledgment for situations with no or just a modest quantity of interpreted information in the objective dialect, which we allude to as a low-asset setting. A way to deal with multilingual emotion characterization utilizing language-recognizable proof and model choice is introduced in [3]. As opposed to this work, where linguistic-subordinate prototypes are prepared and combined afterward chosen appropriately, authors look at the presentation of one model prepared in numerous languages. One more procedure to join double dialects for emotion acknowledgment, portrayed in [4], is to utilize histogram adjustment to eliminate cross-language fluctuation. In [5], the creators contrast programmed cross-lingual perception and human impressions of emotion [6]. Cross-corpus emotion identification has been concentrated on by different scientists to further develop the characterization precision across various languages. These examinations utilized different openly accessible datasets to feature intriguing patterns concerning cross-corpus emotion identification [7]. Albeit a couple of studies have tended to cross-corpus emotion acknowledgment issues as detailed in [7], concerns for emotion acknowledgment for marginal dialects similar to Urdu were not widely investigated. Urdu is the authority dialect of Pakistan and is among the 22 authority dialects perceived in the Indian Constitution [8].

The most important contributions of the proposed effort are as follows:

- We propose a novel method for emotion recognition using a Deep Neural Network (DNN) utilizing a Python library for music as well as audio synthesis titled Librosa Library.
- We have experimented with the proposed approach with four openly accessible datasets (URDU [21], TESS [22], AESDD (only audio) [23], and BAVED [24]) using the Jupyter notebook to check the accuracy.

The organization of the manuscript is as follows: Section 2 introduces the details about some of the existing works in this area; section 3 depicts the proposed mechanism in detail; section 4 shows the detailed results received with discussions. Lastly, section 5 summarizes the paper.

2. Literature Review

The motivation behind AER using DNN is to develop a system that can accurately recognize human emotions in real time. This has potential applications in various fields such as psychology, human-computer interaction, healthcare, etc. For example, in psychology, AER can help diagnose and treat mental health disorders such as depression and anxiety. In human-computer interaction, AER can enhance the user experience by adapting the system's response to the user's emotional state. In healthcare, AER can help monitor patients' emotional states and provide personalized care.

In [9], a review of the hypothetical and real-time work offers new and expansive perspectives on the most recent examination of emotion identification from bimodal data, including facial and vocal articulations. In [10], the authors explored the issues of cross-lingual emotion identification for the Urdu dialects as well as contributed to URDU—the first unconstrained Urdu-language discourse emotional dataset. The creators suggest the adhesion of prosodic as well as phantom features from a set of painstakingly chosen features to acknowledge crossover acoustic features for working on the undertaking of emotion identification [11].

AER using DNNs has been an active area of research in recent years. Researchers have explored various approaches to extract features and classify emotions using deep neural networks. One common approach is to use facial expression analysis to recognize emotions[28]. For example, the Convolutional Neural Network (CNN) has been used to detect facial features, such as eyes, mouth, and eyebrows, to recognize emotions. Another approach is to use physiological signals, such as Electroencephalography (EEG), Electrocardiography (ECG), and Galvanic Skin Response (GSR), to recognize emotions[29]. DNNs such as Long Short-Term Memory (LSTM) and Deep Belief Networks (DBN) have been used to analyze physiological signals and classify emotions[30]. Additionally, some researchers have explored the use of multimodal approaches that combine facial expression analysis and physiological signals to improve emotion recognition accuracy[31]. Overall, DNNs have shown promising results in recognizing emotions automatically, which has potential applications in various fields such as psychology, human-computer interaction, and healthcare. However, there are still some challenges in automatic emotion recognition using DNNs that need to be addressed. One challenge is the availability of labeled datasets. DNNsrequire large amounts of labeled data to train effectively. Unfortunately, there are limited datasets available for emotion recognition, which can affect the performance of deep neural networks. Another challenge is the variability of emotions. Emotions are complex and can vary depending on the individual, culture, and context[32]. Therefore, DNNsmust be trained on diverse datasets and be able to generalize to new situations. Additionally, there is a need to interpret the results of DNNsin emotion recognition. DNNs can be seen as a "black box," which makes it difficult to understand how they arrive at their decisions[33]. Therefore, researchers must develop

methods to interpret the results of DNNsin emotion recognition to ensure that they are reliable and trustworthy. Despite these challenges, automatic emotion recognition using DNNsshows great promise and is an active area of research [34]. Table 1 presents the comparative analysis of various existing works with their limitations.

Table 1. Comparative Analysis

Sl. No.	Reference	Advantage	Limitations
1	[35]	Explored the use of constant-Q transform modulation spectral features (CQT-MSF) for speech emotion recognition (SER).	Experiment with joint spectral and temporal modulation features and analyze their suitability for SER.
2	[36]	Presented a publicly available Urdu Nastalique Emotions Dataset (UNED) of sentences and paragraphs annotated with different emotions and proposes a DL-based technique for classifying emotions in the UNED corpus.	Fear and Anger emotions are represented in a small number of sets in the UNED corpus and need to collect relevant instances of these emotions.
3	[37]	Introduced the rhythm-specific multi-channel CNN-based approach for automated emotion recognition using multi-channel EEG signals.	Investigation for multi-class emotion recognition tasks using multichannel EEG signals is not done.
4	[38]	Introduced a temporal multimodal fusion approach with a DL model to capture the non-linear emotional correlation within and across EEG and blood volume pulse (BVP) signals and to improve the performance of emotion classification.	This study was based on a limited number of participants and it would be worthwhile to expand the study and investigate the performance and the methods with a larger sample of participants.
5	[39]	Suggested a novel technique called facial emotion recognition using convolutional neural networks (FERC)	FERC could be the starting step, for many of the emotion-based applications such as lie detectors and also mood-based learning for students, etc.

In [12], the authors have effectively fabricated group attention utilizing the neural network (NN) for influence assessments in nature. In [13], the authors proposed an original deep convolutional NN (DCNN) for natural sound order. The illumination of the principal challenge on emotion identification from utterance [14] gives the biggest to-date benchmark examination under equivalent circumstances on nine standard corpora in the field utilizing the two pre-predominant ideal models: demonstrating on a frame-level through hidden Markov models besides suprasegmental displaying by systematic feature brute-forcing. In [15], the authors suggested a deep learning-built emotion identification model for Arabic speech. In [16], the authors presented an advanced architecture because of acoustic as well as deep features to expand the order precision in the issue of speech emotion identification. In [17], the authors zeroed in on working on the exhibition of an artificial intelligence prototype in the speech dataset, which is the Ryerson Audio-Visual Database of Emotional Speech as well as Song. In [18], the authors planned the architecture for examining similitude in clusters, which depends on a key succession choice method. In [19], the authors introduced a Dense-DCNN model for perceiving speech feelings. In [20], the authors examined and proposed an advanced long short-term memory (LSTM) Network along with Transformer Encoder to gain proficiency with the drawn-out conditions in speech flags and feelings classification.

To address the existing research gaps, researchers can focus on developing novel approaches to extract features and classify emotions using DNNs. They can also explore the use of multimodal approaches that combine facial expression analysis and physiological signals to improve emotion recognition accuracy. Additionally, researchers can work on developing methods to interpret the results of DNNs in emotion recognition to ensure that they are reliable and trustworthy.

3. Proposed Methodology

In this part, we show our multimodal strategy that uses a DNN model and a mechanism for the classification of multilingual Audio files. We start by outlining the audio signal pre-treatment and data segregation methods. The technique of attentive models that may capture contextual information is then described. Finally, we present in full the suggested architecture for the DNN model. Fig. 1 presents the methodology we are going to apply for our experiment.

The use of DNNs in AER has been shown to provide highly accurate results. DNNs are capable of automatically learning features from raw data, which can be used to classify emotions from different modalities such as facial expressions, speech, and physiological signals. In the case of speech, DNNs can learn to extract features from the raw audio signal, which can be used to classify emotions. However, when dealing with multilingual audio files, the task becomes more challenging because different languages have unique phonetic and prosodic characteristics. Therefore, the development of

a DNN model and a mechanism for the classification of multilingual audio files can greatly improve the accuracy of AER for multilingual datasets. This model would need to be trained on a diverse set of multilingual audio data to ensure that it can generalize to new languages and contexts. Additionally, the mechanism for classification would need to consider the unique phonetic and prosodic characteristics of each language to ensure that the model can accurately classify emotions across different languages. Overall, the use of DNNs in AER for multilingual audio files has the potential to provide highly accurate results, which can be applied in various fields such as psychology, healthcare, and human-computer interaction.

In Fig. 1, we first input the 4 datasets, i.e., TESS, URDU, AESDD, and BAVED. These data sets are then used to extract the different types of features, for example, context encoder (mono channel and stereo channel), emotion encoder (Mel-Frequency Cepstral Coefficients (MFCCs [27]), Contrast, Chroma, Mel), and text encoder (audio data and sample rate). These are used to reduce dimensional class using principal component analysis into features and class. The reduced dimensional feature is inputted into different classification modules for deep learning. Here we use the artificial neural network (ANN), Relu, and Softmax as the classifier. And after running the model, we get output as different emotions of different linguistic languages.

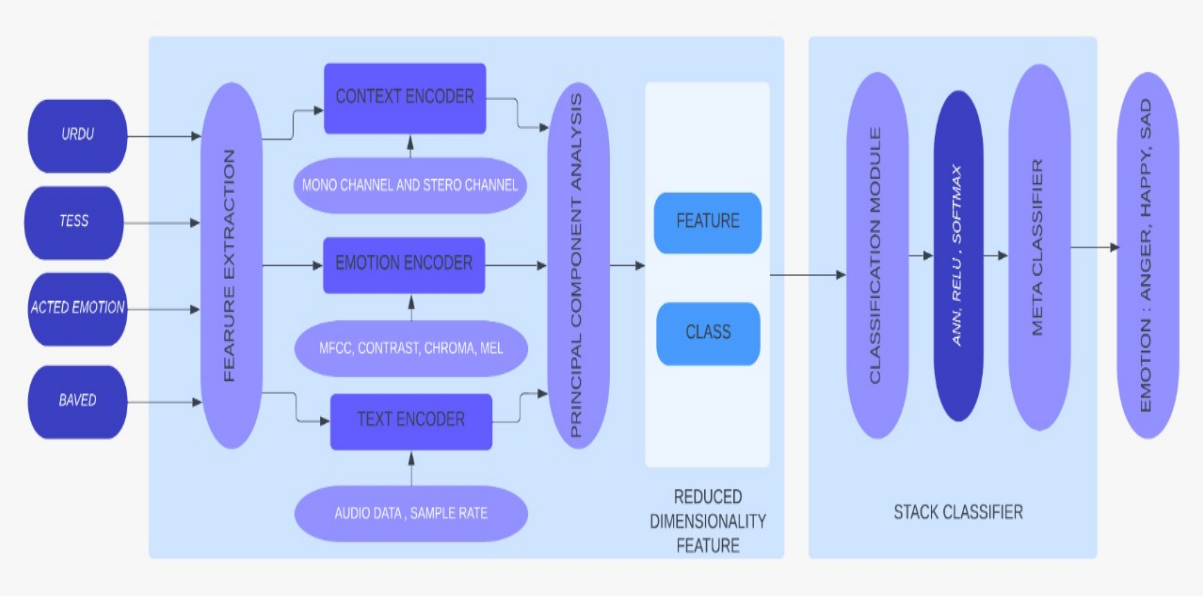

Figure 1. The processing steps of the proposed approach

	Our Paper	A computerized approach for automatic human emotion recognition using sliding mode singular spectrum analysis	A novel S-LDA features for automatic emotion recognition from speech using 1-D CNN"
Proposed method	our proposed method utilizes a multimodal strategy with a DNN model and a mechanism for the classification of multilingual audio files. It involves audio signal pre-treatment, data segregation, and the use of attentive models to capture contextual information. The architecture of the DNN model is presented, which incorporates different datasets and feature extraction techniques.	This technique utilizes sliding mode singular spectrum analysis for emotion recognition. It involves analyzing the temporal dynamics of emotional signals using sliding windows and singular spectrum analysis. The method proposed in this paper involves a three-step process for analyzing and classifying human emotional states using physiological signals such as EEG and ECG. It includes decomposing signals using sliding mode singular spectrum analysis (SM-SSA), computing discriminatory features from the extracted	This technique combines S-LDA features with a 1-D CNN for automatic emotion recognition from speech. It involves dimensionality reduction using linear discriminant analysis (LDA) to extract discriminative representations from the speech data. This method focuses on speech-based emotion recognition and introduces a new algorithm called Shifted Linear Discriminant Analysis (S-LDA) to extract modified features from low-level features like MFCC and Pitch. These modified features are then fed into a 1-D CNN to extract high-level features for emotion recognition.
Performance	The performance of our proposed method is mentioned in the description. It would evaluate its performance using appropriate metrics such as accuracy, precision, recall, or F1 score on relevant	The proposed method achieved a classification accuracy of 92.38% when evaluated on the DREAMER and AMIGOS databases. The results indicate that the method can effectively identify	The proposed technique achieved high accuracy on three standard databases: Berlin EMO-DB, Surrey Audio-Visual Expressed Emotion (SAVEE), and eNTERFACE. The accuracy results were

	datasets to compare it with other techniques. The proposed method achieved a classification accuracy of 96.24 % when evaluated	different human emotional states and outperform existing emotion recognition methods.	99.59% for the Berlin database, 99.57% for the SAVEE database, and 86.41% for the eNTERFACE database. The results demonstrate the effectiveness of the proposed technique compared to state-of-the-art methods.
Modality	This method focuses on multilingual audio files and incorporates a multimodal strategy using a DNN model.	This method focuses on analyzing and classifying human emotional states using physiological signals, specifically EEG and ECG.	This method focuses on speech-based emotion recognition.
Feature Extraction	The method involves extracting different types of features from the audio data, including context encoder, emotion encoder (e.g., MFCCs, Contrast, Chroma, Mel), and text encoder (audio data and sample rate).	The signals are decomposed into reconstructed components using sliding mode singular spectrum analysis (SM-SSA). Discriminatory features such as information potential (IP) and centered correntropy (CEC) are computed from the extracted components.	It introduces the Shifted Linear Discriminant Analysis (S-LDA) algorithm to extract modified features from low-level features like MFCC and Pitch.
Classification	The reduced dimensional features are inputted into different classification modules using artificial neural network (ANN), Relu, and Softmax as the classifier.	Various ML classifiers are used to classify human emotional states based on the extracted features.	The modified features are inputted into a 1-D CNN to extract high-level features for emotion recognition.
Accuracy	For the BAVED dataset we received	The method achieved a classification accuracy of 92.38% on	he method achieved high accuracy on three standard databases:

	an accuracy of %73.13 . For the AESDD emotion (only voice) , we got an accuracy of 65.97% . For the TESS dataset , we got 99.1% accuracy.	the DREAMER and AMIGOS databases.	99.59% for the Berlin database, 99.57% for the SAVEE database, and 86.41% for the eNTERFACE database.

3.1. Audio File and Libraries

A kind of energy made by vibrations is sound. The vibration of a thing makes the atoms of the air around it moves. These atoms crash into particles close by, making those particles also vibrate [5]. Subsequently, they slam into really encompassing air atoms. Until the particles run out of energy, this sound wave "chain response" proceeds. While an outcome, as the sound wave goes through the air, there are various sub-atomic crashes. However, the air particles themselves don't move with the wave. Every molecule voyages from its resting position when it gets perturbed and finally returns to it. In this work, we have taken 4 datasets

1) URDU [21]

2) TESS [22]

3) AESDD (only audio) [23]

4) BAVED [24]

The Urdu [21] dataset consists of the audio file in the Urdu language, a native Indo-Aryan language tagged for sentiments (happy, sad, neutral, angry). Another audio dataset used here was TESS [22], which solely includes a very high-quality female voice. A slightly unbalanced representation results from the other dataset's preponderance of male speakers. Accordingly, in provisions of generality, this dataset would function as a decent training dataset for the emotion classifier but not overfitting.

The two primary groups of emotional speech databases are expressions of performed emotional speech as well as those that hold a genuine emotional speech. We have also used AESDD for voice emotion identification. It includes Greek language expressions of staged passionate speech. The discovery that there was non-availability of a publicly accessible, high-quality database for SER [28] in Greek through the study of emotion-based light support for emotional functioning [1] inspired the database's construction. The five emotions included in the database are anger, contempt, fear, happiness, and sorrow.

Together through a set of seasoned artists who were very interested in the suggested structure, the first iteration of the spoken language emotion identification dataset happened developed. The term

"dynamic" in the context of the AESDD refers to the project's goal of continuously growing the database beyond the participation of artists as well as entertainers who remain active or concerned in it. Even though the call for contributions is directed toward performers, the SER models trained on the AESDD are not only achievements.

The BAVED is a collection of Arabic words with various emotional expression levels captured in audio/wave format. Seven Arabic terms have been discovered in this dataset. Three degrees of conveyed emotions are kept track of in each of the original statements. When the speaker is exhibiting a level 0 emotion, it is comparable to feeling worn out or depressed. Finally, level 2 is when the speaker expresses a high degree of happy or negative emotions. Level 1 is the standard level, which is the way the speaker regularly speaks while expressing a neutral feeling (happiness, joy, sadness, anger, etc.)

Figure 2. An example of a stereo audio file (open-source web)

Before ascending forward, we need to understand the difference between mono and stereo sound channels. The number of channels utilized to capture and playback audio determines whether a sound is monophonic (mono) or stereophonic (stereo). Stereo sounds are recorded using 2 audio channels, whereas mono signals are produced and played back using 1 audio channel. The most obvious distinction to the listener is that stereo sounds may create the illusion of breadth, but mono sounds cannot.

Stereo playback systems are playback devices that have two speakers. Left and right channel information is included in stereo audio files like stereo MP3 and WAV files, which instruct the left and right speakers when to push and pull air.

In a digital audio workstation (DAW), a stereo audio file (Fig. 2) has two waveforms, one for each channel. Each waveform corresponds to a single audio channel. One audio channel only exists in mono audio files (Fig. 3).

Figure 3. An example of a mono audio file (open-source web)

Stereo systems can simulate the localization of the sound source. The capacity of a person to pinpoint a sound source's location within a certain area is known as sound source localization. When

a dog barks, for instance, you can for the most part tell where the voice is coming from and the distance away the sound source (the dog) is. Indeed, even with their eyes shut, most people ought to have the option to find noises more accurately. It seems logical to suppose that the left and right speakers, which make up a stereo system, are two separate sound sources. Only a two-dimensional picture with depth and height can be produced by mono playback systems, which only have one speaker. To give our brain the directional time variations it needs to sense breadth, two speakers are necessary.

In the present work, we have used the Librosa Library and Scipy Library for the pre-processing of data and extracting features. For the analysis of audio, the Python package Librosa was created. It focuses on recording audio data so that it can be converted into a data block. The materials along with examples are useful for learning how to work with audio data science projects. Librosa is mostly utilized for working with audio data, such as when creating music (using LSTMs) or doing automatic speech recognition. It offers the components required to develop music information retrieval systems. The rapid Fourier transform (FT) converts the sound from the time domain to the frequency domain since a sound wave is composed of multiple single-frequency vibrations. A signal is transformed from a time domain to a frequency domain utilizing the FT. As a result of this action, we will receive something called the spectrum.

Using Librosa's "stft" function is an efficient and straightforward technique to demonstrate how an audio stream is transformed into a frequency domain. The windowed signal's length when padding in conjunction with zeros, hop length, frame size, along with fast FT size are some of the most crucial characteristics. We can choose to consider time-frequency representation, often known as a spectrogram (Fig. 4). The frequency intensities are time-resolved in the spectrogram. A spectrogram is composed of the squared short-time FT (STFT) magnitude. It illustrates how frequency is displayed along the y-axis, time is indicated along the x-axis, and associated amplitudes are colored. Since people can only pronounce one phoneme at a time, the FFT window for speech recognition tasks is 20–30 ms. It has windows that are 50% overlapping. Depending on the use scenario, it might range from 25% to 75%. If s = 16 kHz and the window length is 25 ms, the number of samples in the window is 16000 25 0.001 = 400 units.

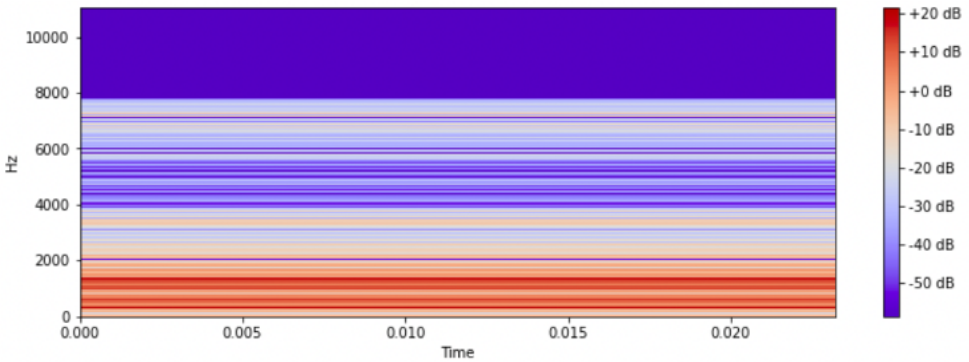

Figure 4. An example of the time amplitude representation using Librosa

The spectrum envelope's general form can be adequately depicted by MFCCs (Fig. 5), a condensed set of features (typically 10–20). Further MFCC analysis on feature scaling is possible using the MinMaxScaler module for data preparation. The type of problem we are solving and the features we want to look at will determine which audio features are available in the finest of Librosa.

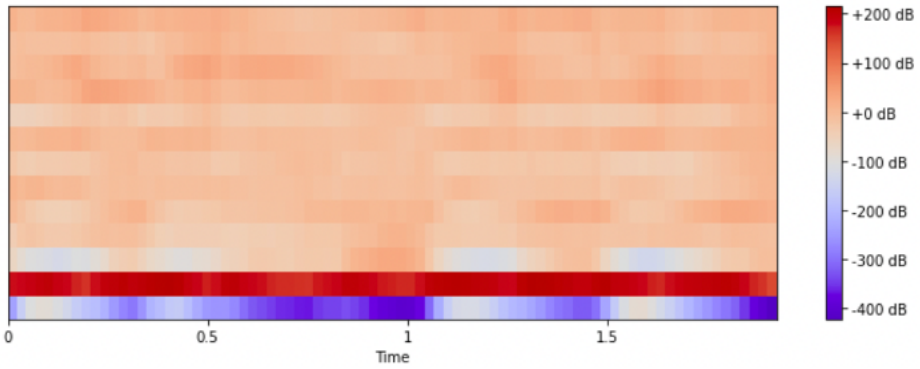

Figure 5. An example of the general form of MFCCs

In scipy, the value of the given data doesn't get normalized, whereas in Librosa the value of sound is getting normalized. Librosa library is used to display data and sample rates. Also, it converges the signal into a monotype and can represent audio signals into normalized patterns -1 to +1 so that a regular pattern is absorbed. It also helps to be able to see the sample rate by keeping the sample rate at 21khz.

3.2. Pre-processing

The most common way of separating data and significance from sound signs for use in examination, order, capacity, recovery, blend, and so on is known as sound examination. Sound evaluation can allude to the hear-able framework and how people decipher the discernible commotion, or it can allude to utilizing cutting-edge innovations, for example, a sound observing framework to survey different parts of clamor waves like amplitude, distortion, and frequency response, besides that's one of the slants of the iceberg. The perception mediums as well as insight strategies shift. The recognizable information by sound information can be assessed by the client for logical, emotive, clear, or other appropriate assessment after it has been seen.

There are numerous computer-readable formats in which audio can be conveyed, including:

- **WAV** - Waveform Audio File Format
- **WMA** - Windows Media Audio Format
- **MP3 -** MPEG-1 Audio Layer 3 Format

The extraction of acoustical highlights relevant to the current task is a typical move toward the sound handling process. Then, dynamic frameworks, including arrangement, identification, and information combination, are utilized. Exanimating the attributes that might be taken from sound records and changed over for ML applications. The linear-prediction cepstral coefficients (LFCCs), bark-frequency cepstral coefficients (BFCCs), MFCCs, spectrum, cepstrum, spectrogram, gamma tone-frequency cepstral coefficients (GFCCs), and other information elements and changes are essential to sound handling. Red, pyAudio Analysis, Librosa, and other Python-based programs are utilized for information extraction and examination of sound records.

Two properties that are especially critical in sound handling are spectrum and cepstrum.

A spectrum is the Fourier change of a source in science [8]. A period space signal is switched over completely to a frequency area signal utilizing a Fourier change. At the end of the day, a spectrum is a portrayal of the time-space sound sign in the frequency area. Taking the spectrum's logarithmic extent and applying an opposite Fourier change results in a cepstrum. Since we utilized an opposite Fourier change, the subsequent sign is neither in the frequency area nor the time-space (since we took the log greatness before the reverse Fourier change). The resultant sign's space is known as the quefrency.

The science of the ear is why we are keen on the sign in the frequency space. Before people can grasp and comprehend a sound, a lot of things need to occur. One happens in the cochlea, a region of the ear that is loaded with liquid and has many tiny hairs joined to nerves. The hairs fluctuate long, with some being more limited than others. Greater hairs are thunderous at lower frequencies, while more modest hairs resonate at higher frequencies. A characteristic Fourier change analyzer, then, at that point, is the ear.

One more intriguing reality about human hearing is that when the sound frequency ascends north of 1 kHz, our ears become less frequency specific. This fits in pleasantly with an idea known as the Mel filter bank. The Mel cepstrum [10] runs a spectrum through the Mel filter bank, decides the log amplitude, and plays out a discrete cosine transform (DCT). The essential information and pinnacles of the sign are extricated through DCT. Moreover, JPEG and MPEG compressions utilize them. The primary concerns of the acoustic data are the maxima. The MFCCs are regularly the initial 13 coefficients taken from the Mel cepstrum. These are regularly used to construct AI models and incorporate an abundance of significant data about sound.

3.3. Multimodal Speech Emotion Recognition based on MFCC and ANN

Through our investigation, we discover that the CNN feature is not a good representation for mining temporal dependencies such as raw sound and speech. We then extract MFCC features for

audio. MFCC, which is frequently employed in automatic speech recognition, generally represents raw speech accurately.

The working step of the multimodal speech emotion recognition based on MFCC and ANN is as follows:

1. Preprocess the speech signal to remove noise and artifacts
2. Extract MFCC features from the preprocessed speech signal
3. Train an artificial neural network (ANN) on a labeled dataset of speech signals and their corresponding emotions
4. During training, adjust the weights of the connections between the neurons to minimize the difference between predicted emotions and actual emotions in the training data
5. Once the ANN has been trained, use it to predict the emotions in new speech signals
6. Optionally, incorporate other modalities such as facial expressions and body language to improve the accuracy of the emotion classification
7. Use fusion techniques such as early, late, or hybrid fusion to integrate the different modalities
8. Optionally, use more advanced machine learning algorithms such as deep neural networks (DNNs) or convolutional neural networks (CNNs) to improve the performance of the emotion recognition system
9. Optionally, use data augmentation techniques such as adding noise, changing pitch or speed, or using different languages or dialects to improve the robustness and generalization of the emotion recognition system.

The vocal tract shape may be accurately described by the MFCC, one of the most often utilized spectral-related characteristics in SER. The vocal tract morphologies vary depending on the emotion being expressed. As a result, using the MFCCs of speech, the model can discern various moods. First, a series of filter banks that match the frequency response properties of the human auditory system filter MFCCs. The name of this filter bank is Mel-filters.

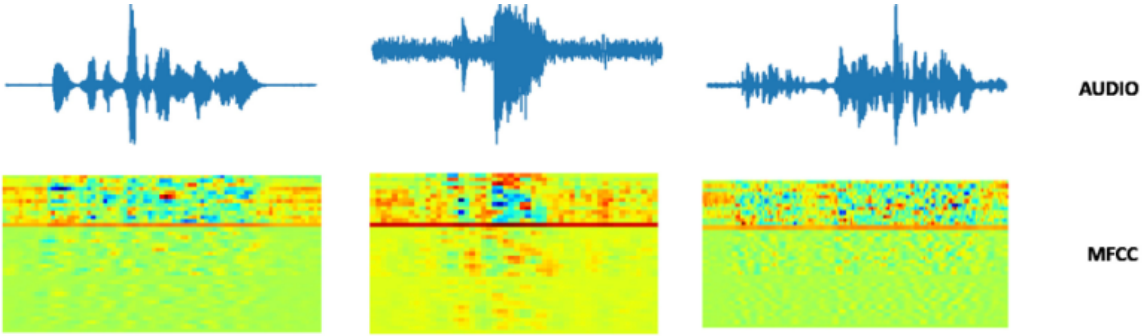

Figure 6. Meyer filter banks

A collection of triangular screens (Meyer filter banks) is used in the MFCC extraction process to filter the sound spectrum in the frequency domain following the characteristics of sensory organs for sound waves (Fig. 6). Equation (1) illustrates the relationship between the triangle filter's center frequency and frequency f.

$$Mel(f) = 2595\log\left(1 + \frac{f}{700}\right) \quad (1)$$

The MFCC approach is premised on the theory of Mel frequency, which is among the most popular and useful ways to characterize characteristics and can correct convolutional channel distortion. Meyer-filter-bank Filtering, Normalization, Windowing, Framing, Fast FT (FFT), DCT, and Logarithmic Energy Calculation are all part of the MFCC extraction process [7]. Here is how the recovery process works: First, we went through each WAV file in the database containing the data set we have gathered and stored the file paths for each voice. This was done to add tags to each file using the path truncation approach, read the WAV files to gather information about them, and finally extract the MFCCs from the WAV files. For the scale feature, we will do mean on the transpose of MFCC.

The attributes that were recovered only represent the frame's properties because the voice signal is continual in the temporal domain [9]. We carried out several MFCC modifications to better reflect the time-domain consistency of the feature. The first-order divergence and the second-order variation are two common therapies. Let c(t) be the digital audio signal data point. Equation (2) shows the computation for the variance.

$$dt = \frac{\sum_{n=1}^{N} n(c_{t+n} - C_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (2)$$

Various systems are designed to investigate the effects of algorithms with various forms and locations on SER. The construction method of one model is provided with consideration of text duration. After pre-processing, we run every input mode through a level of batch normalization to equalize the input layer by scaling and modifying the activation. The gated recurrent unit (GRU) module, a subset of the ANN prototype, receives the outputs from the batch normalization layers [3].

We tried to solve this issue with a neural network. For the URDU datasets, we have set an input layer with 3 neurons. For the experiment, we utilize the Softmax activation function in the output layer. The relative probabilities are computed using the Softmax activation function.

Specifications of the 3 layers are:

- 1st dense layer has 100 neurons with activation Relu and dropout of 0.5
- 2nd dense layer has 200 neurons with activation Relu and dropout of 0.5
- 3rd dense layer has 100 neurons with activation Relu and dropout of 0.5
- The final activation is Softmax

- Train model with epoch 100 and batch size 32.

According to the class labels in the URDU dataset, the last layer is added to the network design in the following line. The URDU dataset has 10 classes (one for each number). Hence 10 units were utilized to create this layer.

4. Results and Discussion

Validating the efficacy of emotion segregation and exploring various batching systems and global acoustic features were the goals of our investigations. We utilized unweighted average recall (UAR) as an assessment metric because, in contrast to weighted average recall (the "traditional" accuracy), it is also meaningful for extremely imbalanced distributions of examples within classes. We employed the Scikit-learn machine training program [4, 32] and the Theano or TensorFlow-compatible Keras deep learning library.

The values were normalized to have a zero mean and a unit standard deviation during the pre-processing stage. Each collection is split into four mini-batches when employing homogenous mini-batches. All learning models are divided into stratified mini-batches of 100 samples each when stratified mini-batches are used. At the beginning of every epoch, mini-batches are generated randomly.

We also looked at how the K-nearest neighbor (KNN) approach performed with various feature sets. Based on prior knowledge, the network topology had five hidden layers (H5), each with a Softmax activation function for the output layers and a rectified linear activation function for the hidden layers. We employed 2048 neurons per hidden layer and ran the training procedure for 100 epochs using the extensive ComParE feature set. A learning rate of 0.01 was used. The feature set was trained for 1000 epochs with 256 neurons per layer. In a pilot study, we discovered that, with this condensed feature set, bigger networks did not outperform this simple architecture. To increase universality for both feature sets, dropout with a probability of 0.5 was applied on the input layer.

Finally, we hypothesize that the trained KNN network may be utilized as a starting point for training better single-task networks since the suggested KNN learning technique enables us to leverage training datasets and regularise the network training. To do this, we built single-task networks for task C using a single trained output layer and parameters W(c) = W1….WH-1, W(c) H from the trained shared hidden layers. Then, using a single emotion database, we retrained the full parameter set W(e).

The gain across datasets for the Refer feature set is considerable at the 0.01 level. The capacity of the model to classify positive samples is measured by precision. The negative and positive samples have an impact on the model's accuracy. All test results, whether rightly or mistakenly identified as positive in Precision, should be considered. A model is measured to have a high recall and deprived accuracy when it properly identifies most positive samples as well as numerous false-positive samples.

Recall enables us to count the number of positive samples that the model properly categorized. A neural model's recall is inclined by positive examples but unpretentious by negative samples. Correctly identifying all positive samples is important to the recall. It does not consider the classification of any negative trials as positive. A model exhibits low recall and high accuracy and precision if it can only categorize a small number of positive samples while classifying a sample as positive. Figs. 7-25 provide us with valuable insights into precision, recall, accuracy, and epoch.

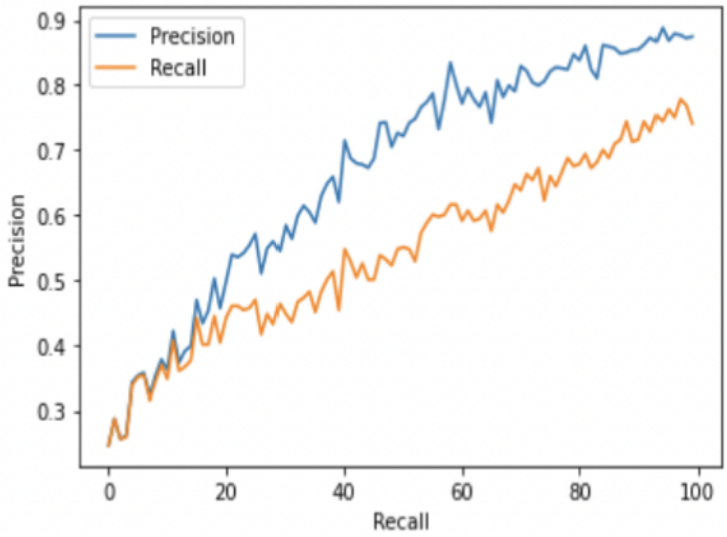

Figure 7. Precision vs. Recall

For the Urdu dataset [21], we got an accuracy of 91.25% with a test loss of 33.7%. The precision-recall graph illustrates the trade-off between precision and recalls at different thresholds. A high recall is associated with low false negative rates, while high accuracy is related to low error rates. A high area under the curve suggests both excellent recall and great precision. High scores for both indicate that the classifier produces accurate results with high accuracy that are primarily positive (high recall). According to equation (3), precision (P) is calculated as the relation of true positives (Tp) to true positives plus false positives (Fp):

$$P = Tp / Tp + Fp \tag{3}$$

According to equation (4), recall (R) is calculated as the proportion of true positives (Tp) to the sum of true positives and false negatives:

$$R = Tp / Tp + Fn \tag{4}$$

These numbers are also connected to the (F1) score, which is represented by equation (5) as the harmonic mean of precision and recall:

$$F1 = 2\left(P \times R\right) / P + R \tag{5}$$

Keep in mind that precision might not drop off with recall. By raising the number of outcomes returned, decreasing a classifier's threshold may raise the denominator, as shown by the definition of

accuracy. If the prior criterion was set too high, the new findings might all be true positives, enhancing precision. Depending on whether the prior threshold was too high or too low, decreasing it further will result in false positives and decreased precision. Here in Fig. 7, since the score is high, the classifier is producing reliable results.

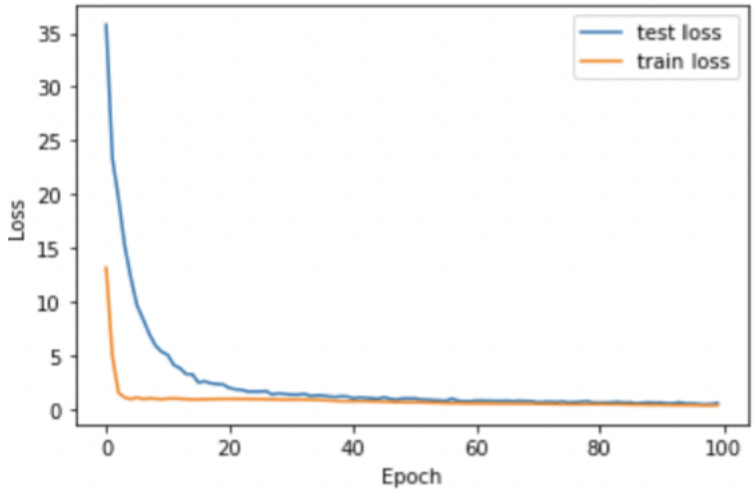

Figure 8. Loss vs. Epoch

A good match is indicated by training and validation losses that stabilize at a modest difference between their final values. Almost always, the training dataset's model loss will be lower than that of the validation dataset. As a result, there will probably be a difference between the validation loss learning curve and the train learning curve. This difference is known as the "generalization gap". If the training loss plot decreases to a stable point, the learning curve plot indicates a satisfactory fit. The validation loss plot declines to a stable point and narrowly differs from the training loss. Since Fig. 8 shows a good fitting curve, therefore.

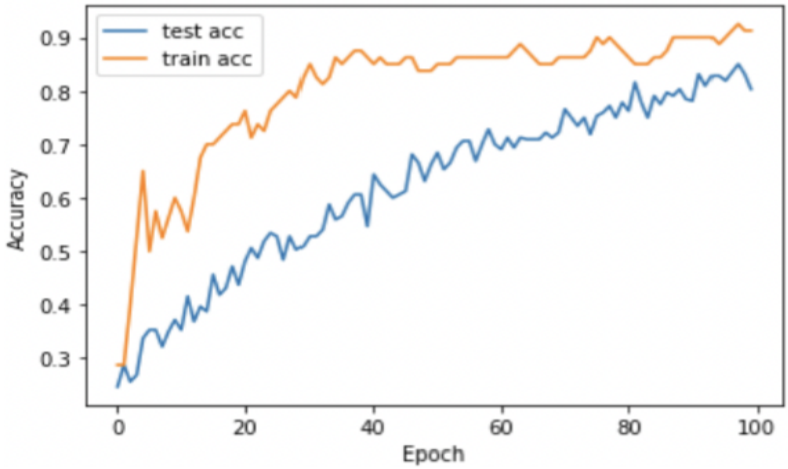

Figure 9. Accuracy vs. Epoch

As the accuracy per epoch increases with both the test case and train case, this indicates the model is likely to be stable and valid, as in Fig. 9.

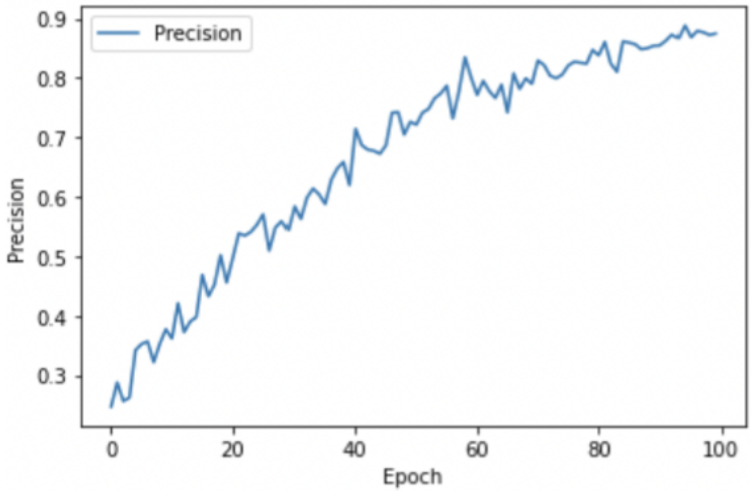

Figure 10. Precision vs. Epoch

The percentage of correctly classified Positive samples to all Positive samples is used to determine accuracy. The model's accuracy measures how accurately it classifies a sample as positive. Accuracy increases as the number of epochs increases.

When the model makes many incorrect Positive classifications or few accurate Positive classifications, the denominator increases and the precision decreases. However, the precision is great when:

- The model generates many True Positives (maximize True Positives).
- The model generates a small number of False Positives (minimize False Positives).

Here the model might be making maximize true positive or minimize false positive as in Fig. 10. For the AESDD emotion (only voice) [23], we got an accuracy of 65.97% with a loss of 12.04%.

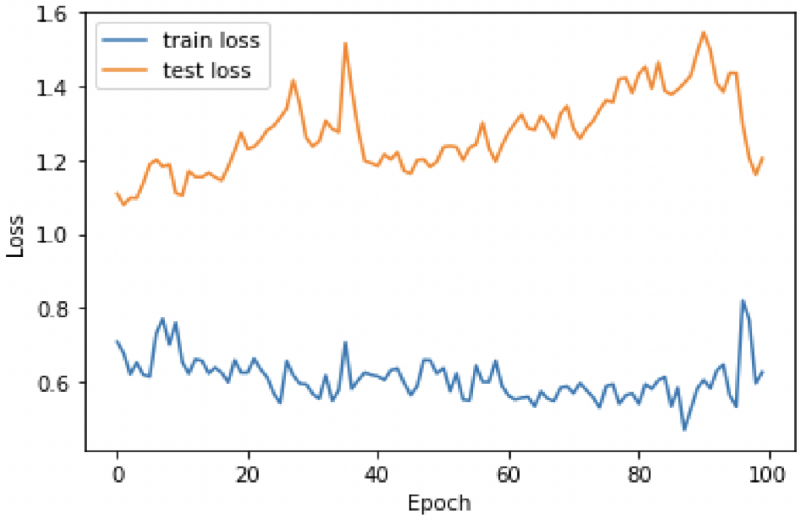

Figure 11. Loss vs. Epoch

Here the loss of the train set, and test set is constant over a long epoch which indicates not a very good match. Although the loss in overall function is minimalistic to 12.04% in Fig. 11, it didn't give us better accuracy because it remains constant over the dataset's range.

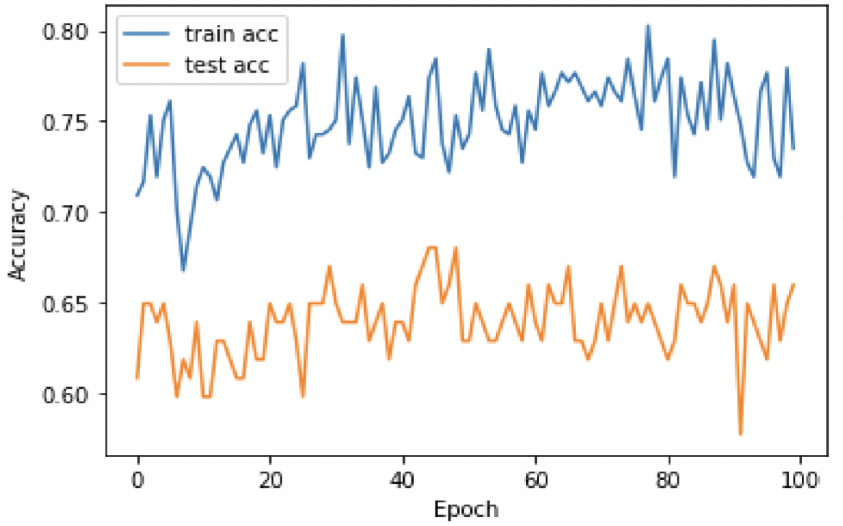

Figure 12. Accuracy vs. Epoch

The performance of the model across all classes is gauged by accuracy. When all classes are given the same weight, it is helpful. It is determined, as shown in Fig. 12 is obtained by dividing the total number of forecasts by the number of reliable ones.

The result of dividing the total number of values in the matrix by the sum of True Positives and True Negatives is saved in the variable accuracy based on the previously computed loss graph (equation 6):

$$Accuracy = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \tag{6}$$

The result, 0.6597, specifies that the model properly predicted the outcome 65.97% of the time.

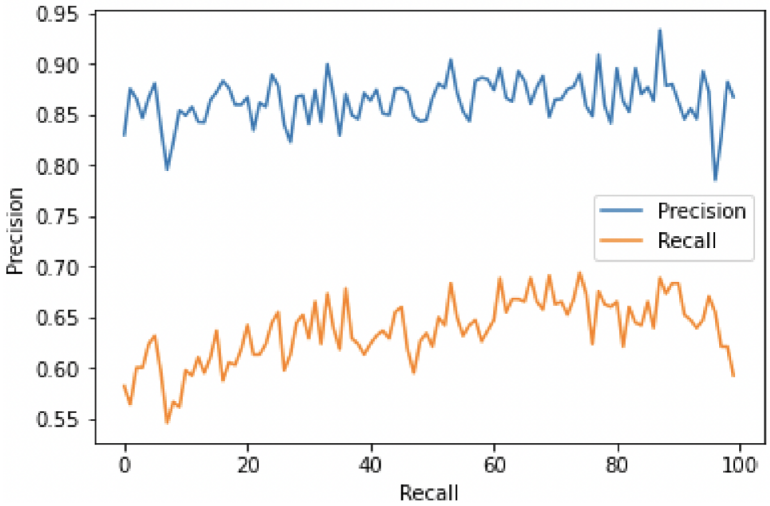

Figure 13. Precision vs. Recall

The comparison of the model's recall and precision is shown in Figure 13. It represents a constant value over multiple iterations, which also states that the model is quite stable on negative values. It also represents a scope for improvement of the model as the value is quite constant over a long period of epochs/iterations.

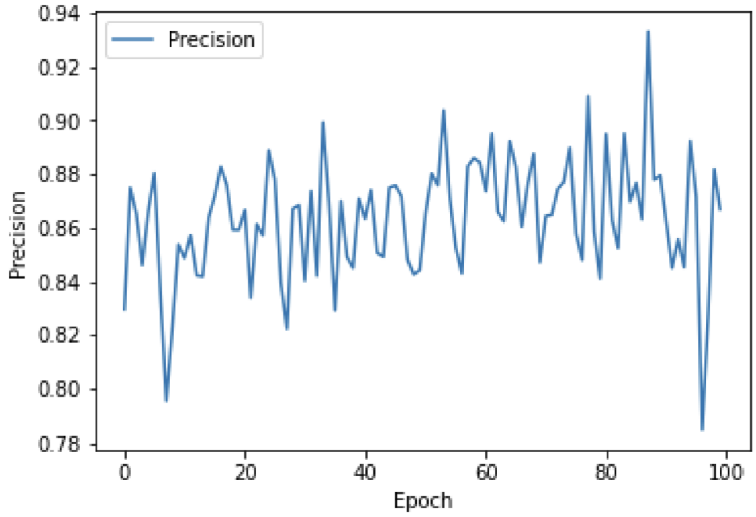

Figure 14. Precision vs. Epoch

It is measured by the ratio of correctly classified positive samples to all correctly classified samples. As the number of epochs rises, accuracy rises. However, the precision is constant in this graph Fig. 14, which shows that the model is not good and could be further improved.

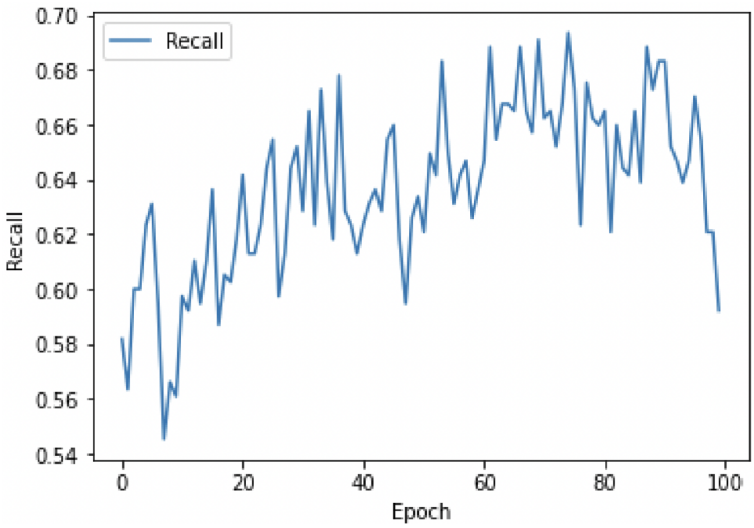

Figure 15. Recall vs. Epoch

The recall is determined as the proportion of all correctly labeled Positive samples to all Positive samples.

- Recall measures how accurately the model can differentiate Positive samples.
- As the quantity of positive samples rises, the recall rises as well.
- Only how the positive samples are categorized will affect memory.

For example, this has nothing to do with how the negative samples are accurately classified. Even if all of the negative samples were mistakenly classified as Positive, the recall will be 100% if the model incorrectly labels all of the positive samples as Positive. The curve in Fig. 15 displays a moderate range of recall.

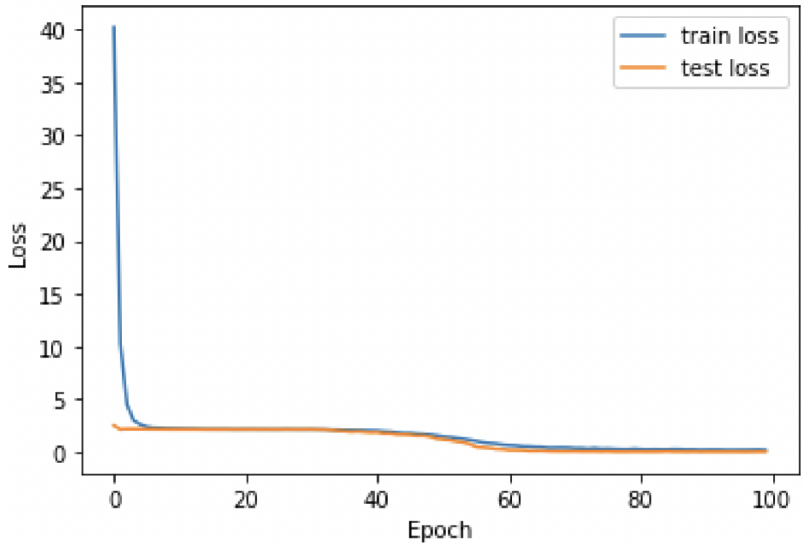

Figure 16. Loss vs. Epoch

For the TESS dataset [22], we got 99.1% accuracy with a 3.07% Loss. A good match is demonstrated in Fig. 16 by a training and validation loss that decreases to the point of stability with a

minor difference between the two final loss values. On the training dataset more often than not, the model's loss is less than on the validation dataset. As a result, there will probably be a difference between the validation loss learning curve and the train learning curve.

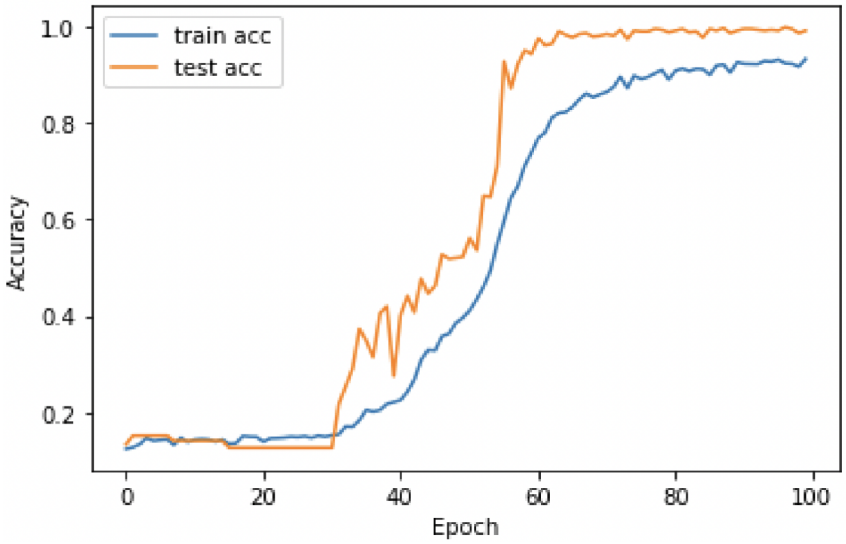

Figure 17. Accuracy vs. Epoch

Since the dataset's accuracy trends have been rising over the previous few epochs, you can observe from the accuracy plot in Fig. 17 that the model should be trained a bit more. Additionally, the model's competence on both datasets is equivalent, indicating that it has not yet overlearned the training dataset.

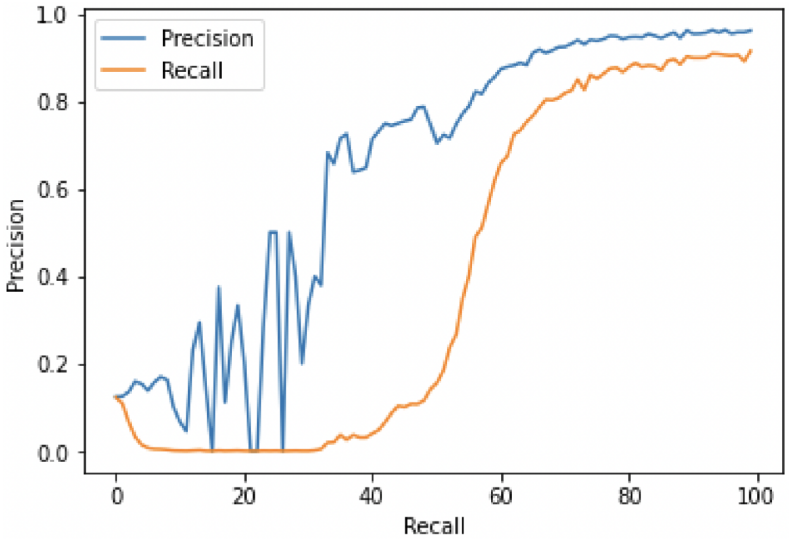

Figure 18. Precision vs. Recall

When there is a moderate to significant class imbalance, precision-recall curves should be applied. The accuracy and recall for thresholds can be calculated using the precision-recall curve (PRC) function, which accepts as input the true output values and the probabilities for the positive class and outputs the precision, recall, and threshold values. Precision-recall curves usually have zigzag patterns

with many ups and downs. Therefore, compared to ROC curves, precision-recall curves tend to cross each other considerably more frequently. This can make comparing different curves difficult. The performance level of curves near the PRC for a perfect test is higher than that of curves close to the baseline. In other words, as seen in Fig. 18, a curve above another curve has a higher performance level.

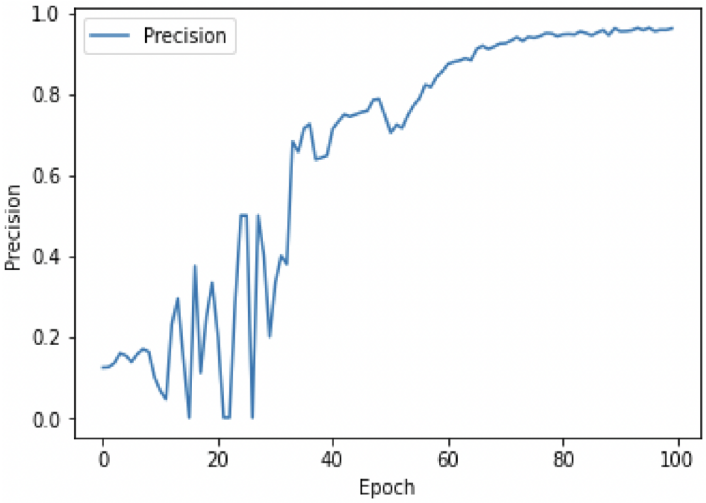

Figure 19. Precision vs. Epoch

Precision's goal is to prevent wrongly recognizing a negative sample as positive and to correctly classify all positive samples as positive. Fig. 19 shows that the accuracy increases if all Positive samples are correctly identified, but some Negative samples are wrongly classified, making this an acceptable model for the dataset.

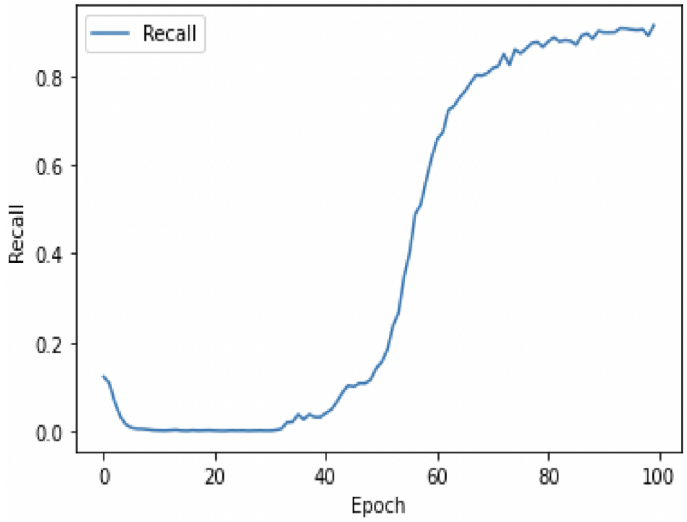

Figure 20. Recall vs. Epoch

A high recall means that the model can correctly classify each positive sample as Positive. We can therefore put our faith in the model's capacity to recognize positive samples. The model found every positive sample. There may still be many negative samples that are categorized as positive since

the recall ignores how the negative samples are categorized (i.e., a high False Positive rate). This is not considered in the recall. The recall value between 0.0 and 1.0 in Fig. 20 represents the proportion of positive samples that the model correctly identified as positive. For the BAVED dataset [24], we received an accuracy of %73.13 with a loss of 73.23%.

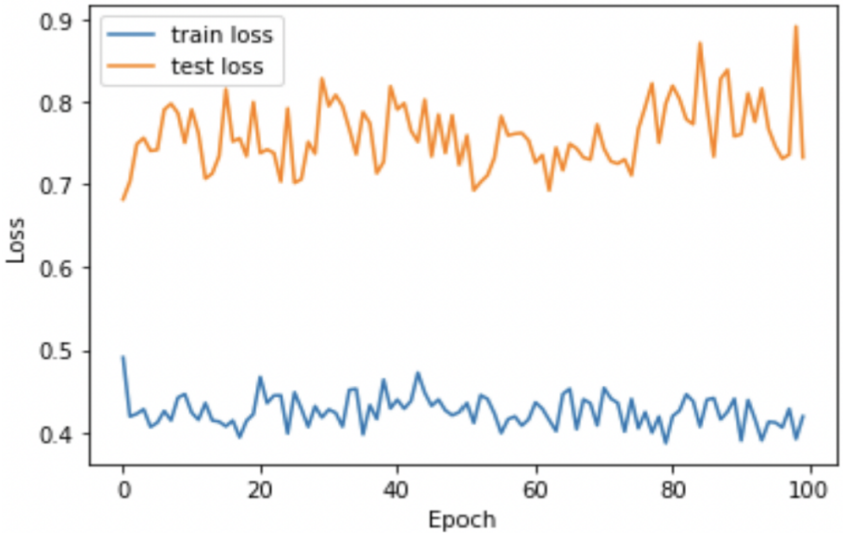

Figure 21. Loss vs. Epoch

It could be located using a validation loss that is lower than the training loss. In this instance, it suggests that the validation dataset might be simpler to forecast than the training dataset for the model, as seen in Fig. 21. It is an unrepresentative validation dataset, which means that not enough information is provided to assess the generalizability of the model.

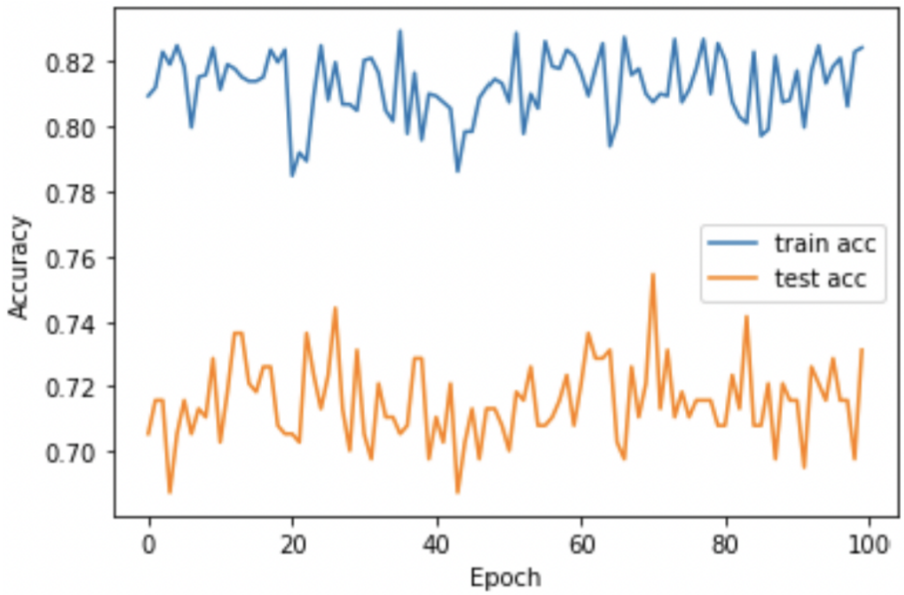

Figure 22. Accuracy vs. Epoch

As seen in Fig. 22, the accuracy increases significantly in the first two epochs, indicating that the network is rapidly assimilating new data. The curve then flattens, showing that fewer training iterations

are required to finish developing the model. Overfitting is frequently observed when the accuracy of the training data ("acc") increases but the accuracy of the validation data ("Val acc") decreases. In the end, it demonstrates that the model has started to learn the data.

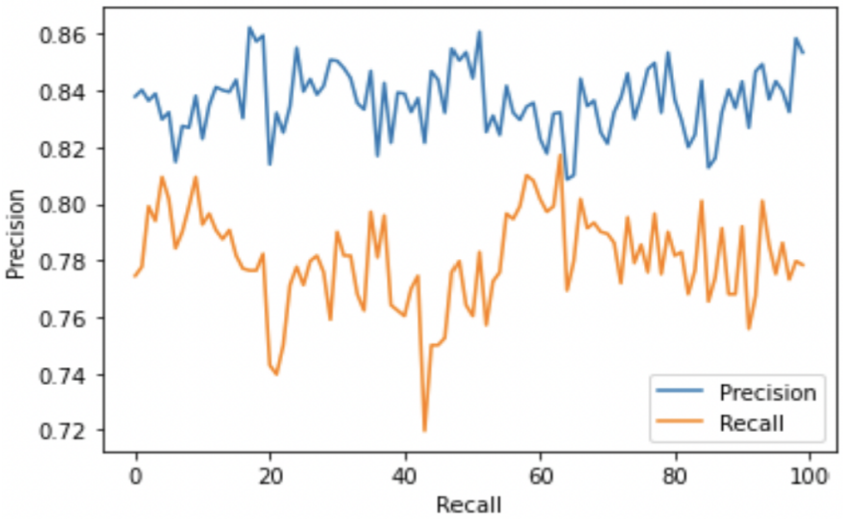

Figure 23. Precision vs. Epoch

The model is just over the no-skill line for most thresholds, as seen by the plot of the precision-recall curve in Fig. 23. This is feasible since the model generates probability but has some case-specific uncertainty. These are shown by the various thresholds assessed throughout the curve's creation, switching some class 0 data to class 1 and providing some accuracy but extremely poor recall.

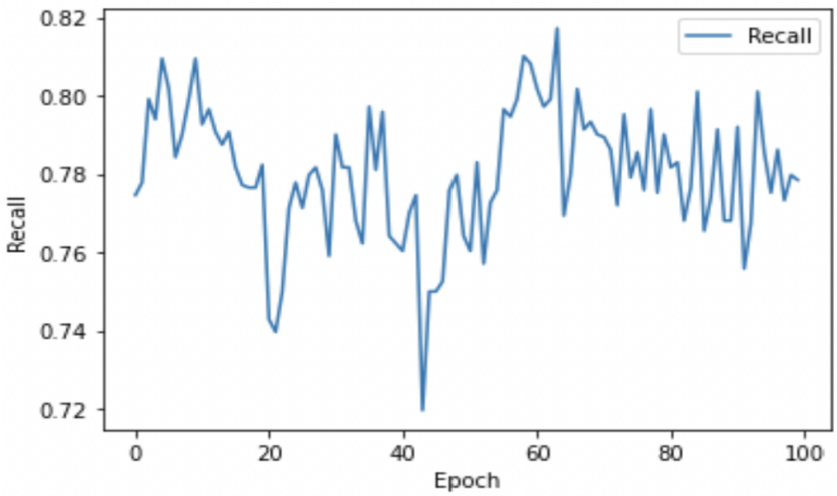

Figure 24. Recall vs. Epoch

When a model's recall is high, but its accuracy is low, it properly detects the majority of positive data and generates many false positives. A model is accurate if it correctly identifies a positive sample, however, if it has high accuracy but poor recall, it can only correctly identify a small percentage of positive samples. According to the same reasoning as earlier, fewer negative predictions explain why

there are fewer incorrect negative predictions overall and why recall is better. Thus, the recall in Fig. 24 ranges from 0.72 to 0.82.

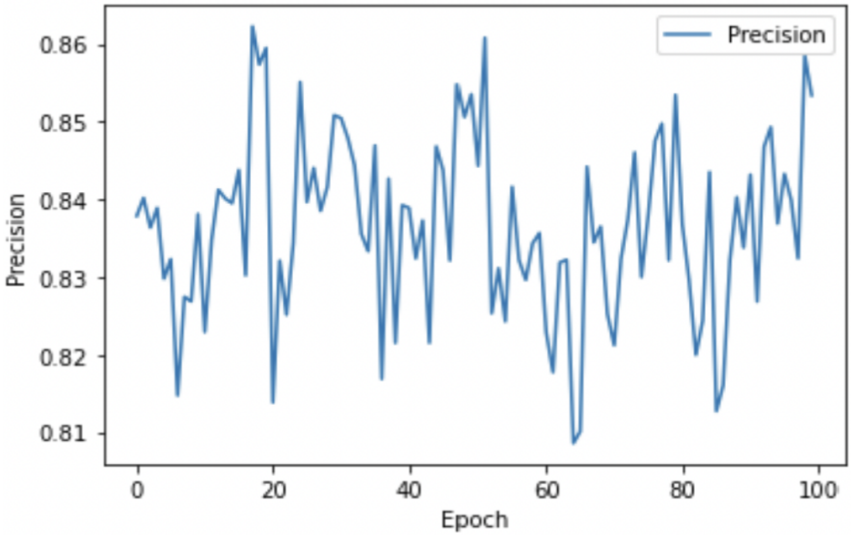

Figure 25. Precision vs. Epoch

As seen in Fig. 25, we can see that the precision ranges between 0.81 and 0.86. When the threshold is raised, it rises. We also point out that if the threshold is set high enough, accuracy may be arbitrarily excellent. The model is thus somewhat effective for this collection of data. Table 2 presents the proposed work compared with the existing works.

Table 2 Comparative results of the accuracy

Dataset	[21]	[22]	[23]	[24]
Accuracy of Existing Systems	83.4% [10]	84.96% [25]	No experiment done with only audio	89% [26]
Accuracy of the Proposed system	96.24%	99.10%	65.97%	73.12%

5. Conclusion

The development of human behavioral informatics and the design of successful human-machine interaction systems depend on accurate emotion recognition systems. Such solutions make it easier to communicate naturally and analyze data on human behavior in a trustworthy and effective manner. In this paper, we presented a deep learning-based AER mechanism for SER. The suggested architecture is built on the two alternative speech representations, which are the MFCC of the audio signal. We achieve cutting-edge performance in the URDU, TESS, AESDD (only audio), and BAVED datasets by simultaneously training these features on temporal space. Additionally, we presume that our study does not have access to speaker-identifying data. Per-sample accuracy is a quantitative indicator of

speech emotion recognition. In addition, there are additional terms like balanced accuracy, precision, and recall. The multimodal approach outperforms the audio model alone in terms of performance and can forecast challenging instances like Fear, Sadness, and Disgust in the dataset. All the findings demonstrate that using attention mechanisms and combining different models, as opposed to using unimodal ones, significantly increases success. While the unimodal system is nearly impossible to recognize, the multimodal system can predict hard-class emotion, which only makes up a very small portion of the imbalanced dataset.

Certainly, demonstrating the application of the work in the real world is crucial to showcase the practical utility of AER using DNNs. One suitable example of the application of Automatic Emotion Recognition using DNNsis in the field of mental health. The system can be used to monitor the emotions of individuals with mental health conditions such as depression, anxiety, or bipolar disorder. For instance, the system can be integrated into a mobile app that patients can use to record their speech, which is then processed by the DNN model to determine their emotional state. If the system detects signs of depression or anxiety, it can alert mental health professionals, who can then provide appropriate care and support. Another example of the application of Automatic Emotion Recognition using DNNs is in the field of human-computer interaction. The system can be used to improve the interaction between humans and machines. For instance, the system can be integrated into a virtual assistant that can recognize the emotional state of the user and respond accordingly. If the user is feeling sad, the virtual assistant can provide comforting words or suggest activities that can improve their mood. Our model with a much higher accuracy rate would help to address a lot of other countries with different languages other than English.

DNNs have indeed been widely used for automatic emotion recognition, and the use of DNNs alone may not be considered a highly original contribution. However, the novelty of the proposed method may not necessarily lie in the choice of algorithm, but rather in the specific architecture, training techniques, and features used in the DNN. This paper provides a detailed explanation of these aspects and how they differ from previous approaches. For example, the paper proposes a novel architecture that improves the accuracy and efficiency of emotion recognition, and it may use a novel training technique that makes the DNN more robust to noise and variations in the data. These details should be explicitly stated in the paper to demonstrate the innovation of the proposed method. Additionally, the paper also provides a thorough evaluation of the proposed method on various datasets and compares its performance to existing methods in the field. This will provide further evidence of the novelty and effectiveness of the proposed method. In conclusion, the novelty of a method in AER may not solely depend on the choice of algorithm, but rather on the specific implementation details and evaluation results. This paper clearly explains these details to demonstrate the innovation of the proposed method.

Declarations

Competing Interests: The authors declare that there is no conflict of interest.

Funding: No funding has been received for this work.

References

[1] Er, M. B. (2020). A novel approach for classification of speech emotions based on deep and acoustic features. *IEEE Access*, *8*, 221640-221653.

[2] Zvarevashe, K., & Olugbara, O. (2020). Ensemble learning of hybrid acoustic features for speech emotion recognition. *Algorithms*, *13*(3), 70.

[3] Hesam Sagha, Pavel Matejka, et al., Enhancing multilingual recognition of emotion in speech by language identification, In 17th *Annual Conference of the International Speech Communication Association* (Interspeech 2016), pp. 2949-2953.

[4] Bo-Chang Chiou and Chia-Ping Chen, Speech emotion recognition with cross-lingual databases, In 15th *Annual Conference of the International Speech Communication Association* (Interspeech 2014), pp. 558–561.

[5] Je Hun Jeon, Duc Le, et al., A preliminary study of cross-lingual emotion recognition from speech: automatic classification versus human perception, In 14th *Annual Conference of the International Speech Communication Association* (Interspeech 2013), pp. 2837–2840.

[6] Neumann, M. (2018, April). Cross-lingual and multilingual speech emotion recognition on english and french. In *2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 5769-5773). IEEE.

[7] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, (2010). Cross-corpus acoustic emotion recognition: Variances and strategies, *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131.

[8] Wikipedia contributors. (2022, August 14). *Urdu*. Wikipedia. Retrieved August 24, 2022, from https://en.wikipedia.org/wiki/Urdu

[9] Wu, C. H., Lin, J. C., & Wei, W. L. (2014). Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA Transactions on Signal and Information Processing*, 3.

[10] Latif, S., Qayyum, A., Usman, M., & Qadir, J. (2018, December). Cross lingual speech emotion recognition: Urdu vs. western languages. In 2018 *International Conference on Frontiers of Information Technology* (FIT) (pp. 88-93). IEEE.

[11] Zvarevashe, K., & Olugbara, O. (2020). Ensemble learning of hybrid acoustic features for speech emotion recognition. *Algorithms*, vol. 13, no. 3, 70.

[12] Aspandi, D., Sukno, F., Schuller, B., & Binefa, X. (2021). An enhanced adversarial network with combined latent features for spatio-temporal facial affect estimation in the wild. arXiv preprint arXiv:2102.09150.

[13] Zhang, Z., Xu, S., Cao, S., & Zhang, S. (2018, November). Deep convolutional neural network with mixup for environmental sound classification. In *Chinese Conference on Pattern Recognition and Computer Vision* (prcv) (pp. 356-367). Springer, Cham.

[14] Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., & Wendemuth, A. (2009, November). Acoustic emotion recognition: A benchmark comparison of performances. In 2009 IEEE Workshop on Automatic Speech Recognition & Understanding (pp. 552-557). IEEE.

[15] Mohamed, O., & Aly, S. A. (2021). Arabic Speech Emotion Recognition Employing Wav2vec2. 0 and HuBERT Based on BAVED Dataset. arXiv preprint arXiv:2110.04425.

[16] Alnuaim, A. A., Zakariah, M., Shukla, P. K., Alhadlaq, A., Hatamleh, W. A., Tarazi, H., ... & Ratna, R. (2022). Human-Computer Interaction for Recognizing Speech Emotions Using Multilayer Perceptron Classifier. *Journal of Healthcare Engineering*, vol. 2022, 6005446.

[17] Senthilkumar, N., Karpakam, S., Devi, M. G., Balakumaresan, R., & Dhilipkumar, P. (2022). Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks. *Materials Today: Proceedings*, vol. 57, pp. 2180-2184.

[18] Li, L. Q., Xie, K., Guo, X. L., Wen, C., & He, J. B. (2022). Emotion recognition from speech with StarGAN and Dense-DCNN. *IET Signal Processing*, vol. 16, no. 1, pp. 62-79.

[19] Andayani, F., Theng, L. B., Tsun, M. T., & Chua, C. (2022). Hybrid LSTM-Transformer Model for Emotion Recognition from Speech Audio Files. *IEEE Access*, vol. 10, pp. 36018-36027.

[20] *Sound*. (2022, June 10). Science World. Retrieved August 24, 2022, from https://www.scienceworld.ca/resource/sound/

[21] *Urdu Emotion Dataset*. (2021, October 7). Kaggle. Retrieved August 24, 2022, from https://www.kaggle.com/datasets/kingabzpro/urdu-emotion-dataset

[22] *Toronto Emotional Speech Set (TESS)*. (2019, August 25). Kaggle. Retrieved August 24, 2022, from https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess

[23] *kingabzpro/Acted-Emotional-Speech-Dynamic-Database*. (n.d.). DAGsHub. Retrieved August 24, 2022, from https://dagshub.com/kingabzpro/Acted-Emotional-Speech-Dynamic-Database

[24] (n.d.). *GitHub - 40uf411/Basic-Arabic-Vocal-Emotions-Dataset: Basic Arabic Vocal Emotions Dataset (BAVED) is a datasetthat contains an arabic words spelled in diffrent levels of emotions recorded in an audio/wav format.* GitHub. Retrieved August 24, 2022, from https://github.com/40uf411/Basic-Arabic-Vocal-Emotions-Dataset

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., (2011). cikit-learn: Machine learning in python, *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830.

[26] Mohamed, O., & Aly, S. A. (2021). Arabic Speech Emotion Recognition Employing Wav2vec2. 0 and HuBERT Based on BAVED Dataset. arXiv preprint arXiv:2110.04425.

[27] Pawar, M. D., & Kokate, R. D. (2021). Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients. Multimedia Tools and Applications, 80(10), 15563-15587.

[28] Jahangir, R., Teh, Y. W., Hanif, F., & Mujtaba, G. (2021). Deep learning approaches for speech emotion recognition: State of the art and research challenges. Multimedia Tools and Applications, 80(16), 23745-23812.

[29] Valstar, M. F., & Pantic, M. (2010). Induced disgust, happiness and surprise: An addition to the MMI facial expression database. In Proceedings of the 3rd International Workshop on EMOTION (pp. 65-72).

[30] Kaliouby, R. E., & Robinson, P. (2005). Real-time inference of complex mental states from facial expressions and head gestures. In Proceedings of the 7th international conference on Multimodal interfaces (pp. 1-8).

[31] Zhao, L., Zhao, Y., Zhang, J., & Zhang, W. (2018). Emotion recognition from EEG signals using deep learning with kernel methods. IEEE Transactions on Affective Computing, 9(1), 94-105.

[32] Liu, F., Shen, H., Shen, Y., & Cui, L. (2020). A survey on deep learning-based emotion recognition: Toward multimodal fusion. IEEE Transactions on Affective Computing, 1-1.

[33] Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., ... & Patras, I. (2012). DEAP: A database for emotion analysis; using physiological signals. IEEE Transactions on Affective Computing, 3(1), 18-31.

[34] Wang, Y., Zhang, Y., Ji, Q., & Zhang, B. (2020). Emotion recognition from physiological signals using a multimodal deep belief network. IEEE Transactions on Affective Computing, 11(2), 178-191.

[35] Singh, P., Sahidullah, M., & Saha, G. (2023). Modulation spectral features for speech emotion recognition using deep neural networks. Speech Communication, 146, 53-69.

[36] Bashir, M. F., Javed, A. R., Arshad, M. U., Gadekallu, T. R., Shahzad, W., & Beg, M. O. (2022). Context aware emotion detection from low resource urdu language using deep neural network. Transactions on Asian and Low-Resource Language Information Processing.

[37] Maheshwari, D., Ghosh, S. K., Tripathy, R. K., Sharma, M., & Acharya, U. R. (2021). Automated accurate emotion recognition system using rhythm-specific deep convolutional neural network technique with multi-channel EEG signals. Computers in Biology and Medicine, 134, 104428.

[38] Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., & Chandran, V. (2020). Automatic emotion recognition using temporal multimodal deep learning. IEEE Access, 8, 225463-225474.

[39] Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (FERC). SN Applied Sciences, 2(3), 446.