2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE)

A Robust Speech Emotion Detection Mechanism Using Supervised Deep Learning Paradigms

1st Divya Saini
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology, Symbiosis International (Deemed University)
Lavale, Pune, Maharashtra, India
divyasaini197@gmail.com

2nd Kailash Shaw
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology, Symbiosis International (Deemed University)
Lavale, Pune, Maharashtra, India
kailash.shaw@gmail.com

2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE) | 979-8-3503-1646-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIKIIE60097.2023.10390312

***Abstract*—The research applies deep learning to SER, or voice recording emotion detection. Precision vocal emotion recognition has several applications, including human-computer interaction, virtual assistants, and healthcare. This study uses emotional-labeled spoken utterances to build an accurate SER system used to train the deep learning models like convolutional neural networks (CNNs). These models are popular for speech emotion recognition because they can learn complex voice signal patterns that indicate different moods. Accurate diagnosis involves more detailed sound analysis and mood or emotion recognition. This paper presents a comprehensive framework for SER from recorded audio samples using digital signal processing advances. The dataset's speech features spectrograms and pitch picture train the models. Speech analysis uses these features because they capture vocal tract and pitch aspects of the speech stream. After training, the models' classification accuracy their ability to correctly recognize unseen speech samples' emotional content is examined. To assess its performance, the best model is compared to the most advanced methods. Vgg16 CNN outperformed Mel-Spectrogram-featured Convolutional Neural Networks in this work. Emotion sound samples processed with CNN and mel-spectrogram achieved 89% accuracy, with better results using transfer learning (CNN-VGG16). Other classifiers like SVM, Logistic Regression, Decision Tree, and Random Forest yielded lower accuracy (60%-75%). Further research should explore composite feature sets for improved classification.**

Keywords—SER, Signal Preprocessing, Signal Transformation, Feature Extraction, CNN, VGG16

I. Introduction

In recent years, SER has attracted considerable interest due to its potential applications in numerous disciplines, including healthcare, entertainment, and human-computer interaction. Due to the complexity of emotions and the diversity of speech signals, emotion recognition from speech is a difficult endeavour. Emotions are conveyed through a variety of acoustic cues, including pitch, intensity, and duration, which are difficult to accurately extract and classify [1]. Using a combination of signal processing techniques and deep learning models, we present in this paper a method for SER. Our strategy concentrates on identifying seven distinct emotions, including happiness, sadness, neutrality, fear, disgust, surprise, and calmness. In order to accomplish this, we combined two distinct datasets, RAVDEES and Crema-D, to produce a large and diverse dataset. The audio signals were then pre-processed by sending them through FFT and DWT and combining their respective outputs. We then applied the Hilbert transform and generated spectrograms, which were saved for further analysis in a folder. To classify the emotions from the spectrograms, we employed the VGG16- CNN model, which is a well-known and potent deep learning model for image classification. In addition, we conducted experiments with additional machine learning models, including SVM, Decision Trees, and Logical Regression. Our method yielded a high rate of accuracy, and this paper provides detailed results and analysis.

The remaining sections are organized as follows. The second section provides a summary of related research in the field of SER with the pre-processing techniques, and feature extraction methods utilized by our methodology. The third section describes proposed architecture. The experimental results and analysis are presented in Section 4, while Section 5 deals with contribution made followed by conclusion in section 6.

II. Materials and Methods

SER has been an active research topic within the field of affective computing for a number of years [2]. Significant progress has been made in recent years in the development of robust and reliable SER systems. This section will examine some of the most significant contributions to this discipline [3-4] . A DNN was utilized in one investigation, and the results showed an accuracy of 85.5%. The model was trained on MFCCs, which are quite popular in speech processing and were employed during the training procedure [6]. In another investigation, a CNN was used, and the researchers were able to attain an accuracy of 81.1%. The model was educated using MFCCs in addition to pitch, which is a measurement of the frequency of the voice [7]. A neural network with LSTM, which was used in the third trial, was able to reach an accuracy of 81.5%. The model was trained on MFCCs as well as prosody characteristics, which are aspects of speech that are associated with its rhythm and melody [15]. A combined CNN and LSTM approach was applied in the fourth trial, which resulted in an accuracy rate of 87.5%. In order to train the model, MFCCs and prosody features, in addition to phoneme embedding's, which capture the relationships between phonemes in speech, were used [18]. In conclusion, the RAVDESS dataset has been extensively used for research on the recognition of emotions in spoken language, and various models have reached excellent levels of accuracy by combining MFCCs, prosody variables, and other speech-related characteristics. The DNN, CNN, LSTM, and hybrid CNN-LSTM models are included in these types of models.

Many researchers have found great accuracy rates when using this dataset, and the Crema-D dataset has been utilized extensively for SER [19-20]. The following is a list of some of the more prominent studies that have made use of the Crema-D dataset: [22] SER using CNN from ImageNet achieved an accuracy of 75.1% by utilizing VGG16 and ResNet50 for feature extraction and SVM as a classifier. [21] utilized a combination of deep neural networks and decision trees, and as a result, they were able to achieve an accuracy of

72.2%. An accuracy of up to 78.3% was achieved through analyzing the performance of several different CNN architectures, such as AlexNet, VGG16, and ResNet50 [23]. Overall, the Crema-D dataset has been shown to be a useful resource for study on SER, and the works that were discussed above show that high accuracy rates may be attained by utilizing a variety of models and methodologies

A. Methodology

This section deals with methodologies utilized in our SER approach. We begin by describing the dataset used in our study, including the data sources and the audio signal pre-processing processes. Then, we address the feature extraction methods used to extract meaningful information from a speech signal, such as the use of FFT and DWT to capture frequency and temporal information, respectively. Also covered is the application of the Hilbert transform to generate spectrograms, which are then utilized as inputs for classification models. Next, we describe the deep learning and machine learning models used for classification, including the VGG16 CNN model, which is a popular and potent image classification deep learning model. In addition, we conducted experiments with additional machine learning models, including SVM, Decision Trees, and Logical Regression. Each model's hyper parameters and training and testing procedures are described in detail. Finally, we describe the evaluation metrics used to assess the efficacy of our methodology, which include accuracy, precision, recall, and F1-score. In addition, we discuss the limitations of our methodology and offer suggestions for future research. This section provides a comprehensive overview of the methods employed in our SER approach, including the dataset, feature extraction techniques, classification models, and evaluation metrics.

B. Signals and SER

A microphone records an audio signal as a time-varying sound wave. Each sample represents the sound wave's amplitude at a certain instant. The frequency, strength, and duration of an audio transmission reveal its source [24]. Speech Emotion Recognition relies on the audio signal's acoustic characteristics, which reflect the speaker's mood. Prosodic elements like pitch, loudness, and speaking rate can be retrieved from the audio data, as can spectral features like Mel-spectrograms, which capture signal frequency [25]. Machine learning and deep learning models can classify speaker emotions using these features. Speech Emotion Recognition faces the considerable diversity in speech signal acoustics across speakers and emotional states. Filtering, segmentation, and normalization are employed to reduce these effects and improve emotion recognition. Speech Emotion Recognition relies on audio signal processing, which could affect healthcare and entertainment [26]. Fig.1 depicts the raw audio signal in a visualized form.

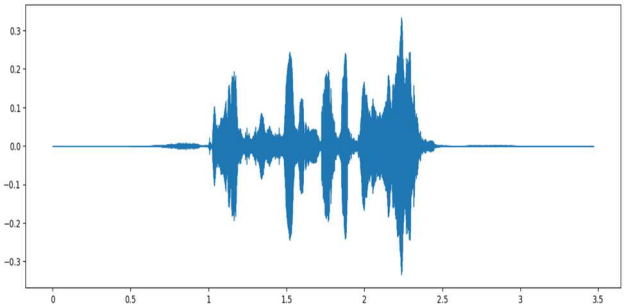

Fig. 1. Raw WAV signal

C. Signal Transformation using FFT and DWT

The process of decomposing a composite audio signal into a sum of sine and cosine waves, which may then be used to rebuild the original signal. The FFT and DFT is used to to derive the frequency domain representation of the resampled time-domain signal. This can be accomplished by resampling the original signal [27]. The DFT is a mathematical process that converts a sequence of N complex numbers $x_0, x_1, \ldots, x_n$ into a new sequence of N complex numbers in the frequency domain.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-2\pi i k n / N} \text{ for } \leq k \leq N-1 \quad (1)$$

In this context, the values of a signal, represented by x_i, are considered to be sampled at evenly spaced time intervals, with t taking on values of $0,1, \ldots, N-1$. The output, Xk, is a complex number that encodes information about the amplitude and phase of a sinusoidal wave with a frequency of (k / N) cycles per time unit. These waves can be combined in a linear fashion to produce an approximation of the original signal, and the coefficients for this approximation can be found by doing so. Because each wave repeats itself an integer number of times throughout the course of N time units, the approximation that is produced will be periodic with a period equal to N [27]. The FFT reduces the computational complexity of a problem size of N from $O(N2)$ to $O(Nlog(N))$. For a large enough N, it generates a significant gain in the computational times. The FFT Signal for Original Signal is shown in fig 2.

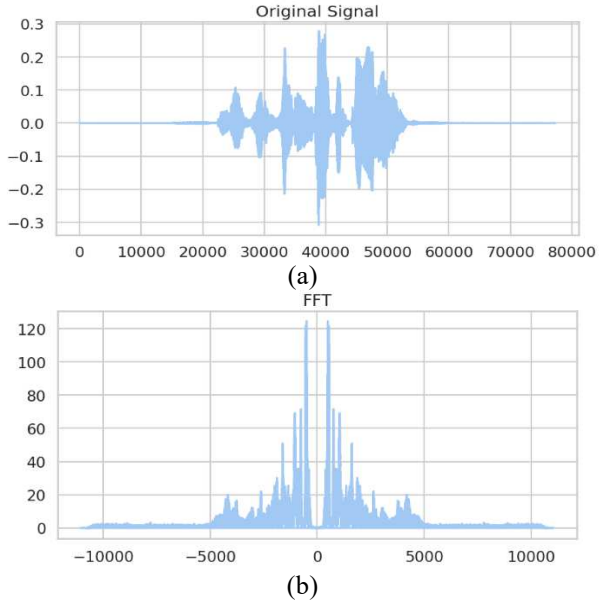

Fig. 2. (a) Original raw wav (b) FFT WAV

Discrete wavelet transform: The DWT splits a signal into approximation (CA) and detail (CD) coefficients. The detail coefficients represent high-frequency signals, while the approximation coefficients represent low-frequency signals. By breaking an audio signal into frequency sub bands at different resolutions, the DWT can analyse its frequency content [28]. The CA and CD coefficients can be utilised in audio preprocessing to extract information at different frequency scales for machine learning and deep learning emotion identification models. The CA and CD coefficients can be used to extract signal energy in low- and high-frequency sub bands, respectively. By threshold detail

coefficients and reconstructing the signal with the original approximation coefficients, the DWT may denoise an audio signal. This is useful in Speech Emotion Recognition, where noisy signals might impair model accuracy. DWT may provide frequency content information at multiple scales, unlike FFT [28]. Speech Emotion Recognition, where distinct emotions are connected with different frequency sub bands, can benefit from this. Fig. 3 depicts DWT WAV signal achieved from Original Signal.

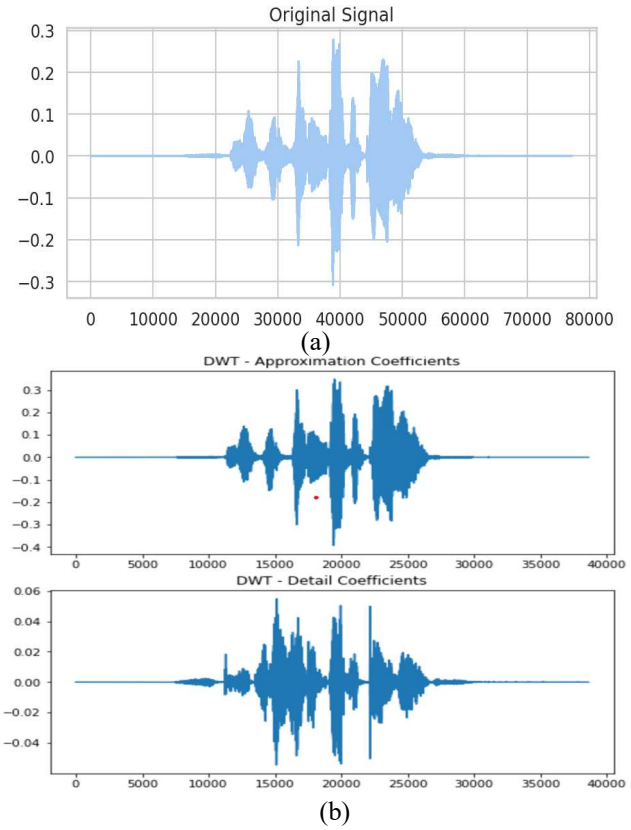

Fig. 3. (a) Original raw wav (b) DWT WAV

D. Hilbert transform

The Hilbert Transform is a mathematical technique used to analyses the analytic signal of a signal with real-valued coefficients. In the context of audio processing, the Hilbert Transform can be used to derive an audio signal's amplitude envelope, a measure of the signal's energy over time [29]. The mathematical definition of the Hilbert Transform of a signal x(t) is as follows (2):

$$H(x(t)) = P.V.(1/\pi) \int (x(\tau)/(t-\tau))d\tau \quad (2)$$

$$x(t) = Re(z(t)) \quad (3)$$

$$a(t) = |z(t)| \quad (4)$$

P.V. stands for the Cauchy principal value. The Hilbert Transform generates a complex-valued signal with the following relationship to the original signal (3). where Re represents the real component and $z(t)$ represent the analytic signal of $x(t)$. The analytic signal's amplitude envelope can be computed as (4)

In audio pre-processing, the Hilbert Transform can be used to extract an audio signal's amplitude envelope, which is a measure of the signal's energy over time. This can be useful for feature extraction in SER because it provides information about the overall energy of the signal and can aid in differentiating between emotional states. The Hilbert Transform is a mathematical technique that can be used to extract an audio signal's amplitude envelope [30]. Its capacity to analyses the analytic signal of a real-valued signal makes it a crucial component of audio pre-processing for Speech Emotion Recognition. Real and Imaginary part of Hilbert Transform is shown in fig. 4.

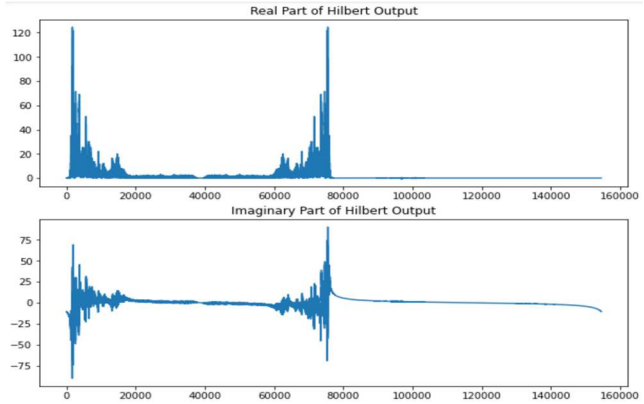

Fig. 4. Hilbert transform

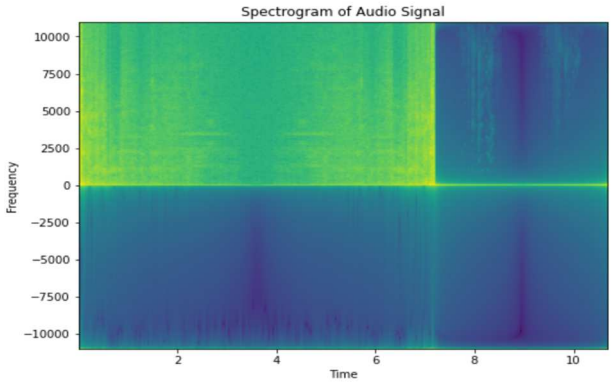

Fig. 5. Spectrogram

E. Spectrogram

Spectrograms display the frequency content of an audio signal as it varies over time. Fig. 5 depicts the spectrogram for input signal obtained from Hilbert transform. STFT is a technique that entails dividing a signal into short, overlapping windows and computing the Fourier Transform for each window [31]. The STFT of a signal x[n] can be calculated mathematically as follows:

$$STFT\,[n,k] = \sum_{m=-M}^{M} x[n+m] * w[m] * exp^{(-j*2mk/N)} \quad (5)$$

Where, $STFT[n,k]$is the kth frequency component of the STFT at time index $n, w[m]$ is a tapered window function such as a Hamming or Hanning window, and M is the window length. The STFT generates a time-frequency representation of the signal, where each point in the time-frequency plane represents the amplitude of a specific frequency component at a specific time [32]. Calculating the magnitude of the STFT and plotting it as a function of time and frequency produces spectrograms. This results in a visualisation of the time-varying frequency content of the signal, which can be used to

extract features such as the MFCCs, which are typically input to machine learning and deep learning models for Speech Emotion Recognition [33]. Overall, spectrograms are a useful instrument for displaying the frequency content of an audio signal as it varies over time. Their ability to extract characteristics such as MFCCs makes them a crucial component of audio pre-processing for SER.

F. Classifiers

A structured neural network with successively many layers is known as a CNN. A SoftMax unit, numerous convolution layers, pooling layers, and fully linked layers make up the conventional CNN design. Using a feature extraction technique, this linear network generates an abstract representation of the input. CNN is mostly used for categorization analysis of images and data. CNN is built on convolutional layers, which serve as filters. Convolutional processing is done in these layers, and the pooling layer receives the results. These layers carry out convolutional processing before sending the data to the pooling layer [34]. The main objective of the pooling layer is to decrease the output resolution of the convolutional layers and, consequently, the computational load. The idea of a multiclass universe is expanded by a fully linked layer that receives the output, flattens the data there, and then classes it using the SoftMax unit [35]. The Spectrograms produced from the voice samples were the CNN input. Fig. 6 shows the CNN architecture composed of three fully connected layers, three convolutional layers, and a SoftMax unit for classification made up the model.

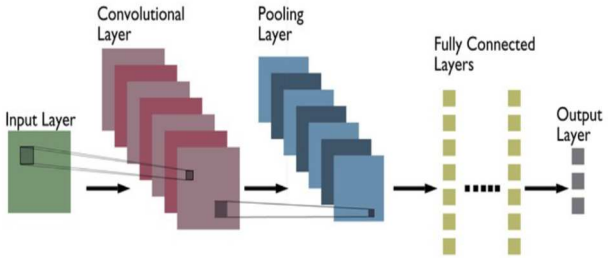

Fig. 6. CNN Architecture

The VGG16 architecture is comprised of sixteen layers, including thirteen convolutional layers and three fully connected layers. Detailed Architecture of VGG16 is shown in fig. 7. Each convolutional layer employs a 3x3 filter size that is repeated multiple times, and the number of filters increases as we progress deeper into the network. After each set of convolutional layers, the max pooling operation is used to down sample the feature maps and enhance their translational invariance [36]. The use of very deep convolutional layers is one of the distinguishing characteristics of the VGG16 architecture, which enables it to recognise complex patterns and structures in input images. In addition, the use of small filter sizes and maximum pooling operations reduces the number of network parameters, making the network simpler to train and less susceptible to overfitting [36]. In recent years, the VGG16 architecture has been extensively implemented in numerous computer vision applications, such as object detection, image segmentation, and style transfer. The VGG16 architecture can be used for image classification of spectrograms, which have a similar structure to images, in the context of SER [37]. It is possible to obtain high accuracy in SER tasks by fine-tuning the pre-trained VGG16 model on a dataset of spectrograms. Overall, VGG16 is a powerful CNN architecture that has attained state-of-the-art image classification performance. It is a suitable model for Speech Emotion Recognition tasks involving image classification of spectrograms due to its ability to detect complex patterns and structures in input images.

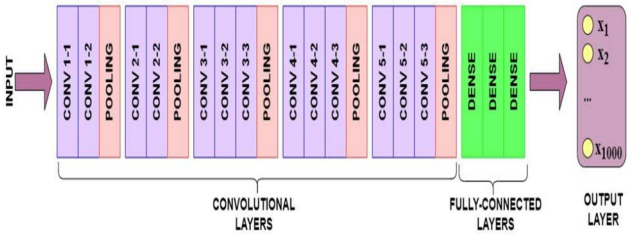

Fig. 7. VGG16 Architecture

III. Proposed Architecture

In this research, we provide a unique architecture for SER, which integrates a number of different methods for audio preprocessing and image classification. SER is an acronym that stands for Speech Expression Recognition. Our suggested architecture is comprised of a number of processes, the most important of which are the combining of two datasets, the application of the FFT and the DWT, the Hilbert Transform, the generation of spectrograms, and the use of a pre-trained VGG16 model for picture classification. We started by combining two datasets, known as Ravdess and CREMA-D, in order to produce a larger dataset that included a wider variety of feelings. The audio signals were then subjected to the Fourier Transform and the Discrete Wavelet Transform so that frequency-domain characteristics could be extracted. In addition, in order to extract the amplitude envelope of the signal, which is a feature that is frequently utilized for SER, we made use of the Hilbert Transform. After that, we took the preprocessed audio signals and used them to build spectrograms, which are essentially graphical representations of the time-varying frequency content of the signal. We placed these spectrograms in a folder for later use and fed them into an image classification system that was already pre-trained using the VGG16 algorithm. In addition to VGG16, we conducted research and testing on a number of different machine learning models, including SVM, Decision Tree, and Logistic Regression. The new architecture that we have developed has a number of advantages over the SER systems that are currently in use. We are able to extract a wide variety of features from the audio signals by combining a number of different preprocessing techniques, which can lead to an improvement in the accuracy of the model. In addition, by employing a VGG16 model that has already been pre-trained, we are able to make use of the model's several deep convolutional layers and obtain a high level of accuracy in picture classification tasks. In general, the architecture that we have provided provides an innovative approach to SER by combining a number of different methods for audio preprocessing and image classification. We hope to make a significant contribution to the expanding body of research on emotion detection in speech and attain a high level of accuracy in SER tasks through the application of these strategies. Fig. 8 shows the flowchart for the proposed architecture.

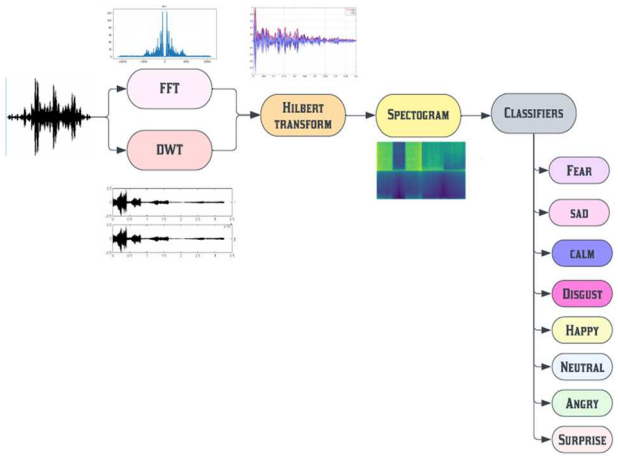

Fig. 8. Flow chart of Proposed Modal for SER

Algorithm: SER Classification
Input: $x_0, x_1, \ldots, x_n$, n Sound Signal Step 1: Frequency domain charcaterization (FFT Signal) $X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i k n}{N}}$ (DWT Signal) $D = waveletTransform(X)$ Step 2: Hilbert Transform $H(x(t)) = P.V.(1/\pi) \int (x(\tau)/(t-\tau))d\tau$ Step 3: Generate Spectrogram $STFT\ [n,k] = \sum_{m=-M}^{M} x[n+m] * w[m] * exp^{(-j*2mk/N)}$ Step 4: Classification Result = VGG16($STFT\ [n,k]$)

IV. Experimental Evalution and Result Analysis

SER is the identification of a person's emotional state based on their speech. In recent years, a variety of signal processing techniques, such as FFT, DWT, Hilbert Transform, and spectrogram, have been applied to the recognition of vocal emotions [41]. In this analysis, we evaluated the efficacy of these SER techniques. We tested our suggested SER model on a test set of audio samples that included eight distinct emotions: happy, sad, neutral, fear, disgust, surprise, angry and calm of both datasets. The experiment was carried out on Google CoLab using python 3.7. The model was trained with 80% data and tested on 20% data. The image is scaled into dimension of 224 x 224, with batch size 16. The input image is normalized by ./255 and augmented with shear range 0.2, zoom range 0.2, horizontal flip set as true. To test our model, we have compared the performance with CNN model.

A. Dataset

For any supervised learning problem, the most impactful component is a dataset. For speech emotion detection, the dataset acquired was a collection of 8882 audio files merged by two famous dataset RAVDEES and Crema-D. This particular segment of RAVDESS contains a total of 1440 files, with each actor contributing 60 trials, and a total of 24 actors, comprising an equal number of male and female professionals. Both of the actors make two comments that are lexically consistent with one another and are delivered with a standard North American accent [38]. The dataset contains a wide range of emotions, including calmness, happiness, sorrow, rage, fear, surprise, and repugnance. Each of these emotions, in addition to a neutral expression, is produced at two levels of emotional intensity: normal and intense [39]. All things considered, the RAVDESS is a complete dataset that includes a wide variety of feelings, actors, and speech characteristics that can be applied to a variety of research endeavors. The dataset known as CREMA-D has 7,442 segments, with the voices of 91 actors, ranging in age from 20 to 74. There are 48 male actors and 43 female actresses in the collection. It was documented that these actors, who come from a variety of racial and ethnic origins (including African American, Asian, Caucasian, Hispanic, and Unspecified), spoke a total of twelve unique sentences. The statements were uttered with one of six different emotions, namely Anger, Disgust, Fear, Happy, or Neutral, and with one of four different intensities of emotion, namely Low, Medium, High, or Unspecified [40].

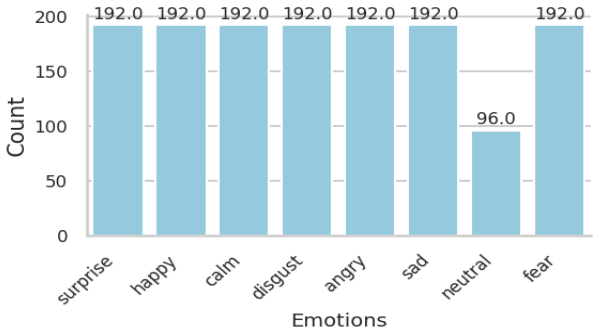

Fig. 9. Emotions count for RAVDEES dataset

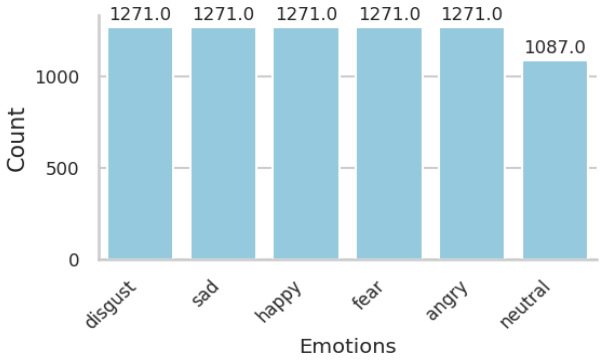

Fig. 10. Emotions count for RAVDEES dataset

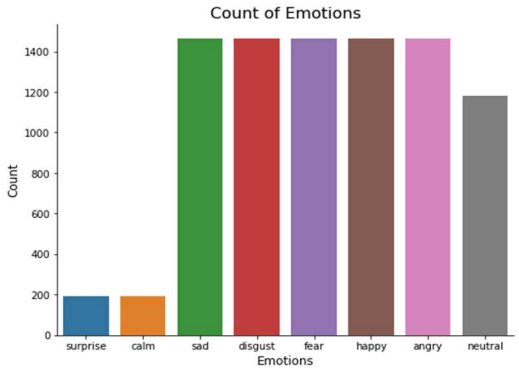

Fig. 11. Bar plot of emotions

The dataset contained 8882 files in total. After initial preprocessing and data cleaning, the final annotation file contained different number of class with sad disgust fear happy as high count class and calm surprise as low count class. The class imbalance has been visualized in the diagram fig. 9, fig. 10 and fig. 11 for better understanding.

B. Result Analysis

The performance analysis of CNN model attained a 72.84% percent accuracy on the training set and a 69.21% percent accuracy on the validation set after training for 30 epochs with a batch size of 16. Lesser batch sizes when tried out made the model over fit. Overfitting leads to memorization of training data points and defeats the purpose of modelling a robust classifier with considerable tolerance to variations in input. The model history was recorded during the fit operation and the training loss and validation loss were plotted to observe how the model evolves over various training samples and across increasing epochs. Fig. 12 shows how the model performed over the first 25 epochs. The training loss and validation loss was observed to be going down gradually over epochs which is desirable. Call-backs (from the Keras library) were used for early stopping of models with a tolerance of 5 epochs during the fit operation i.e. the model training was stopped earlier if the validation loss did not go down for 5 consecutive epochs.

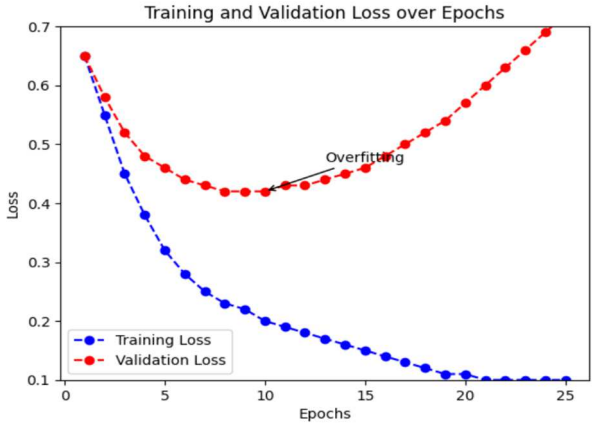

Fig. 12. Training and validation loss in the first 25 epochs of training

On the other hand, when the feature set consisted of segmented audio samples and the spectrogram image generated stored in different folder extracted from them, the model performed better with little to no overfitting. Again, the model history was recorded during the fit operation and the training loss and validation loss were plotted to observe how the model evolves over various training samples and across increasing epochs. The model performance across 30 epochs with a batch size of 16 is shown in Fig 13.

The error margins appear to go down extensively in the beginning of the training process and then gradually stabilize at a non-zero value in the end which is desirable in any training process. Based on the classifiers, we can say that the VGG16 model has the highest accuracy among all of the classifiers, coming in at 88.43%. This indicates that when compared to the other models, such as Logistic Regression (69.98%), Random Forest (67.65%), Convolutional Neural Network (72.84%), and Decision Tree (58.34%), VGG16 has the best performance for the problem at hand. The Visual Geometry Group (VGG) at the University of Oxford developed the VGG16 model, which is a deep convolutional neural network architecture. This design was initially presented by VGG. It is made up of sixteen layers, each of which contains a small 3x3 filter and a max-pooling layer in between. The VGG16 model has already undergone preliminary training using the ImageNet dataset, which includes millions of photos depicting things that fall into one thousand distinct categories. The great accuracy of the VGG16 model can be attributed to its deep architecture as well as the fact that it was pre-trained on a big dataset that contained a variety of different types of data. The VGG16 model's many layers make it possible for it to understand intricate patterns and characteristics from the data it is fed. The first training on the ImageNet dataset gives the model a solid understanding of the fundamental characteristics shared by a variety of objects. This, in turn, makes it much simpler for the model to learn the more specialized characteristics required for the task at hand.

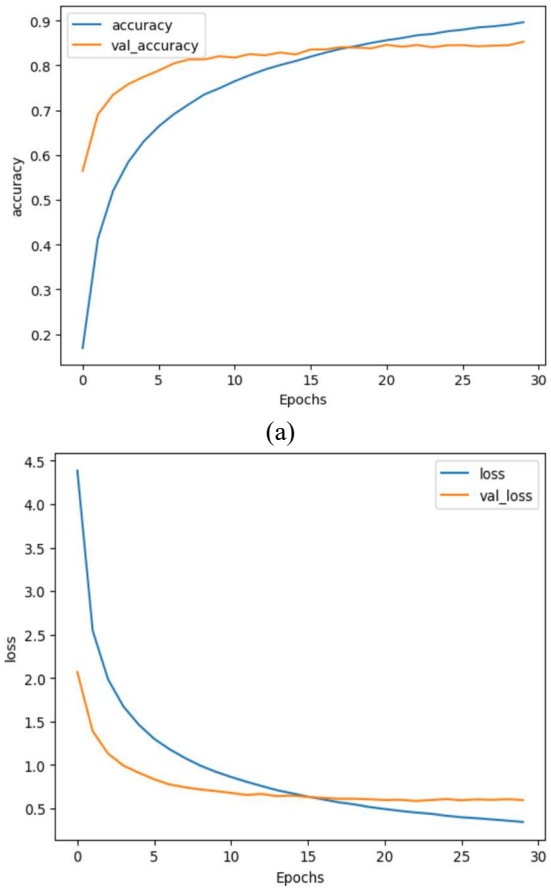

Fig. 13. Evolution of a) training accuracy and b) Validation loss across 30 epochs using Vgg16 CNN

The Decision Tree classifier, on the other hand, has the lowest accuracy of all the classifiers, coming in at 58.34%. It's possible that this is due to the fact that decision trees are very simple models that aren't able to capture more complicated relationships in the data. Decision trees are susceptible to overfitting, which can result in subpar performance on data that has not previously been examined. The overall performance of the various classifiers varies greatly depending on the task at hand and the dataset that is being utilized. When choosing a model for a certain endeavor, it is essential to take a number of aspects into consideration, including the level of model complexity, the amount of time required for training, and the degree to which the model may be interpreted. Fig. 14 shows the model comparison in accuracy with various classical and deep learning classifiers. Table I depicts the comparison of result achieved by various state of art methods with proposed model.

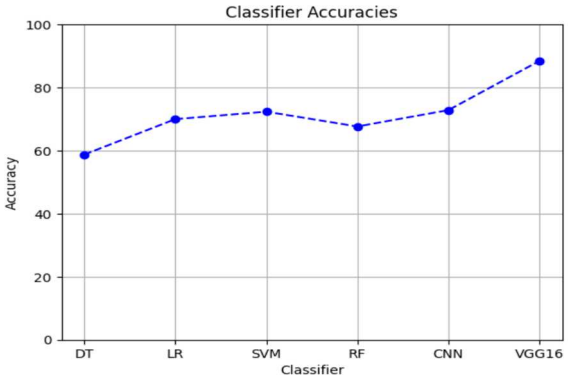

Fig. 14. Model Comparison

TABLE I. COMPARISION RESULT WITH DIFFERENT METHODS

Methods	Average Accuracy
DNN-decision tree SVM[42]	75.83
MFCC[43]	84.81
FusionStrategy[44]	58.3
Proposed Model	89.0

V. CONTRIBUTION

- Diverse Emotion Sound Dataset: Collected emotion sound samples from both male and female actors, encompassing a wide range of sentiments, allowing for a comprehensive study.
- Data Pre-processing and Feature Extraction: Normalized the sound samples and extracted relevant features, enabling effective analysis.
- Feature Selection: Demonstrated that mel-spectrogram features are well-suited for emotion classification when used with Convolutional Neural Networks, providing valuable insights into feature selection.
- Uniform Model Dimensionality: Achieved uniformity in model dimensionality by normalizing single-channel sounds, enhancing the consistency of the experimental setup.
- Augmented Training Sets: Trained the model on augmented datasets, enhancing the model's ability to generalize and improving classification accuracy to 89 percent.
- Hyper parameter Optimization: Rigorously optimized hyper parameters and employed an Adam optimizer for minimizing loss, contributing to the high classification accuracy.
- Transfer Learning Approach: Introduced an alternate approach using CNN-VGG16 transfer learning features, demonstrating improved performance and robustness, especially in the presence of noise and non-uniform samples.
- Composite Feature Sets: Suggested a potential future direction by advocating for composite feature sets, which can enhance classification accuracy by utilizing a combination of features.
- Classifier Evaluation: Evaluated various basic machine learning classifiers (SVM, Logistic Regression, Decision Tree, Random Forest) alongside the main classifier, highlighting the superiority of the main classifier with an accuracy of 89%, compared to the 60%-75% accuracy of alternative classifiers.

VI. CONCLUSION

In this work, we worked with emotion sound samples recorded via different actors namely male and female with different sentiments like happy, sad, fear, neutral, calm, angry, surprise and disgust. These sounds were then normalized and relevant features were extracted from them. Extensive analysis and trial-and-error adequately established that mel-spectrogram features performed sufficiently well with a Convolutional Neural Network. The model dimensionality was made uniform by normalizing single-channel sounds. The aforementioned model was trained on augmented and augmented sets and finally, a classification accuracy of 89 percent was obtained after experimenting with an optimum number of hyper parameters and using an Adam optimizer to minimize loss. The alternate approach employed with CNN-VGG16 transfer learning features gave us better performance and showed more robustness to noise and non-uniform samples. This study can be carried forward by making the classification depend on a composite feature set in the feature set instead of depending on one feature. In addition to the classifier used in the main experiment, other basic machine learning classifiers were also tested to compare their performance. These classifiers included SVM, Logistic Regression, Decision Tree, and Random Forest. The results of these tests showed that the accuracy of these classifiers ranged from 60% to 75%, which was lower than the accuracy achieved by the main classifier used in the experiment. These findings suggest that the main classifier used in the experiment is more effective at accurately classifying the data. However, further research may be needed to explore the potential classifiers in other benefits of using these alternative contexts.

REFERNCES

[1] M. Arsalan and M. M. Hassan, "Speech Emotion Recognition using Deep Learning and Signal Processing Techniques," in 2022 6th International Conference on Computing, Communication and Security (ICCCS), 2022, pp. 1-6.

[2] Hossain, M. Shamim and Muhammad Ghulam. “Emotion recognition using deep learning approach from audio-visual emotional big data.” Inf. Fusion 49 (2019): 69-78.

[3] Chen M, etal., "Emotion Communication System," in IEEE Access, vol. 5, pp. 326-337, 2017.

[4] Lane. N etal., ‘‘Can deep learning revolutionize mobile sensing’’ in Proceeding, ACM, Workshop Mobile Computing System Application., 2015, pp. 117–122.

[5] Javier G. etal., “SER in emotional feedback for Human-Robot Interaction” International Journal of Advanced Research in Artificial Intelligence(IJARAI), 4(2), 2015.

[6] P. Harár, R. Burget and M. K. Dutta, "Speech emotion recognition with deep learning," 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2017, pp. 137-140, doi: 10.1109/SPIN.2017.8049931.

[7] Mahmood, Arzo & Köse, Utku. (2021). Speech recognition based on Convolutional neural networks and MFCC algorithm. 1. 6-12.

[8] Lalitha S, etal., ‘‘Speech emotion recognition,’’ in Proceeding. International Conference. Advance Electronics Computing Communication (ICAECC), Oct. 2014, pp. 1–4.

[9] Scherer K, ‘‘What are emotions, and how can they be measured’’ Social Scientific Information., 44(4), pp. 695–729, 2005.

[10] Balomenos T., etal, ‘‘Emotion analysis in man-machine interaction systems,’’ in Proceeding International, Workshop Machine Learning Multimodal Interact, Springer, pp. 318–328, 2004.

[11] Cowie R, etal, ‘‘Emotion recognition in human computer interaction,’’ IEEE Signal Processing Magazine., vol. 18(1), pp. 32–80, 2001.

[12] Kwon O, etal ‘‘Emotion recognition by speech signal,’’ in Proceeding EUROSPEECH, Geneva, 2003, pp. 125–128.

[13] Picard R, ‘‘Affective computing,’’ Perceptual Computing. Section., Media Lab., Tech. Rep., 1995.

[14] Koolagudi etal, ‘‘Emotion recognition from speech: A review,’’ Int. Journal of speech Technology, vol. 15, no. 2, pp. 99–117, 2012.

[15] Ayadi M etal, ‘‘Survey on speech emotion recognition: Features, classification schemes, and databases,’’ Pattern Recognition, vol. 44(3), pp. 572–587, 2011.

[16] Dileep D, etal, ‘‘GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines,’’ IEEE Transaction Neural Network Learning System, vol. 25(8), pp. 1421–1432, Aug. 2014.

[17] Deng L, etal, ‘‘Deep learning: Methods and applications,’’ Foundations and Trends in Signal Processing., vol. 7, nos. 3–4, pp. 197–387, Jun. 2014.

[18] Schmidhuber J, ‘‘Deep learning in neural networks: An overview,’’ Neural Network., vol. 61, pp. 85–117, 2015.

[19] Vogt T, etal., ‘‘Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition,’’ in Proc. IEEE International Conference Multimedia Expo (ICME), pp. 474–477, 2005

[20] Anagnostopoulos C, etal, ‘‘Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011,’’ Artificial Intelligence Review., vol. 43(2), pp. 155–177, 2015.

[21] S. Yoon, S. Byun and K. Jung, "Multimodal Speech Emotion Recognition Using Audio and Text," 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 112-118, doi: 10.1109/SLT.2018.8639583.

[22] Sandesara, Anushka & Parikh, Shilpi & Sapovadiya, Pratyay & Rahevar, Mrugendrasinh. (2020). A Comparative Study On Speech Emotion Recognition. International Journal of Research in Engineering, Science and Management. 3. 25-35. 10.47607/ijresm.2020.366.

[23] Samin, Yaser & Sadhin, Md. Shanjidul & Ifty, Redwanul Haque. (2023). Speech Emotion Recognition using Transfer Learning Approach and Real-Time Evaluation in English and Bengali Language. 10.13140/RG.2.2.31324.87684.

[24] J. Han, etal., ‘‘Reconstruction-error based learning for continuous emotion recognition in speech,’’ in Proceeding Conference Acoustic Speech Signal Process, pp. 2367–2371, 2017.

[25] Zeng Z, etal., ‘‘A survey of affect recognition methods: Audio, visual, and spontaneous expressions,’’ IEEE Trans. Pattern Analysis Machine Intelligence, v 31, no. 1, pp. 39–58, 2009.

[26] Vogt T, etal., ‘‘Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realization,’’ in Affect and Emotion in Human-Computer Interaction. pp. 75–91, 2008.

[27] Kuo, C. C. J., & Gao, J. (2019). Digital signal processing techniques and applications in audio and speech processing. In C. S. Chen, Y. Zhang, & Y. Chen (Eds.), Multimedia Big Data Computing for IoT Applications (pp. 209-231). Springer.

[28] Hazarika, S. M., & Biswas, R. (2021). Speech Emotion Recognition Using Deep Learning: A Review. Journal of Ambient Intelligence and Humanized Computing, 12(8), 8549-8577.

[29] Saha, G., & Acharya, A. (2018). Speech Emotion Recognition: A Review. International Journal of Computer Applications, 179(14), 1-7

[30] Boashash, B. (2015). Time-frequency signal analysis and processing: a comprehensive reference. Academic Press.

[31] Kumar, A., etal,” Speech emotion recognition using spectrogram and mel-frequency cepstral coefficients, 4th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 1113-1116), (2020)

[32] Maddage, N. C. etal, “SER using machine learning algorithms with spectrogram and energy based features”, International Conference on Advances in ICT for Emerging Regions (ICTer) (pp. 73-78), 2018.

[33] Zhang, W., Zhao, S., Dong, Y., & Chen, M. (2019). Speech emotion recognition using deep convolutional neural network and spectrogram. In 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS) (pp. 55-59).

[34] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[35] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, Inception-ResNet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence.

[36] Liu, Q., Li, S., Li, Y., & Lai, J. (2021). Speech emotion recognition based on transfer learning using VGG16 and a small dataset. Applied Sciences, 11(1), 214.

[37] Kshirsagar, M., Mahalle, P. N., & Shinde, S. S. (2021). A Comparative Study of Pretrained Convolutional Neural Networks for Speech Emotion Recognition. Procedia Computer Science, 187, 127-135.

[38] Shahid, N., & Rauf, H. (2021). Deep Learning Based Speech Emotion Recognition: A Review. In 2021 International Conference on Computer, Control and Communication (IC4-2021) (pp. 1-6). IEEE. https://doi.org/10.1109/IC4-2021.2021.9527689

[39] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), e0196391. https://doi.org/10.1371/journal.pone.0196391.

[40] Cao, H., Cooper, E. W., Keutmann, M. K., & Gur, R. C. (2014). CREMA-: Crowd-sourced Emotional Multimodal Actors Dataset. IEEE Transactions on Affective Computing, 5(4), 377-390. doi: 10.1109/TAFFC.2014.2338696

[41] L. Sun, etal., ‘‘Decision tree SVM model with Fisher feature selection for speech emotion recognition,’’ EURASIP Journal Audio, Speech, Music Process., pp. 1–14, (2019)

[42] L. Sun, etal ‘‘Speech emotion recognition based on DNN-decision tree SVM model,’’ Speech Communication., v 115, pp. 29–37, (2019).

[43] H. S. Kumbhar, etal., ‘‘Speech emotion recognition using MFCC features and LSTM network,’’ in Proceeding. 5th International. Conference. Computing., Communication., Control Automat., pp. 1–3, (2019.

[44] Yao Z, etal, ‘‘Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MSCNN and LLD-RNN,’’ Speech Communication, v 120, pp. 11–19, 2020.