National College of Ireland

Development of Speech emotion recognition using Deep Neural Network Architecture for children with Autism Spectrum Disorder

MSc Research Project
Data Analytics

Bijalben Prafulchandra Bhagat
Student ID: x20167326

School of Computing
National College of Ireland

Supervisor: Dr. Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Bijalben Prafulchandra Bhagat
Student ID:	x20167326
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Christian Horn
Submission Due Date:	16/12/2021
Project Title:	Development of Speech emotion recognition using Deep Neural Network Architecture for children with Autism Spectrum Disorder
Word Count:	8339
Page Count:	23

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

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

Signature:	
Date:	31st January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

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

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

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

Development of Speech emotion recognition using Deep Neural Network Architecture for children with Autism Spectrum Disorder

Bijalben Prafulchandra Bhagat
x20167326

Abstract

Autism spectrum disorder (ASD) is mostly diagnosed in children who find it very difficult to show the human emotion and socialize themselves. In order to help such childrens, our proposed framework utilizes the deep learning architecture for classifying the various emotions from audio data. After several steps of data collection, data processing, model training and evaluation, we have obtained the better results using Deep neural network architecture. Where the existing model was compared with the convolutional neural network and custom model (combined layer of LSTM and Bi-LSTM). The main purpose of this study, is to identify the optimal model for developing a speech emotion recognition system for children with autism specturm disorder in identification of human emotions.

1 Introduction

Since the advent of humans, social interaction has been a crucial part of our lives. The social interaction among humans had led to the development of societies. Particularly, the interaction may be verbally or non-verbally. Most human interactions rely on the verbal mode of interaction, but the non-verbal mode of interaction remains to be the baseline. This interaction varies such as it may be actions, gestures or postures which deliver the emotional aspect behind every interaction channel. It ultimately aligns with the emotional behaviour or intelligence of humans during social interaction. Emotional behaviour directly contributes to the motive for every communication in the human interface. Emotional intelligence is the most crucial part of the human intellect, and it develops since birth through brain-environment interaction. As human grows from babyhood to childhood, the brains also develop the emotional intelligence to interact. Various types of emotions develop in these growing years that can be either common emotions or complex emotions. Although common emotions such as happiness, sadness, fear, anger, disgust, and surprise develop through the basic and routine type of interaction whereas the complex emotion such as jealousy, pleasure, regret is the combination of the common emotions which are developed over time. At a certain stage of childhood, emotional intelligence is developed in comparison to adulthood, gaining a sense of maturity. Again, this depends on the brain-environment interaction since the early stage of life where maturity may be gained. In this, humans gain the ability to interact and understand the emotions of others and themselves. But in some cases, a human is incapable to interact with the

environment and understand emotions. This hinders the social interaction of a person and the development of emotional intelligence. Such inability may be caused due to the neurodevelopment disorder called autism spectrum disorder (ASD).

A person with ASD generally experiences disruption to emotional behaviour and social interaction in an open environment. There is no exact cause of this disorder, but it may appear due to various reasons such as genetic, non-genetic or lesser environmental interactions. There are various types of ASD's such as Asperger's syndrome, Rett syndrome, childhood disintegrative disorder (CDD), Kenner's Syndrome and Pervasive development disorder (PDD-NOS). In this disorder, emotion recognition is the greatest challenge that further hinders the development of interactive behaviour in a person. Although, this disorder when detected earlier in the person can be overcome through the various strategic interactive development approach. It is usually diagnosed in children, who appear incapable of proper interaction. Through the effective approach of interaction, the children can develop the behaviour when reaching adulthood. Although the emotional interaction cannot be compared with the intelligence of the person. In the preliminary stage, the main motive is to train and guide the ASD-driven children to interact with the visual stimuli. For this, it is necessary to recognize the emotional behaviour in them which can help enhance the cognitive behaviour. Therefore, various approaches to recognize the emotion were utilized such as eye-tracking (ET), electroencephalogram (EEG), gesture pattern, or speech pattern. Through each channel such as eye, brain, gesture, or speech the emotions are recognized. Over time, multiple researchers in the study utilized advanced learning mechanisms such as machine learning and deep learning. In this paper, we have studied the approach to recognize the emotion through the speech of a person using word embedding techniques. With this, the emotions of the child can be effectively recognized which can help develop a connection and interaction-able behaviour. Every emotional instance recognized can be utilized to train the emotional intelligence of the child to interact with the environment effectively.

For the recognition of the emotion, the audio dataset needs to be sourced initially which may consist of unwanted noises and distortion apart from the speech modalities. These audio datasets can be either sourced from the real-time environment or the pre-build datasets such as the audio emotion. In this research, we have collected the standard audio dataset from multiple process and performed certain set of proe-processing in order to achieve the better results. After certain steps of data pre-processing model need to be trained. The following model (Deep neural network (DNN), Convolutional neural network (CNN) and custom model) has been trained over the audio data for classifying the human emotion. The performance of these models have been evaluated by calculating the Precision, Recall, F1-Score, Loss and accuracy measures.

1.1 Research Question

- Which algorithm can correctly identify the human emotions from audio-speech data?

2 Literature Review

In this section, the studies by various other researchers are discussed. This section is further divided into the sub-sections such as Study on Autism Spectrum Disorder, emotion recognition using eye-tracking, EEG, gesture, and speech.

2.1 Study on Autism Spectrum Disorder

Kokkinaki et al. (2021) extensively studied the emotional expression in children experiencing the autism spectrum disorder. In this study, the approach to interpreting facial emotional expressions was studied using novel methodologies. A two-group mechanism based on pre and post-test designs was implemented to evaluate the basic facial emotional expression in ASD-induced children. These basic emotions were happy, sad, fear and anger. The children ranging from age of five years to seventeen years were taken into the consideration with some criteria. These criteria were the presence of comorbidity, lack of verbal ability, and the intellectual level less than the normal level. This paper implemented two measures which are the facial emotion recognition task (FERT) and the basic emotions production task (BEST). Also, it was noted that there was no difference between the two facial emotion recognitions systems which are the humanoid-robot mechanism and computer mechanism. Amy and the team analyzed the impact of autism spectrum disorder on the lives of children Stedman et al. (2019) The authors here stated that the research for this topic has been underrepresented and various measures shall be considered to check the remedial ways. Over time various researchers have studied the domain but were limited with the certain advance which couldn't help further. Here, in the paper, various reports were analyzed regarding the autism spectrum disorder. Considering the type of studies, randomized control trial (RCT) contributed the major share of study standing at 49.7%. Whereas in the case of a type of intervention, pharmacological based research for ASD stood at 50.7%. On the other hand, the study for the type of intervention was proactively researched on core ASD symptoms which stood at 57.8%.

Chauhan et al. (2019) did a systematic review and an analysis on the prevalence of autism spectrum disorder in children under Indian geographic. The authors for the analysis utilized various repositories such as PubMed, OvidSP, and Embase. Through these research databases, various studies were sourced and extracted for the further detailed examination of the study. Furthermore, the data were analyzed using software such as STATA MP12. The author studied the four variants namely, for both urban and rural populations. In the final take, the author stated the urge for advanced research and implementation for the autism spectrum disorder in Indian children. Hodges et al. (2020) deeply studied the fundamentals of autism spectrum disorder and further evaluated it. In the paper, the definition, epidemiology, causes, and clinical evaluations were precisely interpreted. The increased diagnosis of ASD in the current situations has brought various researchers to deeply study the cause and remedies using advanced technological implementations. This disorder is a neurological concern and deprived environmental and social interaction. Also, the paper showed a detailed approach for the clinical evaluation with the strategic periodic assessment. Different variants of ASD were interpreted such as DSM-4 and DSM-5. Furthermore, both the variants were compared with the detailed evaluation of the fundamentals.

2.2 Emotion Recognition using Eye Tracking

Lim et al. (2020) studied the approach of emotion recognition using the eye-tracking mechanism. In social interaction and communication, eye movement or gazing shows multiple types of emotions. Through the eye, various emotions can be conveyed in the combination of verbal means of communication. This can be also helpful for the interpretation of emotional behaviour in ASD-induced children. This paper provides extensive research on eye-tracking-based emotion recognition which serves various domains. Different challenges, occurrences, mechanisms, and methodologies have been discussed in this paper. This system relies on the factor of eyes such as pupil dimension, pupil position, EOG, fixation duration, the length between sclera and iris, motion speed, and pupillary responses. On the other hand, Tarnowski et al. (2020) analyzed the eye-tracking system for the emotion recognition mechanism. Here various evoking emotions were captured through the visual stimuli and sound stimuli. Primarily, the video was shown with different audio tracks to the person. 30 different participants were considered for the experiments. In this investigation, six basic emotions were considered such as happiness, sadness, anger, surprise, disgust, and fear. Samples with neutral emotion evocation were also considered in the study. Furthermore, 18 different characteristics were considered and assessed for the study. In addition to this, three different classes were considered such as high arousal and low valence, low arousal and high valence and high arousal and high valence. This model uses the support vector algorithm (SVM) which attained an accuracy of 80% with the leave one subject out validation algorithm.

Zheng et al. (2020) proposed a model for emotion classification under four classes in virtual reality utilizing pupillometry. The emotion classifications were majorly studied using the EGG or ECG signals but the study using the eye-tracking system was limited. Therefore, the author chooses to study this novel approach and thereby increasing the study base. The pupillometer calculated the variance in the pupil diameter during the emotion evocation. This will help to divide the emotion into four different types of classes. This concept relied on the working of Russell's Circumplex Model of Emotions. The video is shown on a 360-degree screen which is generally in VR devices, and then through these emotions are captured. Furthermore, in this model algorithms such as SVM, KNN and RF are utilized. Haber et al. (2019) analyzed the gaze patterns in an emotion recognition task for children with autism through wearable smart glasses. In the study, 16 children with autism spectrum disorder and 17 children without autism spectrum disorder were considered. Through the efficiency of the system was evaluated and compared. These wearable smart glasses were integrated with the custom eye-tracker. During the investigation, the researcher presented the images to the children sourcing them from the online repositories. Their responses were then captured for further evaluation. Thereafter, the real-time application responses were captured and these were then compared with later evaluations of the efficacy of the model. Although this model couldn't outperform the existing algorithms and therefore need further enhancement in the study. There needs to be further research study in the field of ASD phenotype. But this advanced technology can be implemented in a real-time environment with a better approach to study.

2.3 Emotion Recognition using EEG signals

Gannouni et al. (2021) proposed a novel approach of emotion classification through EEG signal using a zero-time windowing-based epoch estimation and relevant electrode identification. Through this paper, the author aimed to improve the efficacy of the traditional approach and the usage of the algorithms in widespread. This model recognizes the brain signal in a new and dynamic way which enhances the performance. In addition to this, the epoch connection in the model is also increased tremendously to accelerate the output. In this, the zero-time windowing mechanisms extract the spectral information of the samples by utilizing the numerator group delay function to efficacy detect the dedicated epoch. This study utilized the DEAP dataset which achieved an accuracy of 89%. The accuracy was enhanced by 8% when compared to the existing traditional. The proposed approach also outperformed the traditional approach which had fallacies such as being limited to three or four emotions.

Similarly, Bazgir et al. (2018) also utilized a machine learning approach to classify emotion through EEG. Considering this mechanism, the EEG signals are initially divided into the packets called alpha, beta, theta, and gamma frequency bands utilizing the discrete wavelets transform (DWT). Then the spectral features are extracted from it. During the process of feature extraction, a mechanism such as principal component analysis (PCA) is used. Thereafter, advanced learning algorithms such as K-Nearest Neighbours (KNN), Support Vector Machine (SVM) and Artificial Neural Network (ANN) are utilized for classification. This study also utilized the DEAP dataset the study. In the final take, the SVM with radial basis function (RBF) is evaluated through the appropriate metrics. The evaluation showed that the model for the arousal showed 91.3% of accuracy whereas the model for the valence showed 91.1% of accuracy. This proposed model outperformed the existing algorithms. Wang and Wang (2021) reviewed the study of emotional feature extraction and classification using the EEG signals. For feature extraction, methods such as domain analysis, frequency domain analysis and time-frequency domain analysis were utilized. For the machine learning algorithms, SVM, KNN and Naive Bayes (NB) were utilized in the study and for the deep learning algorithms, Neural Network (NN), Long Short-Term Memory (LSTM) and Deep Belief Network (DBN) were utilized. The accuracy for this ranged from 57.5% to 95.7% for machine learning whereas for deep learning it ranged from 63.38% to 97.56%. On the other hand, Topic and Russo (2021) utilized the feature mapping technique of a deep learning algorithm for emotion classification. This proposed model uses the feature map of EEG signals which are TOPO-FM and HOLO-FM. Also, the investigation utilized the samples of the datasets such as DEAP, SEED, DREAMER and AMIGOS. This proposed model works efficiently with any type of dataset.

2.4 Emotion Recognition through gesture

Naik and Mehta (2018) utilized a deep learning technique for emotion detection through handover face gestures. The author examined the existing research initiatives, which had the greatest challenge hand disruption during the detection of face for emotion classification. Therefore, in the study, the researchers omitted the distortion caused by the hand gestures. But to overcome these challenges and improve the overall accuracy, the au-

thor in the study proposed to interpret the face with the incorporation of hand gestures. This model further utilized advanced learning algorithms such as CNN and RNN. Wu et al. (2020) proposed a generalized zero-shot framework for emotion recognition using body gestures including the hand. In the baseline model, the author used Bi-directional Long Short-Term Memory (Bi-LSTM) with the self-attention mechanism. Piana et al. (2016) proposed a framework for the adaptive body gesture representation for automatic emotion recognition. Various approaches have been discussed in the paper included the research of various others. Mellouk and Wahida (2020) proactively reviewed facial emotion recognition utilizing deep learning algorithms. Furthermore, the insight of various existing trends was discussed in the paper. In addition to this, the author suggested different ways to improve the conventional models.

2.5 Emotion Recognition using Speech Modalities and Word Embedding

A study by Atmaja et al. (2019) offers a categorized emotion identification system using voice characteristics and word embedding on the IEMOCAP dataset, which has 10039 words. To keep the sample data consistent, the fraction of expressions for each classification is almost the same. The authors employ speech fragments for voice-based emotion identification, which are formed by removing pauses in a sentence and recovering the audio characteristic; for literary sentiment classification, word embedding is used as a feature vector, and hybridization of the two attributes is offered for improved efficacy. Text is symbolized by two linear LSTM layers, but auditory expression detection is represented by entirely linked layers that are fused in an earlier fusion procedure by fully interconnected systems to produce predicted human emotions. The suggested technique yields 75.49% accuracy, which is higher than the precision achieved by solely using audio or only using word embedding, which is 58.29% and 68.01%, correspondingly. In another study, Yoon et al. (2018) utilized a subjectively deep twin cyclic transceiver approach that combines simultaneously text and auditory control messages to achieve a better comprehension of speech data. They employed dual RNNs to capture transcripts from speech and word fragments, and then a feed-forward neurological architecture to incorporate data from numerous resources to predict affective classifications. Audio recurrent encoder employed gated recurrent unit to forecast subcategories of audio signals; meanwhile text recurrent codec matched to a sequence of symbols with Natural Language Framework and went via word-embedding layer to anticipate using the SoftMax activation function. This approach analyses speech data via impulses to linguistic specifications, capable of making greater use of relevant data than systems that merely emphasize auditory characteristics. When contrasted to the IEMOCAP database, the conceptual model outperforms previous techniques in assigning data to all four emotional manifestations, with performance levels ranging from 68.8% to 71.8%.

Damian and his team proposed an approach based on the Support Vector Machine (SVM), which has been demonstrated to have high degrees of reliability in language processing when classifying inputs Valles and Matin (2021). Distinct audio samples should be created to construct emotion detection algorithms. The spoken dataset used in this investigation was the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS),

which was used to create the approach. The Libros module will be used to vectorize audio features. The zero-crossing rate (ZCR) and the first 26 Mel-frequency Cepstral Coefficients (MFCCs) are calculated and utilized to train the machine learning technique. The resulting SVM model had a 77 percent accuracy rate. This approach succeeds well when an extensive ambient noise is introduced to the RAVDESS recorded audio, attaining a forecasting accuracy of 64%. A convolutional neural network (CNN) with auditory word-based embedding has been described for emotion classification. In Huang's study, linear encoding was employed to transform the lower-level features of each voice clip into auditory utterances utilizing the k-means approach Huang et al. (2018). Word2vec is utilized to transform auditory audio vector space patterns about an intake vocal phrase into the CNN-based sentiment framework. The NCKU-ES collection, which comprises seven emotion classifications, was utilized to examine the outcomes of a CNN-based approach for voice synthesis utilizing a five-fold categorization approach. The suggested approach obtains an emotion categorization efficiency of 82.34%, which is an 8.7% increase over the Long Short-Term Memory (LSTM)-based approach, which had to cope with similar challenges of a long carrier frequency. In comparison to fundamental features, auditory word-based synthesis increased speech emotion identification by around 3.4%.

2.6 Research-based on Algorithms

In one study by Murugan (2020), the model for speech emotion recognition was proposed utilizing the Convolutional Neural Network (CNN) framework. In the model, the dataset was acquired from the custom sources called Ravedess for which it was further processed. Multiple categories were considered for the model to recognize. Using the inputted audio file, the spectrogram and the waveform are plotted for further implementations. Then by incorporating the Librosa python library, the Mel Frequency Cepstral Coefficient (MFCC) is extracted and stored. In the final evaluation and assessment of the model, the accuracy achieved by the model stood at 71%. Although the author aimed for higher accuracy which could be enhanced with further tuning and implementation in future work. Yao et al. (2020) proposed a model for a speech emotion recognition system by incorporating three crucial algorithms. The three algorithms considered for this study were recurrent neural network (RNN), deep neural network (DNN), and convolutional neural network (CNN). The architecture of Frame-level low-level descriptors (LLDs), segment-level Mel-spectrograms (MS), and utterance-level outputs of high-level statistical functions (HSFs) on LLDs was also incorporated in the algorithm. For the dataset, the IEMOCAP repository was considered. In the final stage, the author evaluated accuracies that are weighted and unweighted. The accuracy achieved here for weighted is 57.1% and for the unweighted is 58.3%. The fusion of these algorithms showed an amazing output. On the other, Mustaqeem et al. (2020) introduced a clustering-based speech emotion recognition by incorporating learned features and deep Bi-Lstm. The sequences from the given dataset were acquired from the radial-based function network. In addition to this, the obtained sequences were transformed into the spectrogram by utilizing the SIFT algorithms. Then after this Bi-LSTM algorithm was implemented in the model. The datasets used in this model were IEMOCAP, EMODB, and RAVDESS. Although, after evaluating the algorithms using the metrics, the accuracy achieved for each dataset is 72.25%, 85.57%, and 77.02%, respectively.

The comparative analysis for the various studies by different researchers is shown in Table 2.6

Author	Method	Advantages	Future Scope / Disadvantages
Kokkinaki et al. (2021)	This paper implemented two measures which are the facial emotion recognition task (FERT) and the basic emotions production task (BEST)	Extensively studied the emotional expression in children experiencing the autism spectrum disorder	It was noted that there was no difference between the two facial emotion recognitions systems which are the humanoid-robot mechanism and computer mechanism
Zheng et al. (2020)	In this model algorithms such as SVM, KNN and RF are utilized	This concept relied on the working of Russell's Circumplex Model of Emotions	The model can be enhanced
Gannouni et al. (2021)	In this, the zero-time windowing mechanisms extract the spectral information of the samples by utilizing the numerator group delay function to efficacy detect the dedicated epoch	This study utilized the DEAP dataset which achieved an accuracy of 89% and the accuracy was enhanced by 8% when compared to the existing traditional	No disadvantage found
Wu et al. (2020)	The author used Bi-directional Long Short-Term Memory (Bi-LSTM) with the self-attention mechanism	This model allowed to recognize the emotions effectively through body gesture	The study was limited and needed further enhancement

Table 1 : Comparison of Different Works for this Study

3 Methodology

Predicting the emotions from an audio/sound data is a challenging task. The effective of implementation of this technology can be useful for person suffering with ASD (Autism Spectrum Disorder). The main objective of this research is to identify the most optimal deep learning model for predicting the emotions from the audio data. To achieve this objective, a consistent flowchart/ Framework is proposed which consists of procedures that include audio data collection, data cleaning, data preparation, exploratory data analysis and visualization, feature extraction from different audio sources, model initialization, model training, and testing. In this section, a detailed explanation of every step is covered. The proposed framework for audio-emotion recognition is shown in Figure 1.

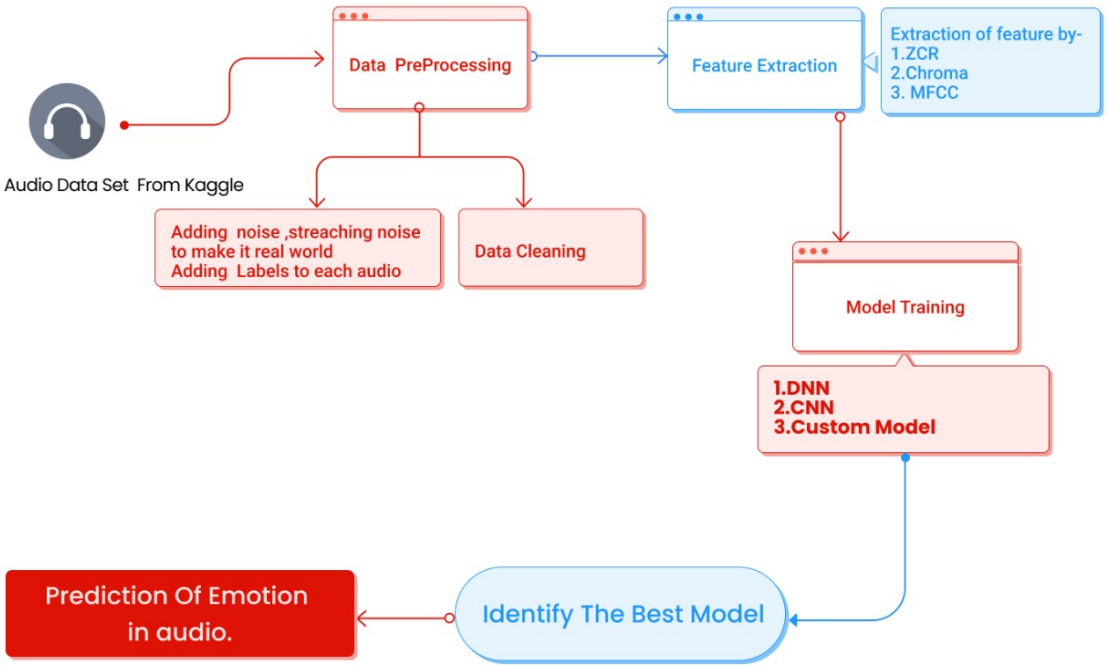

Figure 1: Proposed Emotion Classification from Audio Data

3.1 Data Set Description

Since, the main objective of this research is recognize emotion from audio data, the first step involves is collection of data. Therefore, we have collected the standard audio dataset from Kaggle The details of audio datasets with their size are as Follow:

- Audio-Visual Expressed Emotion (SAVEE) (162 Mb)
  Data Set Link:*Surrey Audio-Visual Expressed Emotion (SAVEE)* (n.d.)
- Toronto emotional speech set (TESS) (281 Mb)
  Data Set Link:*Toronto emotional speech set (TESS)* (n.d.)

- RAVDESS Emotional speech audio (590 Mb)
  Data Set Link:*RAVDESS Emotional speech audio* (n.d.)

- Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) (605 Mb)
  Data Set Link:*CREMA-D* (n.d.)

With the collected dataset, eight cardinal emotions can be detected which are neutral, calm, happy, sad, angry, fear, disgust and surprise.

3.2 Exploratory data analysis and visualization

After successfully data collection, the audio data has been visualised with their Waveplot and Spectrogram along with their labels. Each label is visualized with the help of spectrogram. The spectrogram is an efficient way to visualize a signal's strength/ loudness in points of time at different frequencies present in a particular waveform. In this section visualization of waveform and its respective spectrogram with an amplitude range of 2dB to 10dB is executed. Librosa Library has been used with stft method to plot spectrograms. The following figures 3 show the wave plot of audio from each label with its spectrogram. A bar graph is plotted to analyse the Frequency count of all the labels. From bar graph, it has been observed that the count of calm emotion is minimum. Surprise emotion counts are about 650 while the count of remaining emotions is almost same and higher.

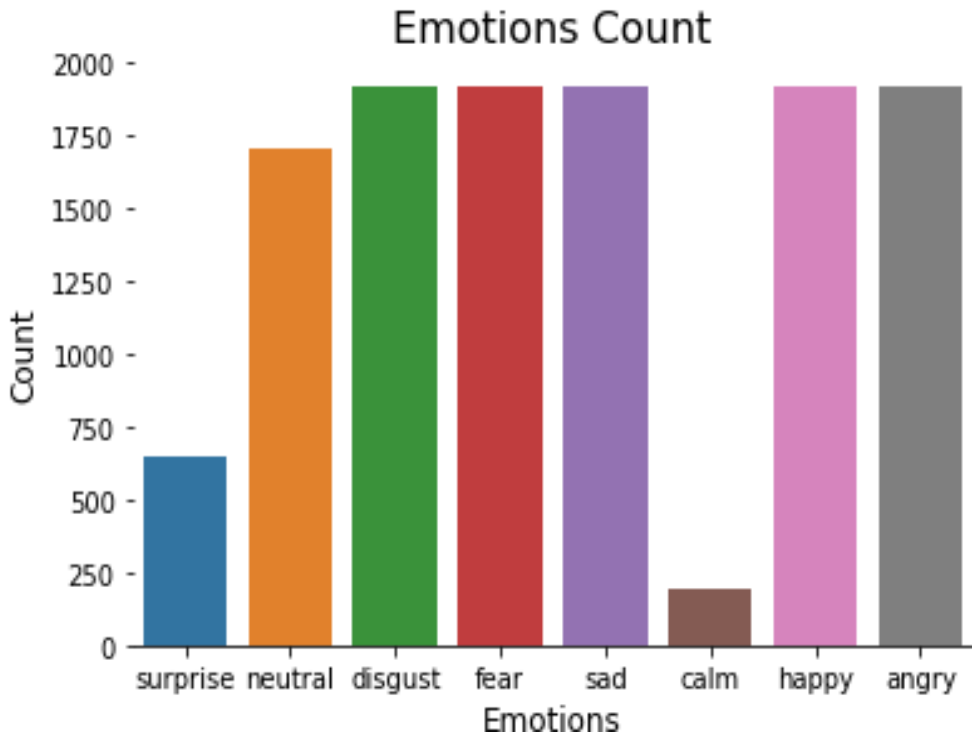

Figure 2: Bar plot showing counts of each label

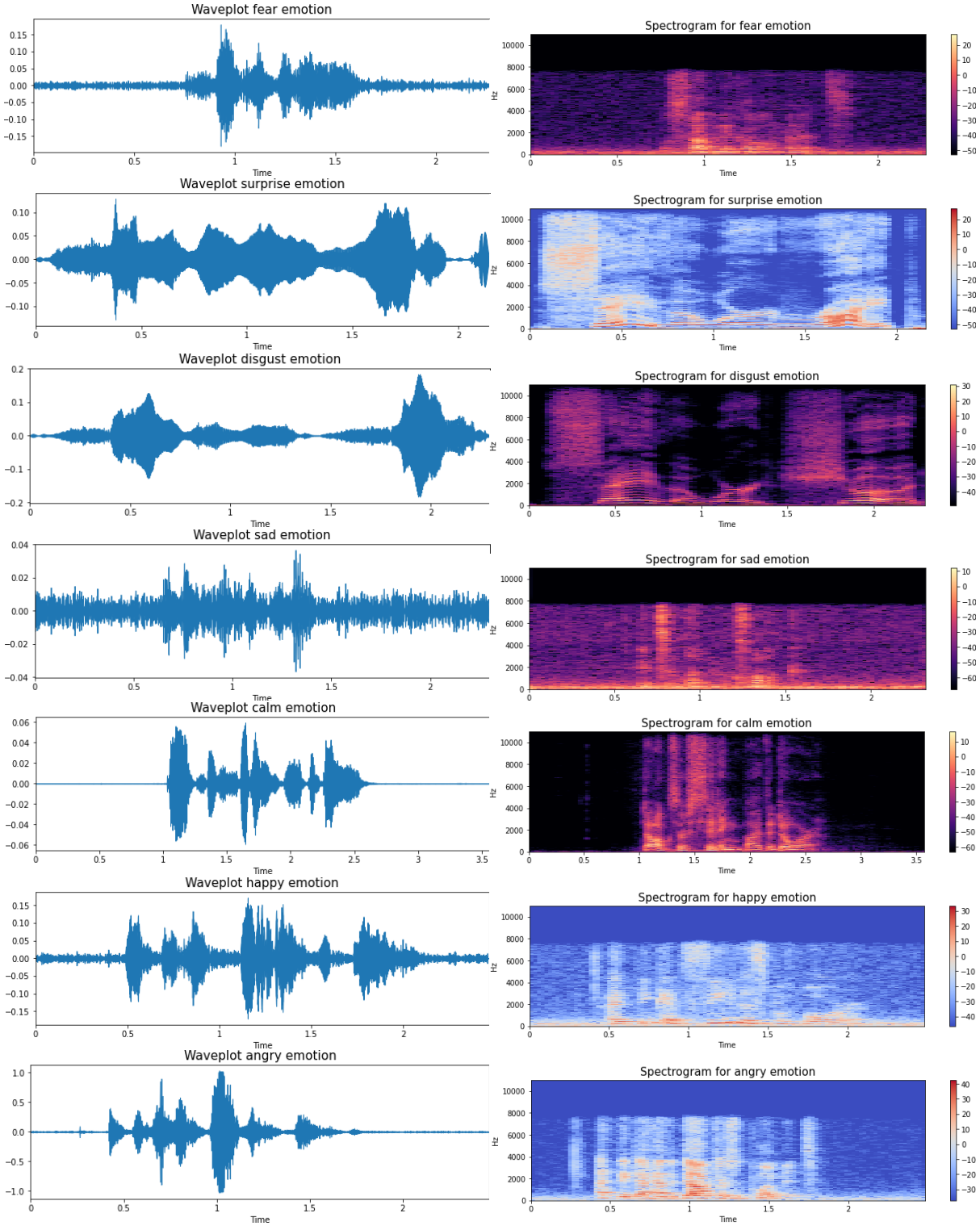

Figure 3: Wave Plots and Respective Spectrograms of all Emotions (Before Pre-processing).

3.3 Data Pre-processing

After collecting the dataset for better analysis and fitting data into deep learning models, pre-processing of the acquired dataset is performed. In pre-processing first step is extraction of audio and labels from the unstructured dataset as audio and labels are

present in string format. Path and emotions have been separated from each data file. After extracting labels and audio from each file concatenation of all data into one single panda's data frame is performed which will be used for later work.
Since normal audio data is not real-world data because real-world data consists of noise, elevated and different pitches, even long stretched voice is also found in real-world audios. As a part of pre-processing Noise, stretching and shifting voice operations are performed on the entire dataset. To perform these operations, the Librosa Library has been used. The audio data after pre-processing is shown in Figure 4.

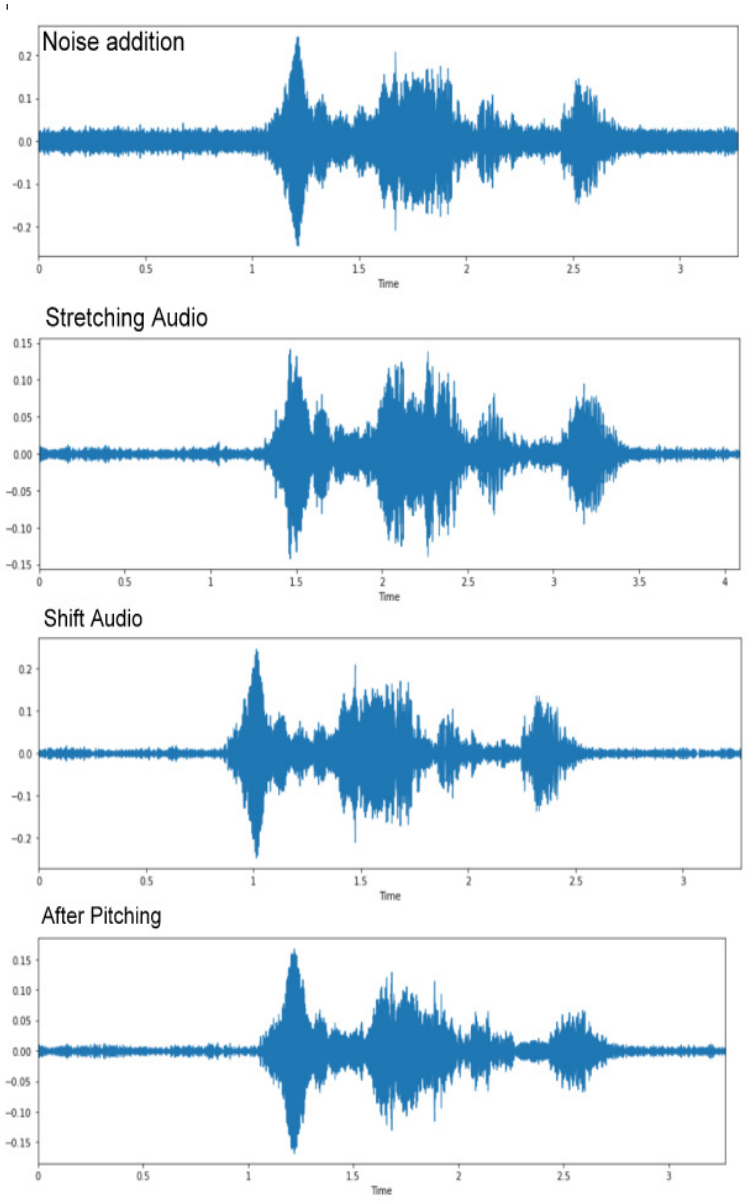

Figure 4: Audio data after Pre-processing

3.4 Feature Engineering

Feature engineering is one of the important processes in which domain knowledge is used to extract features(characteristics). It is used to enhance the performance of models. In

this research audio feature extraction is performed in which several features of an audio signal are extracted in order to perform predictive analysis for achieving better results. Audio feature extraction is very important step, in the for processing the audio signals. It is related to the manipulation of the audio signals and remove unwanted noise from the audio it also balances the time-frequency ranges by converting digital and analog signals suitably. ZCR, Chroma and MFCC are extracted from all audio signals. Zero-crossing rate (ZCR) is the feature of the signal which is the time required for a signal to change from positive to zero to negative or from negative to zero to positive. It is one of the important features in order to extract information from a signal which can be an audio signal. Chroma is a feature of an audio signal which represents the tonal content of a musical audio signal in a condensed form. Therefore, this feature can be considered an important prerequisite for high-level semantic analysis, like chord recognition or harmonic similarity estimation. MFCC stands for Mel-frequency cepstral coefficients. It is a feature extraction technique used in many machine learning models trained on audio data. To evaluate the performance of the real-life audio signals noises are added into them. In real life, the audio signal usually contains noise which is the unwanted sound. Also, it is observed that audios are stretched a bit therefore in the dataset stretched audios are introduced which makes it an important feature for predicting emotions from audios.

3.5 Model Training and Testing

In this research, three different types of deep learning algorithms are used for analysis. The data fed in the algorithm was in the form of audio. where for each algorithm, the 70 percent of the data samples were used for training purposes and the remaining 30 percent of the data samples were used as a testing set. Deep Neural Networks(DNN) Convolution Neural network (CNN) and a custom model are used and each model is trained over google colab with the epoch value as 50.

3.6 Model Evaluation

To identify the most optimal model for predicting the emotions from audio signals evaluation of each model needs to be performed based on some metrics. Since this is a classification task the metrics will be utilized for the evaluation are accuracy, precision, recall, and F1 score, which will be calculated over the test dataset. The model with high accuracy, high PRF (Precision, Recall and F1-Score) with minimum loss will be considered as the optimal model for emotion classification based on audio data.

4 Design Specification

In order to develop audio based emotion recognition system, this research utilizes the three different models which are Artificial neural network (ANN), Convolutional neural network (CNN), and Custom model (with the combined layers of LSTM and BILSTM).

4.1 Deep Neural Network (DNN)

A deep neural network is made up of neurons and just like the human mind, a neural network uses these neurons to process information and present output like recognising

an object or solving an equation. A neural network breaks the input information into tiny pieces stores it in neurons now these neurons processes this information to generate something useful out of it. If you have to break a deep neural network it can be divided into 3 parts the input layer, the hidden layer or the processing layer, where all the processing happens and the output layer, where the result is presented.

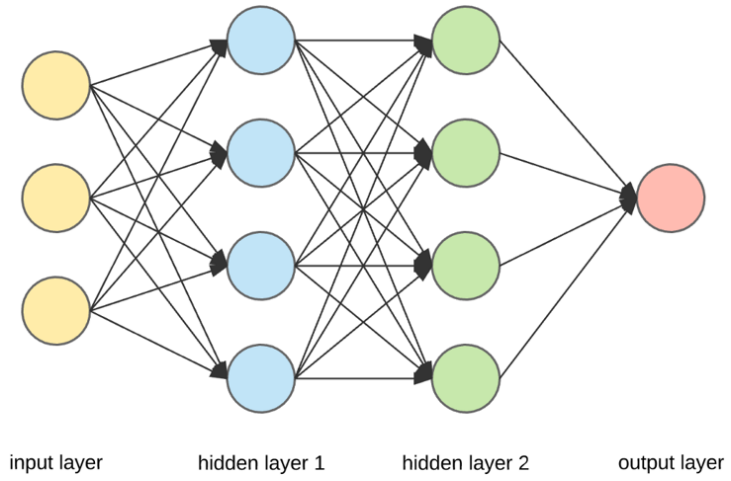

Figure 5: Deep Neural Network

4.1.1 Proposed Deep Neural Network (DNN) Model Architecture

In our model we trained a neural network that is a sequential model, it consists of 4 layers, the first layer is a dense layer with 128 neurons using 'relu' as the activation value and also we applied a batch normalization function at the end of it. The second layer is again a dense layer but here there are 256 neurons and 'relu' is used as the activation value. There is a batch normalization function at the end of it and also a dropout function with a parameter value set to 0.2. The third layer is similar to the first one as there is no dropout function. Here there are 512 neurons with the activation keyword set to 'relu'. The fourth and the last layer used in the neural network model uses dense function with 256 neurons, activation value set to 'relu', batch normalization function at the end of it and also a dropout function with parameter value set to 0.2. At the end of all this the output function is placed that has an activation value set to 'softmax'. Now, this model is compiled with the optimization parameter value set to 'adam' and the loss function value set to 'categorical-crossentropy'.

4.2 Convolutional Neural Network (CNN)

Convolutional neural network or CNN helps in image processing just like humans use the visual cortex for processing whatever the eyes see. A convolutional neural network breaks the image and assigns each broken part into the matrix in the form of sequential rows and columns and then uses this data, processes it and presents output. It uses the values of the matrix to add importance and biases to the image, this helps in recognising images and differentiating them from one another.

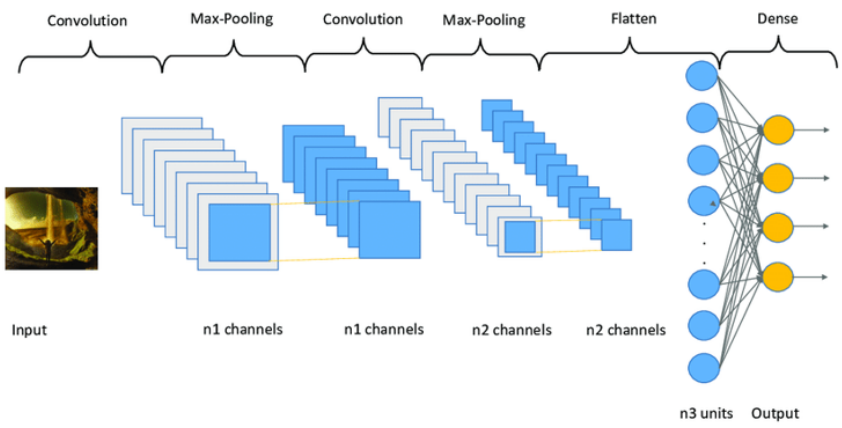

Figure 6: Convolutional Neural Network Architecture García-Ordás et al. (2020)

4.2.1 Proposed Convolutional Neural Network (CNN) Model Architecture

Our model has four convolutional blocks followed by a flattening layer, a dense layer and at last an output layer. The first block is a one-dimensional convolutional layer with 128 neurons, kernel size of 2 and the activation set to 'relu'. There is also a one-dimensional max pool layer with a pool size 2 and padding set to the same. The second block is again a one-dimensional convolutional layer with 256 neurons, kernel size is 2, padding set to same and activation value set to 'relu'. It has a max pool layer that is one dimensional and has a pool size of 2, its padding is also set to 'same'. In the end, it has a batch normalization layer. The third convolutional block has a one-dimensional convolutional layer with 256 neurons kernel size is 5, just like the previous blocks it has padding set to 'same' activation value set to 'relu'. Max pool layer with a pool size of 5, padding set to same and a dropout with the parameter value set to 0.2. Now the fourth block has a one-dimensional convolutional layer but when compared to the rest of the blocks has only 64 neurons with a kernel size of 5 strides set to 1, padding set to 'same', activation value passed is 'relu'. It has a max pool layer with a pool size of 5 and padding is also set to 'same', followed by a batch normalization layer. After all these blocks there is a flattening layer, a dense layer with 512 neurons and an output layer with activation set to 'softmax'. While compiling the model the optimization used is 'adam' and the loss value is set to 'categorical-crossentropy'.

4.3 Long Short Term Memory (LSTM) Neural Network

Long short term memory model focuses on storing important information and neglecting the rest of them. If you pass a sentence as input to the LSTM model it will remember only the words that are affecting the outcome and will neglect the rest of them. LSTM has cells that store information, now, if the algorithm thinks that the information passed is useful, it will be stored in one of those cells and if it's not useful it will be neglected by the algorithm. By doing this LSTM saves a lot of computational power and time. LSTMs are derived from RNNs stands for recurrent neural networks that suffer from the problem of short term memory.

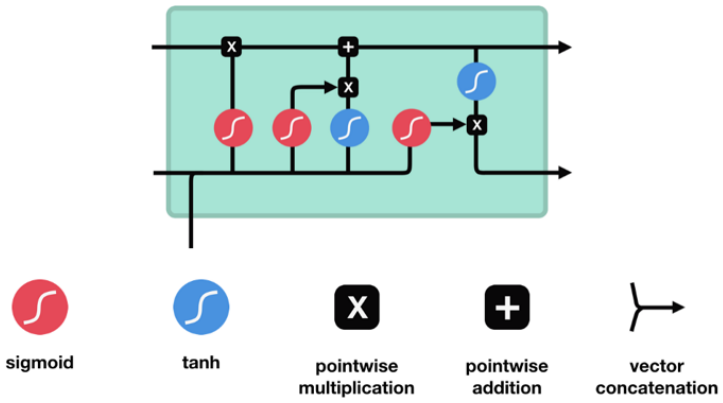

Figure 7: Long Short Term Memory Neural Network Architecture Phi (2020)

4.3.1 Proposed Custom Model Architecture

Our model is the 3rd and the last model and it is an advanced deep learning model. It is sequential with four layers only. The first layer is an LSTM layer with 32 neurons and activation set to 'tanh', it also has a bidirectional LSTM layer with only 8 neurons and activation set to 'tanh'. The second layer is a dense layer with 256 neurons and activation set to 'relu' followed by a batch normalization layer and a dropout layer with the value set to 0.2. The third layer is a dense layer with 512 neurons activation value set to 'relu', followed by a batch normalization function. The fourth and the last layer is very similar to the rest of the layers as it's a dense layer with 256 neurons activation value passed to be 'relu' and followed by a batch normalization function followed by a dropout function with the parameter value set to 0.2. In the end, there is a dense layer with activation set to 'softmax'. In compiling, the optimizer used is 'adam' and the loss function is 'categorical-crossentropy'.

5 Implementations

To implement deep learning models for predicting emotions, first preprocessing of dataset needs to be performed to extract out important features from audios. This dataset is then transformed into the required format which consists application of data scaling and shape change according to each implemented model. After acquiring the required dataset three deep learning models are implemented i.e., Deep Neural Network (DNN/ANN), Convolutional Neural Network (CNN), and a Custom Model, and the best model is selected on the basis of validation accuracy, Precision, Recall, and F1Score. All Models are optimized with ADAM optimization function and categorical cross-entropy as loss function is used. Each model is trained on a training dataset on 50 epochs. In order to implement the proposed work various libraries have been used which includes Librosa, pandas, NumPy, matplotlib, seaborn, Tensor Flow, Keras, etc. This whole experiment is done on the Google colab platform for training purposes with python as a programming language, After evaluating all models the most accurate model is dumped into .h5 file for further use.

6 Evaluation

Since classifying the emotions from audio data is a challenging task, the selection of optimal model becomes very important in order to achieve the better results. Therefore, in this research the performance of every model (Deep neural network, Convolution neural network and custom model) will be evaluated for audio data based on certain metrics such as Accuracy, Loss, Precision, Recall and F1-Score. In this research the custom model consists the combined layer of LSTM and Bi-LSTM. Further, the audio data from the different sources has been merged and splitted into training and test set. The ratio of training and testing has been defined as 7:3. For every model the number of epoch value is considered as 50.

6.1 Experiment 1 / Evaluation Based on Accuracy

Accuracy is a parameter that determines which model is best in recognising patterns and it also represents the relations between independent and dependent variables. In this experiment, Deep neural network model with four layers is implemented. For the Deep neural network the validation accuracy after the end of 50 epochs is found to be 70.43%. The other model deployed is a Convolutional neural network. For this model the validation accuracy after the end of 50 epochs is 64%. The third model used in this project is a custom model that consists of four layer. Custom model has achieved the validation accuracy after 50 epochs is 63.18%. The model having highest accuracy can be selected as the best model. The model comparison graph for accuracy is shown in figure 8.

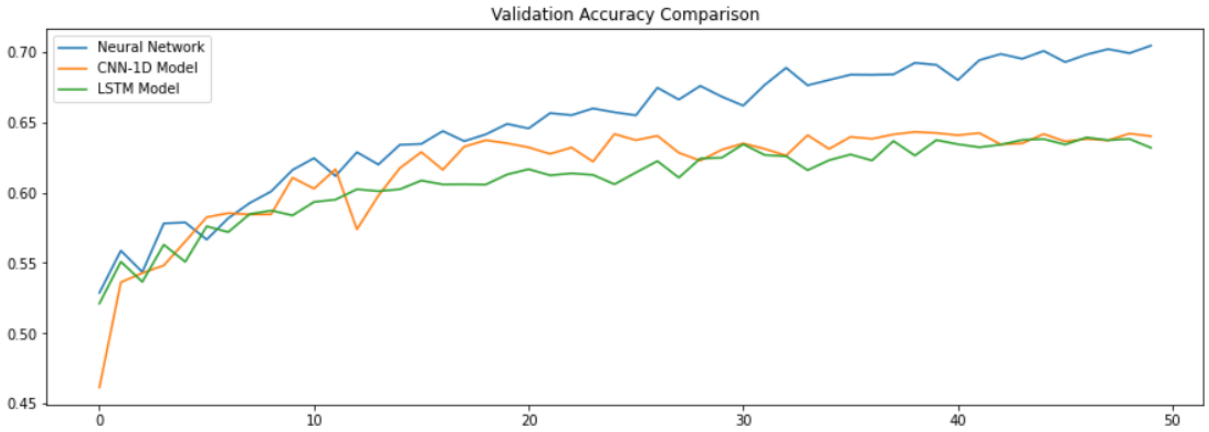

Figure 8: Comparison Of Different Models based on Accuracy

On observing the graph carefully it has been found that highest accuracy has been achieved using Neural network architecture, followed by CNN and Custom model architecture. We have achieved the highest accuracy of 70.43% using simple neural network architecture.

6.2 Experiment 2 / Evaluation Based on PRF Score

Precision is the measurement of how many times a model predicted something correctly out of all the possible outcomes. Recall is the measure of how many times the model

predicted something correctly out of actual correct predictions. Whereas, F1 score is the harmonic mean of precision and recall. For Deep neural network the values of Precision, recall and f1 scores are 0.7762, 0.6385 and 0.7028 respectively. For Convolutional neural network the values of Precision, Recall and f1 score are 0.6708, 0.6131 and 0.6399 respectively and for custom model the values of precision, recall and f1 score are 0.7149, 0.5568 and 0.6310 respectively for each metric. The Comparison of applied models based on PRF is shown in figure 9 , Figure 10 and Figure 11.

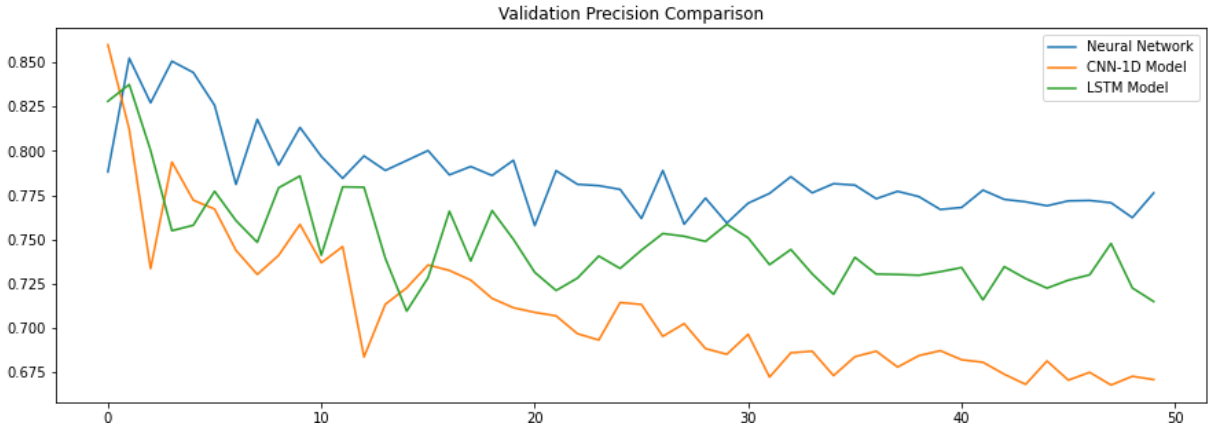

Figure 9: Comparison Of Different Models based on Validation Precision

After carefully comparing the precision score between all the algorithms, we have found that highest precision score has been obtained using Simple neural network architecture followed by Custom model and CNN architecture. In case of Precision, CNN model does not perform well. Whereas, in case of recall score the CNN model outperforms as compared to the custom model and with Deep neural network architecture the highest recall score of 0.6385 is achieved. On comparing the F1-Score, the results are found to similar as recall. In terms of F1-Score Deep neural network outperforms as compared to CNN and Custom model.

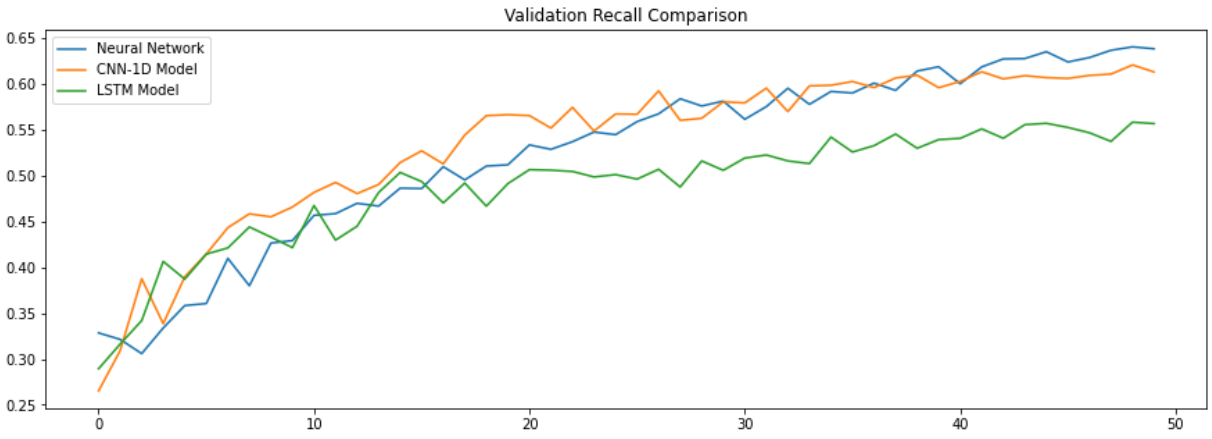

Figure 10: Comparison Of Different Models based on Validation Recall

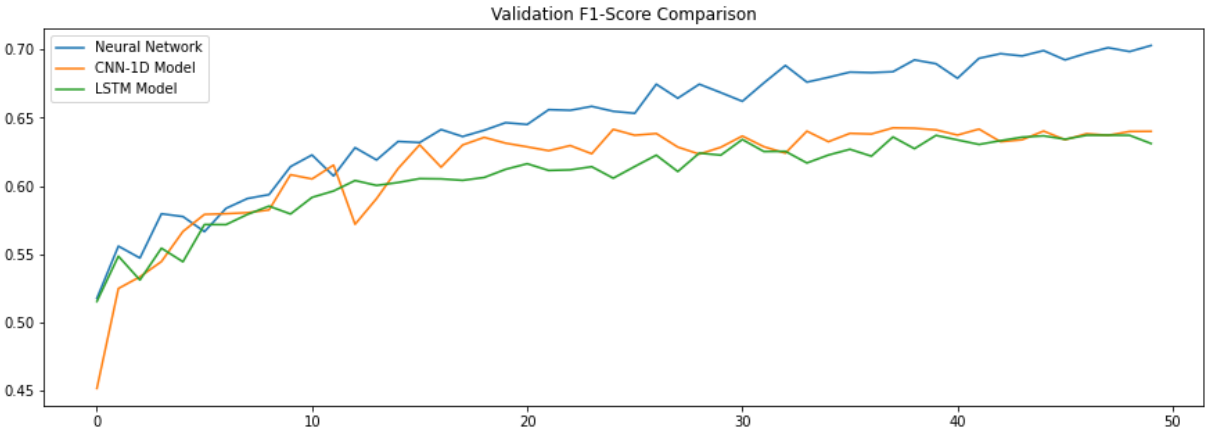

Figure 11: Comparison Of Different Models based on Validation F1-Score

6.3 Experiment 3 / Evaluation Based on the Validation Loss

Validation Loss is the loss calculated on the validation set. When the data set is split it forms 3 subsets that are training set, the validation set and the test set. For Deep neural network the value of validation loss is 0.8505. For Convolutional neural network the value of validation loss is 1.4108 and for Custom model the value of validation loss is 1.1248. The Comparison of applied models based on validation loss is shown in figure 12. A model with minimum validation loss is considered as the optimal model and can perform the better predictions.

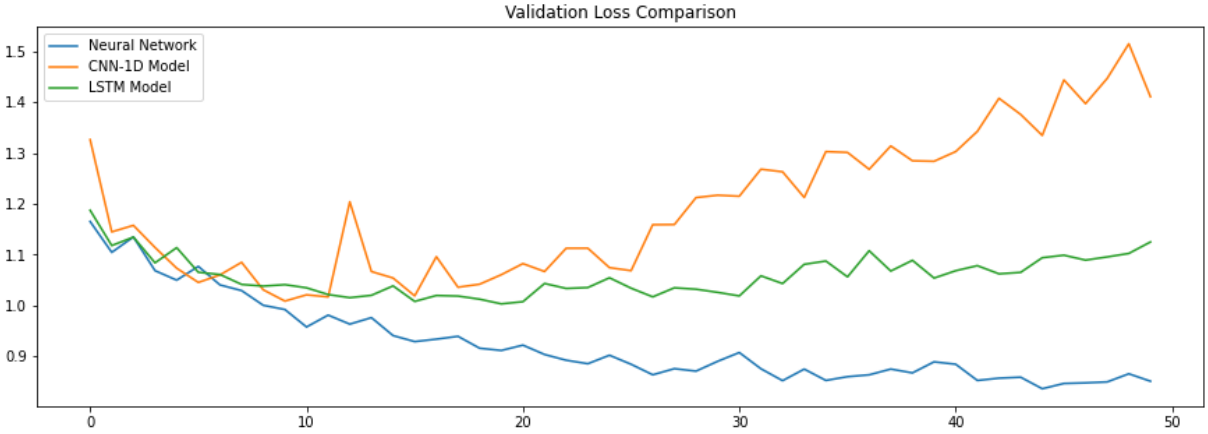

Figure 12: Comparison Of Different Models based on Validation loss

After analyzing the results, we have obtained the minimum loss using Deep neural network architecture, followed by custom model and CNN.

6.4 Discussion

After performing certain set of experiments to predict emotions from audio data, we can say that Deep neural network is an optimal model for predicting emotions from audio signals. The applied models are analysed based on validation accuracy, Precision,

recall and f1 score, so that we can make a comparison between the models. Deep neural network model has obtained highest accuracy score of 70.43%, with a recall score of 0.6385, precision score of 0.7762 and f1 score of 0.7028. For audio analysis, Librosa library utilized in this research. Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems and uses sound file and audio read to load audio files. Therefore, in this research the DNN which is a best performing model will be utilized further to predict the emotions from the different audio set. To prevent models from overfitting dropout layers have been used with a drop of value 0.2/0.3. The balanced dataset used in this analysis and research prevents the results from being biased towards one kind of feature/value. The diagonal of the confusion matrix informs about the Correct predictions. It is mainly used to give a insight about the mis-classified values. There are many classes in our model predictions, which has been misclassified. Where the highest number misclassified value is 193, where the actual emotion is angry and predicted outcome is happy. The lowest number of misclassified values are 6. Where the actual emotion sad is classified as angry. The obtained confusion matrix for deep neural network architecture over the test data is shown in Figure 13 . Furhter after the evaluation, the DNN model has been utilized to predict the emotion on unseen audio data, which have provided us the correct outcomes.

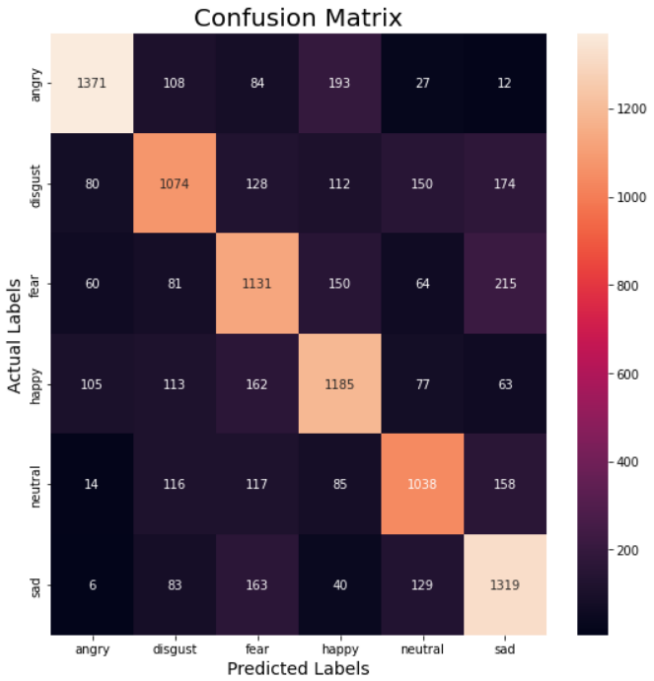

Figure 13: Confusion Matrix for DNN on Test Dataset

7 Conclusion

Identification of emotion from audio data is a challenging task and still an open area of research. In this research project, different deep learning models are implemented to recognize and predict emotions using audio signals. The field of emotion recognition and and analysis using audio is a vast and ever-evolving field of work, where different deep learning models and algorithms are being used to accomplish this task with utmost

accuracy. As our proposed framework is able correctly identify the human emotion from audio data, this can help us to build a speech emotion recognition system which can help the children suffering with autism specturm disorder. In this research, the multiple algorithm for emotion classification based on audio data are tested and evaluated. Where highest accuracy has been achieved using the simple Deep neural network architecture. In order to get more insight about the prediction of DNN, the confusion matix is calculated over the test data, which informs about the mis-classification. In the future scope, the more complex deep learning architectures can be utilized to improve the accuracy of the model. Some of the emotions in the current dataset, were very less in number which makes the data imbalanced. Due to which the model can generate the biased results. Therefore, more number of audio samples can be added to improve accuracy and predictions by models. The DNN model did not have any knowledge about what is being said in the audio or what actually is being said or what subject is actually being discussed. At one instance this can be an advantage, but RNN type networks can be used if we try to go with subject/context based approach.

References

Atmaja, B. T., Shirai, K. and Akagi, M. (2019). Speech emotion recognition using speech feature and word embedding, *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 519–523.

Bazgir, O., Mohammadi, Z. and Habibi, S. A. H. (2018). Emotion recognition with machine learning using eeg signals, *2018 25th National and 3rd International Iranian Conference on Biomedical Engineering (ICBME)*, pp. 1–5.

Chauhan, A., Sahu, J., Jaiswal, N., Kumar, K., Agarwal, A., Kaur, J., Singh, S. and Singh, M. (2019). Prevalence of autism spectrum disorder in indian children: A systematic review and meta-analysis, *Neurology India* **67**: 100.

CREMA-D (n.d.).
URL: *https://kaggle.com/ejlok1/cremad*

Gannouni, S., Aledaily, A., Belwafi, K. and Aboalsamh, H. (2021). Emotion detection using electroencephalography signals and a zero-time windowing-based epoch estimation and relevant electrode identification, *Scientific Reports* **11**(1): 7071.
URL: *http://www.nature.com/articles/s41598-021-86345-5*

García-Ordás, M., Benítez-Andrades, J., García, I., Benavides, C. and Alaiz Moreton, H. (2020). Detecting respiratory pathologies using convolutional neural networks and variational autoencoders for unbalancing data, *Sensors* **20**.

Haber, N., Voss, C., Nag, A., Tamura, S., Daniels, J., Ma, J., Chiang, B., Ramachandran, S., Ouillon, J., Winograd, T., Feinstein, C. and Wall, D. (2019). Towards continuous social phenotyping: Analyzing gaze patterns in an emotion recognition task for children with autism through wearable smart glasses (preprint), *Journal of Medical Internet Research* **22**.

Hodges, H., Fealko, C. and Soares, N. (2020). Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation, *Translational Pediatrics* **9**: S55–S65.

Huang, K.-Y., Wu, C.-H., Hong, Q.-B., Su, M.-H. and Zeng, Y.-R. (2018). Speech emotion recognition using convolutional neural network with audio word-based embedding, *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 265–269.

Kokkinaki, T., Matsuda, S., Chee, W., Lecciso, F., Levante, A., Fabio, R., Caprì, T., Leo, M., Carcagnì, P., Distante, C., Mazzeo, P. L., Spagnolo, P. and Petrocchi, S. (2021). Psychology for clinical settings, a section of the journal frontiers in psychology emotional expression in children with asd: A pre-study on a two-group pre-post-test design comparing robot-based and computer-based training, *Frontiers in Psychology* **12**: 678052.

Lim, J., Mountstephens, J. and Teo, J. (2020). Emotion recognition using eye-tracking: Taxonomy, review and current challenges, *Sensors* **20**: 2384.

Mellouk, W. and Wahida, H. (2020). Facial emotion recognition using deep learning: review and insights, *Procedia Computer Science* **175**: 689–694.

Murugan, H. (2020). Speech emotion recognition using cnn, *International Journal of Psychosocial Rehabilitation* **24**.

Mustaqeem, Sajjad, M. and Kwon, S. (2020). Clustering based speech emotion recognition by incorporating learned features and deep bilstm, *IEEE Access* **PP**: 1–1.

Naik, N. and Mehta, M. A. (2018). Hand-over-face gesture based facial emotion recognition using deep learning, *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)*, pp. 1–7.

Phi, M. (2020). Illustrated Guide to LSTM's and GRU's: A step by step explanation. **URL:** *https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21*

Piana, S., Staglianò, A., Odone, F. and Camurri, A. (2016). Adaptive body gesture representation for automatic emotion recognition, *ACM Transactions on Interactive Intelligent Systems* **6**: 1–31.

RAVDESS Emotional speech audio (n.d.). **URL:** *https://kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio*

Stedman, A., Taylor, B., Erard, M., Peura, C. and Siegel, M. (2019). Are children severely affected by autism spectrum disorder underrepresented in treatment studies? an analysis of the literature, *Journal of Autism and Developmental Disorders* **49**.

Surrey Audio-Visual Expressed Emotion (SAVEE) (n.d.). **URL:** *https://kaggle.com/ejlok1/surrey-audiovisual-expressed-emotion-savee*

Tarnowski, P., Kołodziej, M., Majkowski, A. and Rak, R. (2020). Eye-tracking analysis for emotion recognition, *Computational Intelligence and Neuroscience* **2020**: 1–13.

Topic, A. and Russo, M. (2021). Emotion recognition based on eeg feature maps through deep learning network, *Engineering Science and Technology an International Journal* **24**.

Toronto emotional speech set (TESS) (n.d.).
URL: *https://kaggle.com/ejlok1/toronto-emotional-speech-set-tess*

Valles, D. and Matin, R. (2021). An audio processing approach using ensemble learning for speech-emotion recognition for children with asd, *2021 IEEE World AI IoT Congress (AIIoT)*, pp. 0055–0061.

Wang, J. and Wang, M. (2021). Review of the emotional feature extraction and classification using eeg signals, *Cognitive Robotics* **1**.

Wu, J., Zhang, Y., Zhao, X. and Gao, W. (2020). A Generalized Zero-Shot Framework for Emotion Recognition from Body Gestures, *arXiv:2010.06362 [cs]* . arXiv: 2010.06362.
URL: *http://arxiv.org/abs/2010.06362*

Yao, Z., Wang, Z., Liu, W., Liu, Y. and Pan, J. (2020). Speech emotion recognition using fusion of three multi-task learning-based classifiers: Hsf-dnn, ms-cnn and lld-rnn, *Speech Communication* **120**.

Yoon, S., Byun, S. and Jung, K. (2018). Multimodal Speech Emotion Recognition Using Audio and Text, *arXiv:1810.04635 [cs]* . arXiv: 1810.04635.
URL: *http://arxiv.org/abs/1810.04635*

Zheng, L., Mountstephens, J. and Teo, J. (2020). Four-class emotion classification in virtual reality using pupillometry, *Journal of Big Data* **7**.