A Project Report

On

Emotion Recognition in Speech Using Machine Learning

Submitted
to

Amity University Uttar Pradesh

in partial fulfilment of the requirements for the award of the degree
of

Bachelor of Technology
in
Computer Science & Engineering
By

GROUP 79
Akshay Kumar (A2305218221)
&
Aditya Chandrayan (A12405218070)

under the guidance of

Dr. Sanjay Kumar Dubey
(Associate Professor)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY
AMITY UNIVERSITY UTTAR PRADESH
NOIDA (U.P.)
2022

DECLARATION

We, Akshay Kumar and Aditya Chandrayan, students of B.Tech (CSE) hereby declare that the project entitled "**Emotion Recognition in speech Using Machine Learning**", which is submitted by me to the Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, in partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Computer Science & Engineering has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition

Place: NOIDA
Date:

Akshay Kumar
(A2305218221)

Aditya Chandrayan
(A12405218070)

CERTIFICATE

On the basis of the declaration submitted by Akshay Kumar and Aditya Chandrayan, students of B.Tech (CSE), We hereby certify that the project titled "**Emotion Recognition in speech Using Machine learning**", which is submitted to the Department of Computer Science & Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, in partial fulfilment of the award of the degree of Bachelor of Technology in Computer Science & Engineering, is an original contribution with existing knowledge and faithful record of work carried out by them under my guidance and supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: NOIDA
Date:

Dr. Sanjay Kumar Dubey
Associate Professor
Dept. of Computer Science & Engineering
Amity School of Engineering and Technology
Amity University Uttar Pradesh

ACKNOWLEDEGEMENT

We take this opportunity to express our profound gratitude and deep regards to my supervisor Dr. Sanjay Kumar Dubey for his exemplary guidance, monitoring, and constant encouragement throughout this project. His invaluable comments on system modelling have always inspired me to refine my work. The blessing, help and guidance by him time to time shall carry us a long way in the journey of life on which us about to embark.

We would also like to thank to our Head of Institute Prof. (Dr.) Abhay Bansal, and all the staff of department of Computer Science & Engineering for their regular encouragement and endeavour's.

We are grateful to Amity University, Noida, Uttar Pradesh, for its time and support, which helped us in this project through various stages. Lastly, We would like to thank almighty, our parents, friends for their constant encouragement without which this assignment would not be possible.

.

Akshay Kumar
A2305218221
Aditya Chandrayan
A12405218070

ABSTRACT

Human when communicate they use emotion to convey their message to each other clearly so we can say that emotion is equally important as the content said by humans. The role of emotion in the context which defines the message led to the demand of emotion recognition system and expansion that idea led to demand of human computer interaction system. The advances that were made in Machine learning field lead by many researchers and the requirement of real time speech emotion recognition system for the human computer interaction led to the creation of many datasets and methods to achieve and better the accuracy or existing solution for emotion recognition in speech. Human emotion is very complex in nature and due to which, what feature we select can have major impact on our machine learning model outcome. My work led me to explore many machine learning methods and feature extraction methods for creation of speech emotion recognition system. In this paper we are going to understand and see the approaches discovered by me throughout our major project.

Table of Contents

List of Tables

List of Figures

CHAPTER 1
INTRODUCTION

1.1 Introduction

A great deal of multidisciplinary study has been done which went into understanding speech emotion detection throughout the years. Researchers in the neuroscience field study how the brain receives incoming stimuli by analysing raw audio data. Linguists would rather observe speech that can supply emotional information through semantic and syntactic analysis of the speech, but computational intelligence researchers provide tools to explain the knowledge received from neuroscientists in a mathematical solution. These multidisciplinary research collaborations have showed enormous promise in comprehending voice emotion.

The starting of SER was from psychology not from computer science field as some scientist researched the role of acoustics of human emotion, that is what is the effect of emotion upon the voice which is understood by other people. It was found that the most primitive emotion example love, and anger can be understood by most primitive people which is also shared with animals alike. The language of tone is most common and oldest in our civilization. The idea for letting computer understand the emotion was lead two decades before at the same time when machine learning was also being popular, and many people was discovering many new machine learning methods. Now after two decade of research we can say that speech emotion recognition is matured enough so that major advancement can be done in this area. In today's world there are lot data created by humans which can help us to create good dataset which can used to train model in deep learning. Our report is based on speech emotion recognition that is recognition of emotion by computer and to put it more accurately we can say that enabling computers to recognise emotion from the acoustic characteristics from the speech like pitch, tone, loudness, and spectral distribution of frequency. However, recognising emotions from speech is a difficult process. First and foremost, even for the same feeling, there is unlimited diversity in emotional responses inside and across speakers. The diversity of emotions is another aspect. People may shade or conceal their emotions because they are complex or because of societal factors. Information must be distinguished from other impacts on the voice, such as vocal abnormalities, breathing sound and physical ambiance sound.

There is also a high demand for Voise based system, a computer system which uses intelligence system to mimic emotion and read the content. For computer to understand the emotion and to mimic is quite hard to build so much research is going on in this field which can be also helper or reinforced by emotion recognition system.

Speech emotion detection is challenging due to the nature of emotion and complexity build by people around the world and their culture. Which is solved by using machine learning method to classify the emotion in speech. What has been accomplished thus far is the recognition of clearly expressed performed emotions with excellent accuracy, but still real-world dataset is not created and still we are not able to identify emotion in real time in social environment. Prediction accuracies are based not only on emotion types, but also on the number of emotion classes: from performed emotion, up to eight emotion classes have been effectively identified, but from natural emotions, only five classes can be meaningfully categorized, and even these with accuracies below 50%. The proposed approach has been proven to work on both men and women as well as children. However, thus far, technology has mostly been used in laboratory settings with a peaceful background and a solitary clearly loud s

1.2 Objective

The main goal and objective for the project undertaken are to create speech Emotion Recognition (SER) to aid human to machine interaction with the help of recently advanced machine learning techniques. Machine learning methods has improved very fast in a short period. With present deep learning methods, we can train our model with huge dataset and implement data augmentation techniques or multi-layer models to increase the accuracy of our model. Also, to provides for simple application integration so we can widen the use of speech emotion recognition application in the world.

1.3 Methodology

The methodology that we are going to adapt will use various neural network methods or model for classification and compare the output or accuracy of different neural network model. This section provides a summary of the proposed network architecture, which is made up of two basic building blocks: feature extraction and classification. The tools and datasets which were applied to train the network are then revealed. We also describe how and why the dataset is pre-processed.

Finally, the chosen model parameters are discussed, as well as the reason behind them and their values, as well as a general description of the analysis utilized.

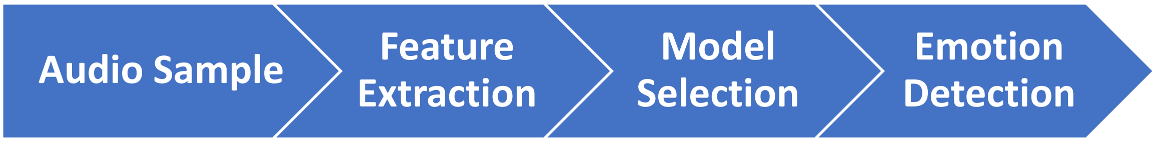


Figure 1.1 Data Flow Diagram

The work done by me in major project is to select the features, which is best to identify the pitch, tone, jitter, loudness etc. and to create a machine learning model then improve it furthermore for better accuracy.

1.4 Relevant question

- Audio segmentation to discover units for analysis, extraction of emotion-relevant information, and classification are all part of the spoken emotion recognition methodology. In the meanwhile, it may be required to sacrifice psychological validity in the case of acoustic units. Since several research have been done with acted emotions in the past, it should be investigated if acted emotions are still relevant. Whether the same features applicable to both, then are good replacements for genuine emotions.
- Afterall obviously, there are many more ways to identify emotions beyond acoustic components, such as word contents, facial expressions, and so on. A holistic approach is equally attractive for automatic emotion identification since people use all available information.
- Evidently, one feature-based emotion detection focused solely on audio file cannot produce good results close to humans, who utilize many more evidence. As a result, multiple feature-based recognition research works have been carried out to see if and how much of each modality contributes additional information to the acoustic properties.
- Hopefully, there will be more constructive applications for the SER models than negative ones. Strengthening essential processes such as emergency phone calls, understanding the emotions of persons contacting, and delivering immediate assistance are examples of good applications.
- The models must be developed in a method that is respectful of race, personal integrity, age, and gender. When gathering dataset for training, it's crucial that all participants understand the project's objectives and scope, as well as range to the use

of their data and models. While using SER in real-world applications, it's also important to be aware of and emphasise the potential of bias when training models on small and biased datasets.

1.5 Importance

Emotion detection advances have a favourable influence on a wide range of applications. Psychology, psychiatry, and neuroscience are among the scientific fields that benefit from automating the emotion detection approach. These cognitive science departments focus on human interaction, in which the topic of research is placed through a series of questions and scenarios, and many conclusions are drawn based on their emotions and replies. Few people are labelled as introverts and are hesitant to communicate, which might be a disadvantage. Conversation is essential for self-expression. Humans communicate efficiently by using the majority of their body and their voice. Hand gestures, body language, tone, and temperament are all employed to communicate one's feelings. Though the linguistic component of communication changes depending on the languages spoken throughout the world, the non-verbal element is the display of emotion, which is most likely universal. As a result, any advanced technology designed to provide a social environment experience must also have the ability to recognise emotional context in speech. The findings of this study will be useful to market participants interested in developing end-to-end SER models or improving the performance of current models. Building models for video labelling or metadata creation are two such applications. The findings might be used to decide on a method for including audio as a technique to make use of all available information in video, either as an independent model or, more likely, as part of an ensemble model with other modalities.

CHAPTER 2
LITERATURE REVIEW

This paper [1] proposed discourse feeling acknowledgment in light of the 2D-CNN model with Log-Mel spectrogram as the information. The proposed technique for the 2D-CNN model with Log-Mel spectrogram had the option to catch the huge discourse signals. The analyses utilized the EMODB information base to assess the proposed model with increased information. The proposed model had the option to deliver higher exactness of discourse feeling than the current models, which additionally involved profound learning strategies for perceiving feelings. The exploratory outcomes showed that the proposed model accomplished a precision of 0.88.

This paper [2] is to perceive feelings in discourse and arrange them in 7 inclination yield classes to be specific, outrage, fatigue, disdain, uneasiness, bliss, misery and impartial. The proposed approach depends on the Mel Frequency Cepstral coefficients (MFCC) and energy of the discourse signals as component information sources and uses Berlin data set of close to home discourse. The elements separated from discourse are changed over into a component vector, which thus are utilized to prepare different arrangement calculations specifically, Support Vector Machine (SVM), Random Decision Forest and Gradient Boosting. Irregular timberland was found to have the most noteworthy exactness and anticipated right feeling 81.05% of the time.

In this paper [3] the TESS dataset that we considered is adjusted. Since it has quiet information, it was simple as far as we were concerned to characterize and highlight the information. The classifier with most extreme precision, AUC, F1 score, kappa, MCC is Random Forest Classifier. The classifier with the least precision and the wide range of various terms is the choice tree classifier. We additionally referenced why the choice tree classifier has low accuracy and the irregular woods classifier has high exactness. Different classifiers that we investigated, i.e., additional trees classifier, light slope supporting machine, multi perceptron classifier, angle helping classifier, have the mid qualities in exactness and different qualities. Taking everything into account, the arbitrary backwoods classifier is the most reliable calculation and can be utilized, in actuality, situations to distinguish an individual's feelings through his discourse.

In this article [5], the best result of recognition rate was 90.05 %, achieved by combining the MFCC and M In this article, the best result of recognition rate was 90.05 %, achieved by combining the MFCC and M In this article, the best result of recognition rate was 90.05 %, achieved by combining the MFCC and M In this article, the best result of recognition rate was 90.05 %, achieved by combining the MFCC and M In this article, the best result of recognition rate was 90.05 %, achieved by combining the MFCC and M In this article, the best result of recognition rate was 90.05 %, achieved by combining the MFCC and M In this article, the best result of recognition rate was 90.05 %, achieved by combining the MFCC and M

In this article [4], the best consequence of acknowledgment rate 90.05 %, accomplished by coming the MFCC and MS Features and for the RNN model in the Spanish close to home data set. In addition, higher precision can be gotten utilizing the blend of additional highlights. Aside from this, looking for vigorous element portrayal is additionally considered as a component of the continuous exploration, as well as effective characterization methods for programmed discourse feeling acknowledgment

In this paper [5] we think about the exhibition of two classes of models. For both, we remove eight hand-created highlights from the sound sign. In the primary methodology, the separated elements are utilized to prepare six customary AI classifiers, while the subsequent methodology depends on profound learning wherein a standard feed-forward brain organization and a LSTM-based classifier are prepared over similar highlights. To determine vagueness in correspondence, we additionally incorporate highlights from the text area. We report exactness, f-score, accuracy, and review for the different analysis settings we assessed our models in. By and large, we show that lighter AI based models prepared north of a couple of hand-made highlights can accomplish execution tantamount to the ongoing profound learning based cutting-edge technique for feeling acknowledgment.

In this paper [6] the concentrate just examined a basic thing where the information expansion network is personal vectors removed by OpenSMILE. Be that as it may, the recommended model actually has a few issues. For instance, the created tests are like the examples utilized in the preparation organization and follow them. Accordingly, in the event that the test information circulation is not the same as preparing information appropriation, the increased information won't be useful. Future explores are necessities to utilize a strategy to sum up this technique.

In this paper [7] Rather than utilizing the normal cross-entropy blunder to prepare generative antagonistic organizations and to eliminate the evaporating inclination issue, Wasserstein Divergence has been utilized to deliver top notch fake examples. The recommended network has been tried to be applied for discourse feeling acknowledgment involving EMODB as preparing, testing, and assessing sets, and the nature of counterfeit information assessed utilizing two Support Vector Machine (SVM) and Deep Neural Network (DNN) classifiers. Additionally, it has been uncovered that extricating and recreating significant level highlights from acoustic elements, discourse feeling acknowledgment with isolating five essential feelings has been finished with satisfactory precision.

This paper [8] proposes a feeling acknowledgment framework in view of discourse signals in two-stage approach, to be specific element extraction and grouping motor. First and foremost, two arrangements of component are researched which are: the first, we extricate a 42-layered vector of sound highlights including 39 coefficients of Mel Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Harmonic to Noise Rate (HNR) and Teager Energy Operator (TEO). Furthermore, the subsequent one, we propose the utilization of the strategy Auto-Encoder for the choice of appropriate boundaries from the boundaries recently removed. Besides, we utilize the Support Vector Machines (SVM) as a classifier strategy. Tests are directed on the Ryerson Multimedia Laboratory (RML).

In this paper [9], the creator explained on causing machines to hear our feelings from one finish to another from the early investigations on acoustic corresponds of emotion8,16,21,34 to the first patent41 in 1978, the principal fundamental paper in the field,10 to the primary start to finish learning system.37 We are as yet learning. In light of this advancement, a preoccupied outline is displayed in Figure 2 introducing the principal elements of a cutting-edge motor. Ideally, current impasses, for example, the absence of rich measures of unconstrained information that take into account adapting to speaker variety, can be survived. After over 20 years into programmed acknowledgment of feeling in the discourse signal, we are presently seeing energizing seasons of progress: information learned highlights, orchestrated preparing material, all-encompassing structures, and learning in an undeniably independent way all of which can be anticipated to before long prompt the ascent of expansive everyday use in numerous wellbeing, recovery, security, and further gainful use-cases close by following quite a while of waiting36-the coming of sincerely clever discourse interfaces.

In this paper [10], a brain network is utilized for recognizing feelings in discourse with Toronto Emotional Speech Set (TESS), Surrey Audio-Visual Expressed Emotion (SAVEE), Ryerson Audio-Visual Database of Emotional Speech and Song(RAVDESS) datasets. The fundamental point is to zero in on perceiving the quiet trademark and the valuable highlights of sound signs. Convolutional brain networks is utilized for removing the significant elements and decrease the ghostly variety that is in signs and one of the brain procedures as repetitive brain networks for learning the worldly connections and tracking down the distinction of letters in order. Along these lines, correctness's determined by this model utilizing the relu enactment work bitterly affect our datasets.

In this paper [11], we apply multiscale region consideration in a profound convolutional brain organization to go to close to home attributes with changed granularities and in this way the classifier can profit from an outfit of considerations with various scales. To manage information sparsity, we direct information increase with vocal parcel length annoyance (VTLP) to further develop the speculation ability of the classifier. Tests are completed on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. We accomplished 79.34% weighted exactness (WA) and 77.54% unweighted precision (UA), which, as far as we could possibly know, is the best in class on this dataset.

In this paper [12] TensorFlow implementation of Convolutional Recurrent Neural Networks for speech emotion recognition (SER) on the IEMOCAP database. In order to address the problem of the uncertainty of frame emotional labels, we perform three pooling strategies (max-pooling, mean-pooling and attention-based weighted-pooling) to produce utterance-level features for SER. These codes have only been tested on ubuntu 16.04(x64), python2.7, cuda-8.0, cudnn-6.0 with a GTX-1080 GPU.

In this paper [13] investigates how to use a DCNN to connect the emotional hole in discourse signals. To this end, we first concentrate three channels of log Mel-spectrograms (static, delta, and delta) like the red, green, blue (RGB) picture portrayal as the DCNN input. Then, at that point, the AlexNet DCNN model pretrained on the huge ImageNet dataset is utilized to learn undeniable level component portrayals on each section separated from an expression. The learned section level highlights are totaled by a discriminant worldly pyramid coordinating (DTPM) methodology. DTPM joins transient pyramid coordinating and ideal Lp-standard pooling to shape a worldwide expression level component portrayal, trailed by the straight help vector machines for feeling characterization. Trial results on four public datasets, that is to say, EMO-DB, RML, eNTERFACE05, and BAUM-1s, show the promising exhibition of our DCNN model and the DTPM technique. Another intriguing finding is that the DCNN model pretrained for picture applications performs sensibly great in emotional discourse include extraction. Further tweaking on the objective profound discourse datasets considerably advances acknowledgment execution.

In this paper [14] a profound convolution brain networks-based approach has been created to gain the powerful highlights for discourse feeling acknowledgment from sound spectrogram information, which is further log-changed and handled utilizing guideline part examination (PCA) brightening. As needs be, a discourse feeling acknowledgment calculation named as PCA-DCNNs-SER is proposed. Starter tests have been led to assess the exhibition of PCA-DCNNs-SER on the IEMOCAP data set. Results show that our proposed PCADCNNs-SER (containing 2 convolution and 2 pooling layers) can acquire around 40% order exactness, which beats the SVM based SER utilizing hand-created highlights. Moreover, the PCA brightening cycle of spectrogram ends up being powerful and support the presentation of DCNNs based SER. Our future work is to join highlight extraction and setting gaining for better feeling forecast from sound information of variable length with an enormous information.

In this paper [15] We proposed to use a DNN to gauge feeling states for every discourse fragment in an expression, develop an utterance level include from section level assessments, and afterward utilize an ELM to perceive the feelings for the expression. Our exploratory outcomes demonstrate that this approach considerably helps the presentation of feeling acknowledgment from discourse signs and it is exceptionally encouraging to utilize brain organizations to gain close to home data from low-level acoustic elements.

In this paper [16] another strategy is proposed utilizing dispersed Convolution Neural Networks (CNN) to naturally gain influence remarkable elements from crude ghastly data, and afterward applying Bidirectional Recurrent Neural Network (BRNN) to get the worldly data from the result of CNN. Eventually, an Attention Mechanism is executed on the result grouping of the BRNN to zero in on track feeling relevant pieces of an expression. This consideration component further develops the arrangement exactness, yet in addition gives model's interpretability. Exploratory outcomes demonstrate the way that this approach can acquire 64.08% weighted exactness and 56.41% unweighted precision for four-feeling grouping in IEMOCAP dataset, which outflank past outcomes detailed for this dataset.

In this paper [17] we propose another technique for SER in light of Deep Convolution Neural Network (DCNN) and Bidirectional Long Short-Term Memory with Attention (BLSTMwA) model (DCNN-BLSTMwA). We first preprocess the discourse tests by information improvement and datasets adjusting. Also, we separate three-channel of log Mel-spectrograms (static, delta, and delta) as DCNN input. Then, at that point, the DCNN model pre-prepared on ImageNet dataset is applied to produce the fragment level highlights. We stack these elements of a sentence into expression level highlights. Then, we embrace BLSTM to get familiar with the significant level close to home highlights for fleeting synopsis, trailed by a consideration layer which can zero in on genuinely pertinent elements. At last, the learned significant level close to home highlights are taken care of into the Deep Neural Network (DNN) to foresee the last inclination. Probes EMO-DB and IEMOCAP information base get the unweighted normal review (UAR) of 87.86 and 68.50%, separately, which are better compared to most famous SER strategies and show the viability of our propose technique.

In this paper [18] The review expects to recognize feelings in the human voice utilizing man-made brainpower strategies. One of the main prerequisites of man-made brainpower works is information. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) open-source dataset was utilized in the review. The RAVDESS dataset contains in excess of 2000 information recorded as discourses and melodies by 24 entertainers. Information was gathered for eight distinct mind-sets from the entertainers. It was pointed toward identifying eight unique inclination classes, including impartial, quiet, blissful, miserable, furious, unfortunate, sickened, and shocked states of mind. The multi-facet perceptron (MLP) classifier, a generally utilized managed learning calculation, was liked for characterization. The proposed model's exhibition was contrasted and that of comparative investigations, and the outcomes were assessed. A general precision of 81% was gotten for grouping eight distinct feelings by utilizing the proposed model on the RAVDESS dataset.

In this paper [19] Different examinations and studies about Emotion Recognition, Deep learning procedures utilized for perceiving the feelings are performed. It is important in future to have a framework like this with significantly more solid, which has vast conceivable outcomes in all fields. This undertaking endeavored to utilize initiation net for tackling feeling acknowledgment issue, different information bases have been investigated. Prepared my model utilizing TensorFlow. Precision pace of around 39% is accomplished. In future, ongoing feeling acknowledgment can be created utilizing a similar design.

In this paper [20] surveys the various methodologies took on to lessen the obstruction of close to home correspondence which are now in presence and what system they utilized at the same time. In this unique circumstance, we likewise present a methodology of utilizing the Recurrent Neural Network which is a piece of Deep learning calculations. The entire course of computerized frameworks which ceaselessly learn, adjust, and improve absent a lot of guidance is truly captivating. Our essential objective is to make a vigorous correspondence framework through innovations that empower machines to answer accurately and dependably to human voices and offer helpful and significant types of assistance in like manner. In this survey, a broad report is made on the different methodologies accessible for discourse feeling acknowledgment that has been done work now. All the model's and exactness perspectives are thought about and are handed-off as per it.

In this paper [21] the creator talks about the ideas and late advances in discourse signal pressure, coding, and transmission, including mental discourse coding. To finish up, the primary goal of this article is to feature late accomplishments and difficulties in view of new AI standards that, over the course of the past 10 years, had a colossal effect in the field of discourse signal handling.

In this paper [22] The principal objective of this paper is to plan a reasonable model of Convolutional Neural Network (CNN) - Stride-based Convolutional Neural Network (SCNN) by taking fewer convolutional layers and wipe out the pooling-layers to increment computational dependability. This end will in general expand the exactness and reduction the computational season of the SER framework. Rather than pooling layers, profound steps have been utilized for the important aspect decrease. SCNN is prepared on spectrograms produced from the discourse signs of two distinct information bases, Berlin (Emo-DB) and IITKGP-SEHSC. Four feelings, furious, cheerful, impartial, and miserable, have been considered for the assessment interaction, and an approval exactness of 90.67% and 91.33% is accomplished for Emo-DB and IITKGPSEHSC, separately. This study gives new benchmarks to both datasets, showing the possibility and importance of the introduced SER strategy.

In this paper [23] outline level highlights are separated from every expression and, a long transient memory is utilized to gain proficiency with these elements outline by outline. Simultaneously, the deltas and delta-deltas of the log Mel-spectrogram are determined and remade into three channels (static, delta, and delta); these three-dimensional elements are advanced by a convolutional brain organization (CNN). Then, the two learned significant level elements are melded and bunch standardized. At long last, a SoftMax classifier is utilized to order feelings. Our PCRN model all the while processes two distinct kinds of elements in lined up with better gain proficiency with the unpretentious changes in feeling. The trial results on four public datasets show the prevalence of our proposed technique, which is superior to the past works.

In this paper [24] the RNNs have been utilized in numerous successive information handling undertakings to learn long haul conditions between the neighborhood highlights. A mix of these two provides us with the benefit of the qualities of the two organizations. In the proposed technique, CNN has been applied straightforwardly to a scalogram of discourse signals. Then, the consideration instrument based RNN model was utilized to learn long haul transient connections of the learned elements. Probes different information like RAVDESS, SAVEE, and Emo-DB show the viability of the proposed SER strategy.

In this paper [25] a Deep Convolutional-Recurrent Neural Network (Deep C-RNN) way to deal with characterize the viability of learning feeling varieties in the grouping stage. We utilize a combination of Mel-Gammatone channel in convolutional layers to initially separate undeniable level unearthly highlights then intermittent layers is embraced to gain the drawn out worldly setting from significant level elements. Additionally, the proposed work separates the feelings from unbiased discourse with appropriate parallel tree diagrammatic representations. The system of the proposed work is applied on a huge dataset covering Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset. At long last, the proposed results which acquired precision over 80% and have less misfortune are contrasted and the best in class draws near, and an exploratory outcome gives proof that combination results outflank in perceiving feelings from discourse signals.

In this paper [26] we research the commotion resistance of various acoustic highlights in distinctive different feelings by looking at the SER arrangement execution on clean discourse signals and uproarious discourse signals. We extricate the range and cepstral boundaries in light of human hear-able qualities and foster AI calculations to characterize four sorts of feelings utilizing these highlights. Trial results across the spotless and uproarious information show that contrasted with cepstral highlights, the hear-able spectrogram-based elements can accomplish higher acknowledgment precision for low sign to-clamor proportions (SNRs), however lower exactness for high SNRs. Gammatone channel cepstral coefficients (GFCCs) beat every one of the separated elements on the Berlin information base of close to home discourse (EmoDB), under each of the four sorts of tried clamor conditions. These outcomes show remuneration connections between here-able spectrogram-based highlights and cepstral highlights for SER with better commotion strength in certifiable applications.

In this paper [27] we examine sound sign denoising strategies in cepstral and log-ghostly areas and contrast them and normal executions of standard methods. The various methodologies are first analyzed for the most part utilizing arrived at the midpoint of acoustic distance measurements. They are then applied to programmed acknowledgment of unconstrained and regular feelings under reenacted cell phone recorded loud circumstances. Feeling acknowledgment is carried out as help vector relapse for constant esteemed forecast of excitement and valence on a sensible multimodal information base. In the examinations, the proposed strategies are found to outflank standard sound decrease calculations by and large.

In this paper [28] proposed technique utilizes formant consideration, commotion entryway filtering, and moving standardization settings to speed up and resilience to affliction. A bunch of syllable-level formant highlights is separated and taken care of into a solitary secret layer brain network that makes expectations for every syllable instead of the ordinary methodology of utilizing a modern profound student to make sentence-wide forecasts. The syllable level forecasts help to accomplish the constant inactivity and lower the totaled blunder in expression level cross-corpus expectations. The investigations on IEMOCAP (IE), MSP-Improv (MI), and RAVDESS (RA) information bases show that the technique files constant inactivity while foreseeing with cutting edge cross-corpus unweighted exactness of 47.6% for IE to MI and 56.2% for MI to IE.

In this paper [29] contrasted with other exemplary cross-corpus SER techniques, the significant benefit of the DDACNN model is that it can extricate hearty discourse highlights which are time-recurrence related by spectrograms and tight the errors between include dissemination of source corpus and target corpus to get better cross-corpus execution. Through a few cross-corpus SER tests, our DDACNN accomplished the cutting-edge execution on three public feeling discourse corpora and is demonstrated to proficiently deal with the cross-corpus SER issue. catchphrases: cross-corpus discourse feeling acknowledgment, profound convolutional brain organization, area transformation.

In this paper [30] the convolutional component of conventional DCNNs bit by bit become more unique, which may not be the best element for discourse feeling acknowledgment. Then again, the shallow element incorporates just worldwide data without the itemized data separated by more profound convolutional layers. As indicated by these perceptions, we plan a profound and shallow element combination convolutional network, which consolidates the component from various degrees of organization for discourse feeling acknowledgment. The proposed network permits us to take advantage of profound and shallow component completely. The famous Berlin informational collection is utilized in our examinations, the trial results demonstrate the way that our proposed organization can additionally further develop discourse feeling acknowledgment rate which exhibits the adequacy of the proposed network.

Chapter 3

BACKGROUND AND TECHNOLOGY USED

3.1 Basics of emotions in speech

To recognise emotions, one must first have a clear understanding of what they are, which would be a philosophy of emotions. In psychology, emotion theories have a lengthy history.

The categorical interpretation of emotions examines utterances that indicate clearly distinct moods and is based on the origins of language.

The concept of universal emotions, or feelings that are aroused in the same manner all throughout the world.

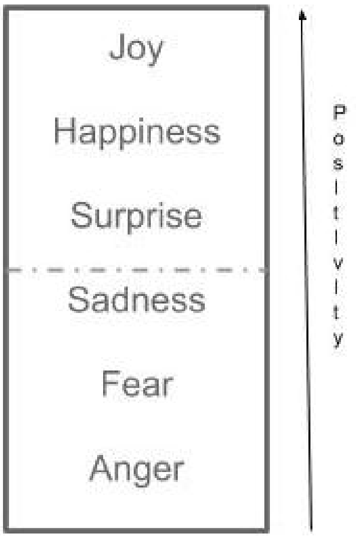


Figure 3.1 Universal emotions

Emotions are represented as points in an n-dimensional space in the dimensional interpretation of emotions, which is a continuous perspective.

3.2 PAD (Pleasure-Arousal-Dominance) model

Models in which an emotion is a point in an emotion space defined by two or three variables in a two-dimensional space, valence/evaluation/pleasure and arousal/activation are generally the variables; in a three-dimensional space, dominance, posture, power, or control are frequently included. On the one hand, because they allow for the simultaneous modelling of several aspects of emotions in a continuous range of values, they provide for a more fine-grained description. However, they simplify things since they just record the aspects stored in the dimensions. Differences between rage and fear, for example, are difficult to explain in a two-dimensional model because both are extremely negative and associated with high arousal. This judgment, called emotion space, is best handled by a three-dimensional model.

Basic emotions can still be represented as a point or region in the emotion space in a dimensional model.

In 2D space, just the valence and arousal dimensions are frequently employed to map out emotions.

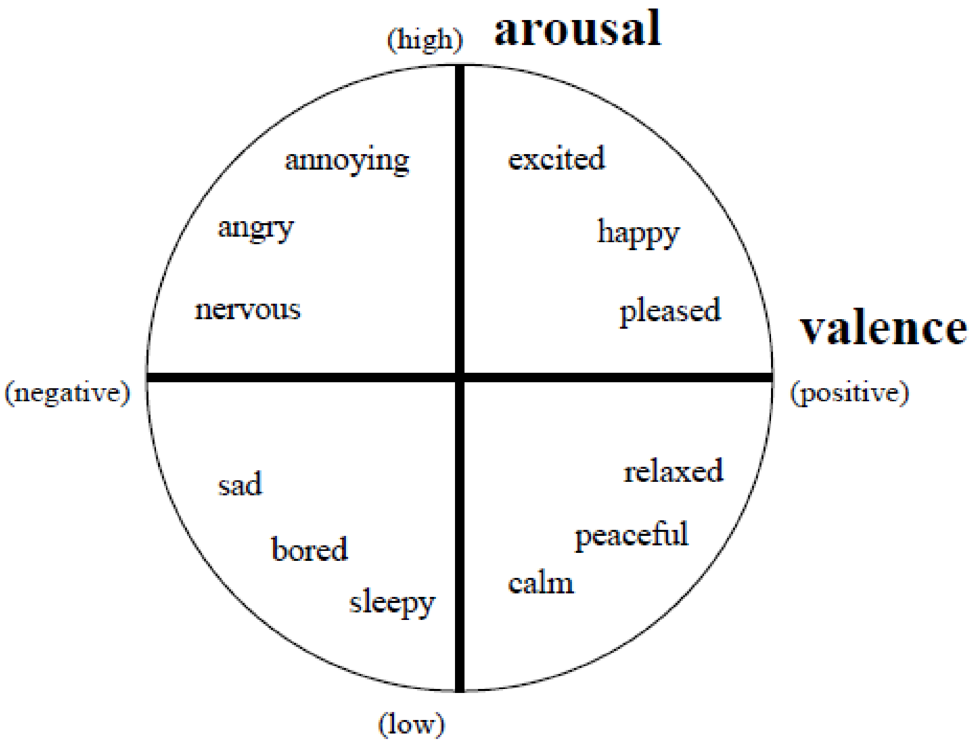


Figure 3.2

1D model of PAD (Pleasure-Arousal-Dominance) model

The "pleasure" dimension, also known as the "valence" dimension, indicates how pleasurable or painful something is. The "arousal" dimension represents one's level of activity or inactivity.

Finally, the third dimension, "dominance," describes how one feels controlled or in charge. Anger, for example, is a prevailing unpleasant emotion of relatively high intensity or arousal. Fear, on the other hand, is unpleasant and arousing, but it is a submissive emotion. Boredom is somewhat unpleasant, has a low arousal level, and is largely of a submissive nature.

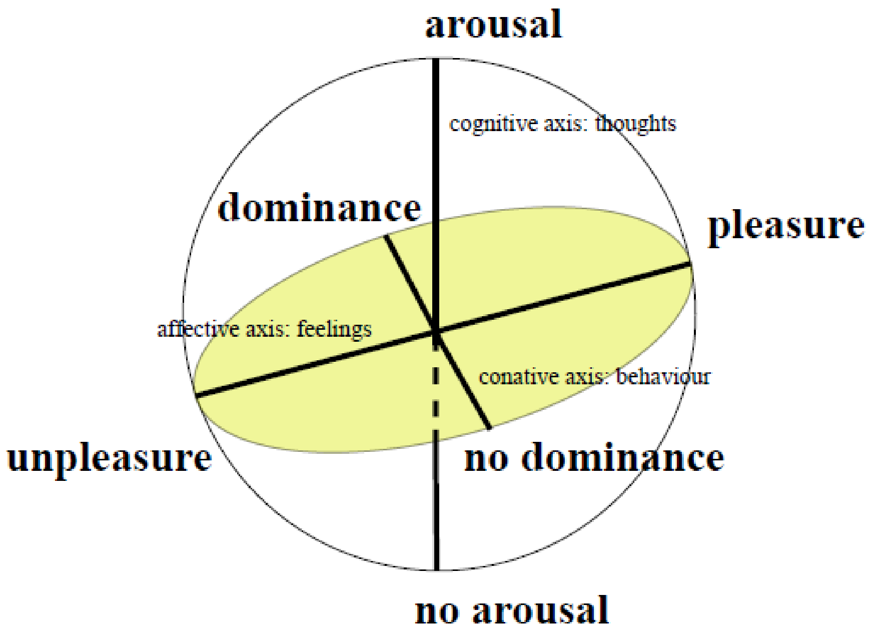


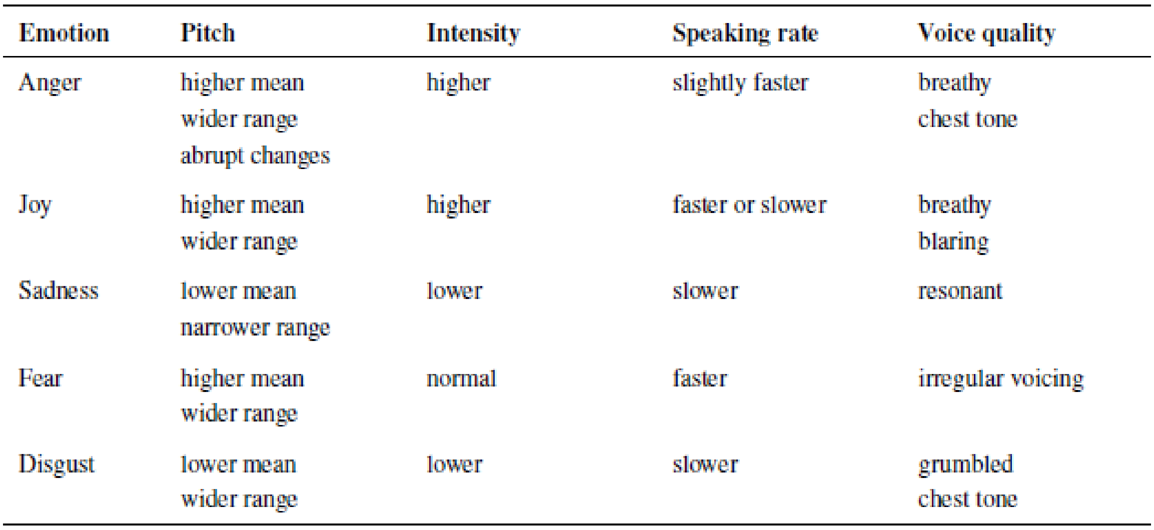
Figure 3.3

2D model of PAD (Pleasure-Arousal-Dominance) model

3.3 Language and speech

Emotional information is embedded in every part of language, including what we say and how we say or pronounce it, with the "how" being even more significant than the "what." The following factors can be examined at all levels of language, from pragmatics to phonetics and acoustics: To begin with pragmatics, a speaker's emotional state is significantly associated with his objective.

Beyond the literal meaning, the tone of a voice, or the phonetic and acoustic qualities of spoken language, is another indication of emotions. A cracking voice, for example, is indicative of excessive arousal. The major topic of this thesis is acoustic emotion identification, and the essential elements will now be addressed in greater depth. Prosody (pitch, intensity, speaking tempo) and voice quality are the vocal factors that have received the most attention in psychological studies and are also intuitively the most essential.



Emotion	Pitch	Intensity	Speaking rate	Voice quality
Anger	higher mean wider range abrupt changes	higher	slightly faster	breathy chest tone
Joy	higher mean wider range	higher	faster or slower	breathy blaring
Sadness	lower mean narrower range	lower	slower	resonant
Fear	higher mean wider range	normal	faster	irregular voicing
Disgust	lower mean wider range	lower	slower	grumbled chest tone

Figure 3.4

Emotion and language related properties

3.4 Speaker-Dependent vs. Speaker-Independent

There are two types of speech recognition systems: speaker-dependent and speaker-independent. Each intended user must train the speaker-dependent systems ahead of time. To recognise the utterance, the system must hear each vocabulary. This rapidly becomes impracticable in big groups of users or for specialised purposes; hence, speaker-independent solutions are employed in these situations. Instead, speaker-independent systems are trained on a huge number of utterances of all vocabulary items. This is best done by hundreds of individuals, which necessitates a lot of resources and effective data gathering methods.

Speaker-dependent if the speaker is recognised by the listener, i.e., there is prior information about the speaker that is used in the recognition process. Even for humans, emotion identification without the use of a speaker is tough.

3.5 Feature Extraction

Feature Extraction is one of the important parts as in this process we can change the audio file input if not careful, so we have to select those features which don’t change the properties of audio file itself. It is important to select right features those which extract all the property related to vocal characteristics of the speaker and store it.

Conversion from frequency (f) to mel scale (m) is given by

$$m=2595\cdot\log(1+f/700) \qquad (1)$$

3.5.1 Mel-Frequency Cepstrum Coefficients (MFCC)

MFCC's are formed from the cepstral representation of the audio clip. They are created by applying the fourier transform of the signal and translating it to the mel scale. The take the discrete cosine transform of the powers of the mel log powers to obtain the MFCC signal [7]

The following is the fundamental technique for creating MFCCs:

1. Hertz to Mel Scale Conversion
2. Take the logarithm of Mel's audio representation.
3. Use Discrete Cosine Transformation on logarithmic magnitude.
4. This provides a spectrum over Mel frequencies rather than time, resulting in MFCCs.

3.5.2 Log-melspectogram

A Mel spectrogram is one that has been converted to the Mel scale.

A spectrogram is a visual representation of a signal's frequency spectrum, where the frequency spectrum is the frequency range that the signal covers. According to studies, people do not detect frequencies on a linear scale, thus the Mel scale matches how the human ear works. At lower frequencies, humans are better at perceiving differences than at higher frequencies.

3.6 Deep Learning

In contrast to natural intelligence displayed by people and animals, artificial intelligence is a vast field that encompasses nearly all intelligence demonstrated by computers. Machine learning is a subset of artificial intelligence that follows a somewhat more stringent set of criteria. Machine learning (ML) has emerged as a new paradigm in recent years, replacing rule-based systems in which computers are now responsible for finding patterns and establishing rules.

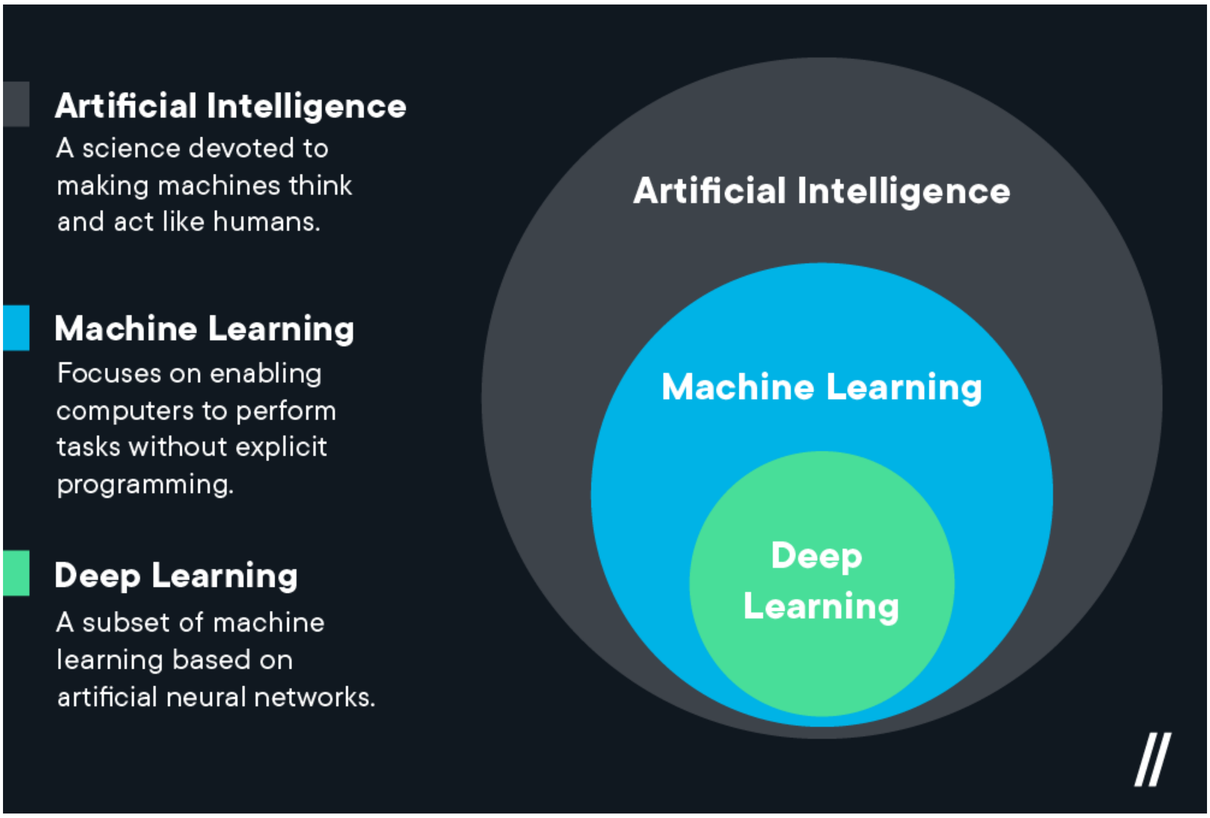

Artificial Intelligence
A science devoted to making machines think and act like humans.
Machine Learning
Focuses on enabling computers to perform tasks without explicit programming.
Deep Learning
A subset of machine learning based on artificial neural networks.
Artificial Intelligence
Machine Learning
Deep Learning

Figure 3.5 - Deep learning, machine learning and AI intersection [32]

3.6.1 CNN (Convolutional Neural Networks)

CNN have been around for a while but they are lately gaining popularity between researchers for non-traditional uses, beating more complicated and advance model on a range of uses for deep learning task or problem including object detection, Human gender detection, differentiating between cat and dog and even identifying emotion in audio file. During 1940s the first Artificial neuron model was introduced by MCCulloch-Pitts, which lead to new area of artificial neural network (ANN) and multi-layer perceptron's (MLPs). Artificial neural network was inspired and closely resembles neuron found in our brain in the term of working. During the year of 1982 two person named as Fukushima and Miyake proposed the concept of Convolutional Neural Networks (CNN), which was called as "Neocognitron" at that time, which was self-organized, hierarchical network and could identify the different pattern on the bases of shapes present in the pattern. But at the time of its discovery, it was hard to train CNN model as it required a lot of graphical power due to which CNN was used only on small to medium dataset at that time but in today's day and age its possible for us to train our CNN model with the help of powerful graphical and processing power.

3.6.2 1D CNN

1D CNN majorly focus on solving 1D data and signal dataset which is advantageous when solving 1D data problem. The computational complexity of 1D CNN is less than 2D CNN as the matrix operation is only 1D which required simple array operation. 1D CNN is able to learn accurately with the small dataset while having low number of hidden layers and neurons for the complex task involving 1D data input.

A convolution is a type of linear operation that combines two functions f and g; for the continuous one-dimensional example, this means:

$$h(x) = (f * g)(x) = \int f(u)g(x - u)$$

Figure 3.6
1D CNN convolution formula

3.6.3 2D CNN

Even though the first CNN model was introduced 20 yeas ago the 2D CNN still share some common property's the first one. 2D CNN has the ability to fuse feature extraction and future classification process into a single learning process, also have the ability to optimize feature extraction and future classification process during their training section directly from the dataset.2D CNN has the capacity to process large input data and compute on them effectively too when compare with others neural network. 2D CNN can optimize itself to the dataset size which means it can accept both large and small dataset equally and work on them. The result of 2D CNN do not get changed by the modification done on dataset like translation, scaling, distortion, and skewing.

The architecture of 2D CNN have 2D planes or layer which contains 2D weighted layers known as Kernel, and the input and output mapping which is known as feature map. 2D CNN model is also made up of convolution and pooling layers where the last polling layer is proceeded by single fully connected layer and followed by output layer to generate CNN classification output.

Because pictures are discrete and two-dimensional, we may substitute a summation for the integral and convolve in two dimensions.

$$h(x) = (f * g)(x) = \sum_{u}\sum_{v} f(u, v)g(x - u, y - v)$$

Figure 3.7

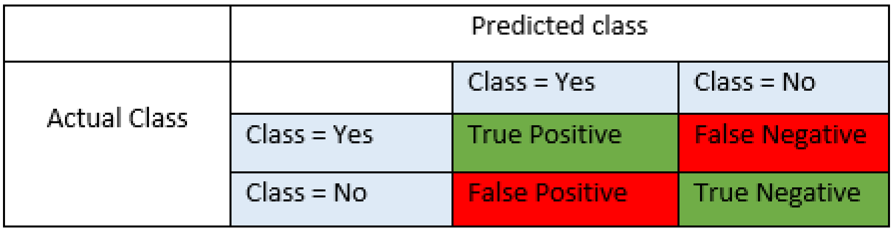
2D CNN convolution formula

3.7 Evaluation Metrics

3.7.1 Confusion Matrix

The capacity to evolve is by far the most significant feature of machine learning models. Machine learning expert evaluate the model's performance after it has been created. Evaluation metrics identify essential model parameters and offer numerical scores that may be used to assess the model's performance. The confusion matrix is the most significant parameter for evaluating the model.

A confusion matrix has four values that are used to compute various metrics and is structured against the actual and expected positive and negative classes. The true positives and true negatives indicate valid predictions in the positive and negative classes, respectively. The observations that were incorrectly forecasted for their respective classes are known as false positives and false negatives.



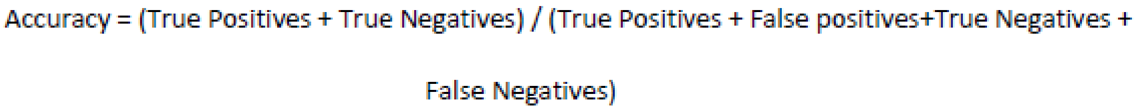
	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Figure 3.8

Confusion Matrix formula

3.7.2 Accuracy

It's the proportion of accurately predicted observations to the total number of observations. Accuracy performs better for datasets with an uniform class distribution, therefore it isn't always a desirable metric to use when assessing a model.

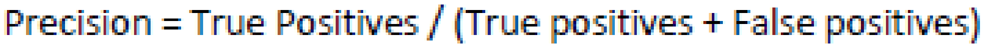


Accuracy = (True Positives + True Negatives) / (True Positives + False positives+True Negatives + False Negatives)

Figure 3.9
Accuracy formula

3.7.3 Precision

It's the proportion of successfully predicted positive observations to total expected positive observations. The higher the precision number, the better and more accurate the model is. Precision can also function with a class distribution that is unequal.

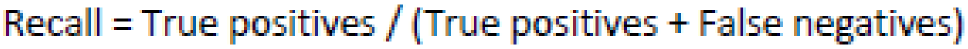


Precision = True Positives / (True positives + False positives)

Figure 3.10
Precision formula

3.7.4 Recall

It's the proportion of accurately anticipated positive observations to total positive observations. A model that has a recall score of 50% or higher is considered to be good. With an unequal class distribution, recall can still function.

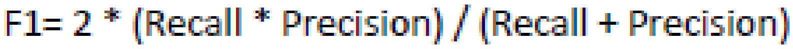


Recall = True positives / (True positives + False negatives)

Figure 3.11
Recall formula

3.7.5 F1 score

Precision and recall are weighted average values. For unequal class distribution, the F1 score is the optimum statistic.



F1= 2 * (Recall * Precision) / (Recall + Precision)

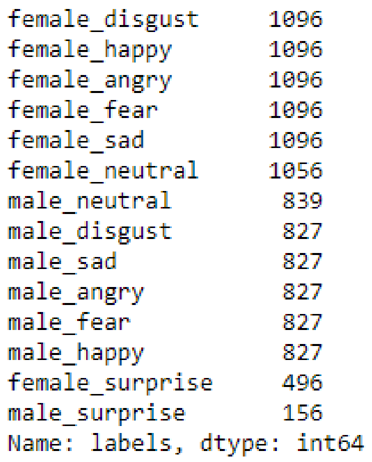
Figure 3.12
F1 score formula

Chapter 4

PRODUCT DESIGN AND IMPLEMENTATION

4.1 Datasets

For Dataset in this paper, we have Combining the four Datasets into one dataset. If we don’t combine these four datasets into one, then there are chances of overfitting the model. The other reason why we have decided to use four combined dataset is so that we have large variety and different circumstances for our dataset and when our model is used on unseen data it will be able to perform better than the model trained on only one dataset.



```
female_disgust     1096
female_happy       1096
female_angry       1096
female_fear        1096
female_sad         1096
female_neutral     1056
male_neutral        839
male_disgust        827
male_sad            827
male_angry          827
male_fear           827
male_happy          827
female_surprise     496
male_surprise       156
Name: labels, dtype: int64
```

Figure 4.1

Combine Dataset

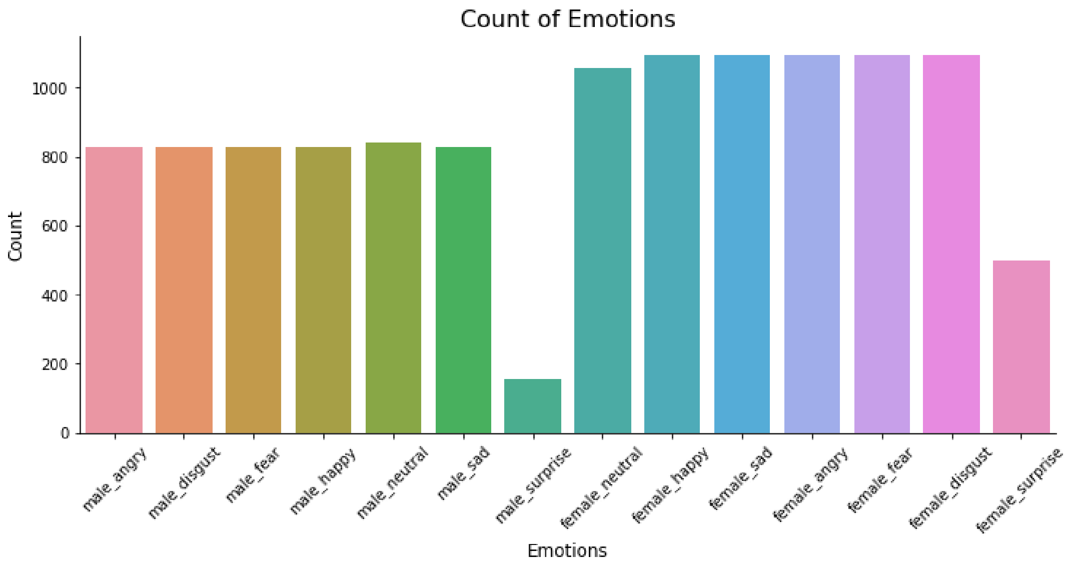
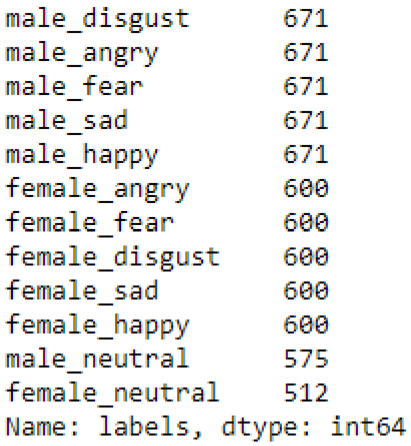


Figure 4.2 - Combine Dataset Bar Graph

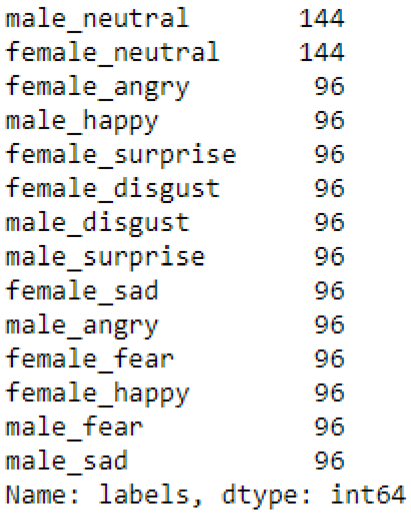
CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) contain 7442 original audio clips from 91 actors in which 48 were male and 43 were female actors aged between 20 to 74 belonging to various culture, race, places, and ethnicities. The actors for datasets spoke from a selected sentence of 12 sentences or lines. These 12 sentences were categorized under 6 different emotions i.e. Anger, Disgust, Fear, Happy, Neutral and Sad with 4 different emotion level to the emotion i.e., Low, Medium, High, and Unspecified. CREMA-D is one of the best datasets as it has large number of speakers which help with many things like the model don’t overfit and there is large no of files to train model too. The naming of the file is done in such a way that first 4 digits tell the actor ID, second 3 alphabet tells which sentence is spoke by actor from the 12 selected sentences, third 3 alphabet tells the emotion the audio file i.e., ‘ANG’ for Anger, ‘DIS’ for Disgust, ‘FEA’ for Fear, ‘HAP’ for Happy/Joy, ‘NEU’ for Neutral, ‘SAD’ for Sad and last 2 alphabet tell about the emotion level of the audio file in the dataset i.e. ‘LO’ for Low, ’MD’ for Medium , ‘HI’ for High ,’XX’ for Unspecified.



```
male_disgust      671
male_angry        671
male_fear         671
male_sad          671
male_happy        671
female_angry      600
female_fear       600
female_disgust    600
female_sad        600
female_happy      600
male_neutral      575
female_neutral    512
Name: labels, dtype: int64
```

Figure 4.3 - CREDMA–D Dataset

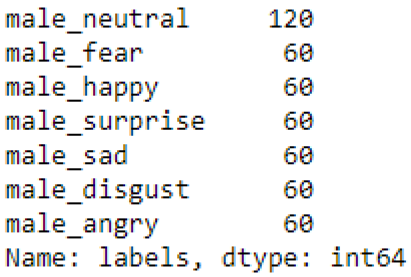
RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song): with 24 professional actors ,12 male speakers and 12 Female speakers, the lexical features (vocabulary) of the utterances are kept constant by speaking only 2 statements of equal lengths in 8 different emotions by all speakers. We got 7356 audio samples ((Total size: 24.8 GB) which were in the wav format. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. We chose this database for the verity of speech available which is produced by trained professional under the suitable condition. The naming of the file are done in such a way that it identify as first digit tells us about Modality (01 = full-AV, 02 = video-only, 03 = audio-only), second one tells us about Vocal channel (01 = speech, 02 = song), third one tells us about Emotion (01 for neutral, 02 for calm, 03 for happy, 04 for sad, 05 for angry, 06 for fearful, 07 for disgust, 08 for surprised), forth one tells us about Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion, fifth one tells us about Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door"), sixth one tall about Repetition (01 = 1st repetition, 02 = 2nd repetition) and the last one seventh tells us about Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).



```
male_neutral       144
female_neutral     144
female_angry        96
male_happy          96
female_surprise     96
female_disgust      96
male_disgust        96
male_surprise       96
female_sad          96
male_angry          96
female_fear         96
female_happy        96
male_fear           96
male_sad            96
Name: labels, dtype: int64
```

Figure 4.4-RAVDESS Dataset

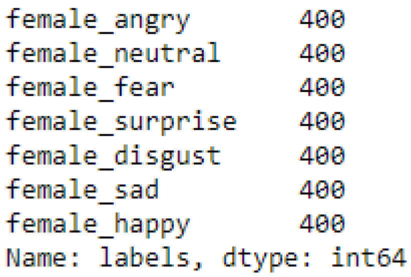
The SAVEE database has been made by recording of four native English male speakers aged between 27 and 31 years. The dataset contains files in four folders where each folder belongs to one person which are identified as DC, JE, JK, KL. Emotion was created psychologically under six category that is anger, disgust, fear, happiness, sadness, and surprise. The naming of the file is such that the prefix letters tell the emotion which are as ‘a’ for anger, 'd' for 'disgust', 'f' for 'fear’, 'h' for 'happiness’, 'n' for 'neutral’,'sa' for 'sadness' and 'su' for 'surprise'.



```
male_neutral     120
male_fear         60
male_happy        60
male_surprise     60
male_sad          60
male_disgust      60
male_angry        60
Name: labels, dtype: int64
```

Figure 4.5-SAVEE Dataset

TESS (Toronto emotional speech set) dataset contains 2800 files which are set of 200 targeted words spoken in the phrases of "Say the word ____" by two native speaking English actresses aged 26 and 64 years old and they were also university educated with musical trading. Phrase in dataset was said to stimulate seven emotion that was anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral while keeping the threshold under the normal range.



```
female_angry       400
female_neutral     400
female_fear        400
female_surprise    400
female_disgust     400
female_sad         400
female_happy       400
Name: labels, dtype: int64
```

Figure 4.6

TESS

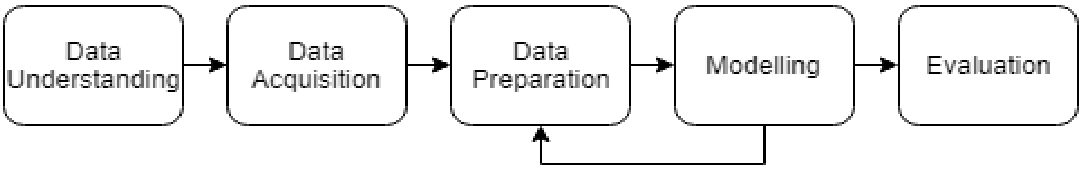


Figure 4.7-Methodology or Workflow

4.2 Data Preparation & Processing

Cleaning and converting raw data before processing and analysis is known as data preparation. It's a crucial stage before processing that frequently include reformatting data, making data changes, and integrating data sets to improve data. For data experts or users, data preparation might be time consuming, but it is necessary to put data in context in order to transform it into insights and reduce bias caused by poor data quality.

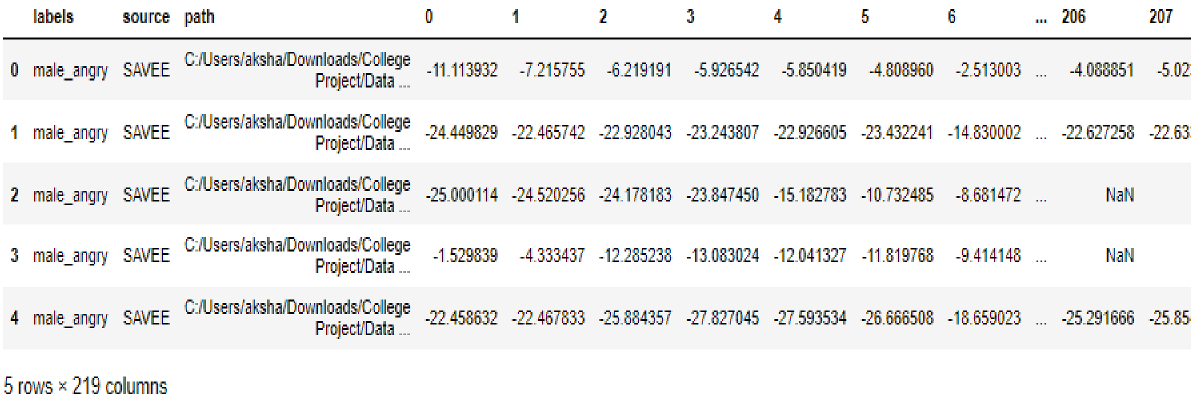
Cleaning up the data takes the longest, but it's necessary for deleting inaccurate data and filling in gaps. Important responsibilities include:

I. Extraneous data and outliers are removed.

II. Missing values are being filled in.

III. Data that follows a standardised pattern.

We will extract all of the information accessible out from audio to serve as a representation for the modelling phase because we would not be directly interacting with the audio files.



	labels	source	path	0	1	2	3	4	5	6	...	206	207
0	male_angry	SAVEE	C:/Users/aksha/Downloads/College Project/Data ...	-11.113932	-7.215755	-6.219191	-5.926542	-5.850419	-4.808960	-2.513003	...	-4.088851	-5.02
1	male_angry	SAVEE	C:/Users/aksha/Downloads/College Project/Data ...	-24.449829	-22.465742	-22.928043	-23.243807	-22.926605	-23.432241	-14.830002	...	-22.627258	-22.63
2	male_angry	SAVEE	C:/Users/aksha/Downloads/College Project/Data ...	-25.000114	-24.520256	-24.178183	-23.847450	-15.182783	-10.732485	-8.681472	...	NaN	
3	male_angry	SAVEE	C:/Users/aksha/Downloads/College Project/Data ...	-1.529839	-4.333437	-12.285238	-13.083024	-12.041327	-11.819768	-9.414148	...	NaN	
4	male_angry	SAVEE	C:/Users/aksha/Downloads/College Project/Data ...	-22.458632	-22.467833	-25.884357	-27.827045	-27.593534	-26.666508	-18.659023	...	-25.291666	-25.85

5 rows × 219 columns

Figure 4.8 - Data Preparation & Processing Output

Step 1 - Identifying emotion labels - We'll start by labelling each element with its accompanying emotion using a Python script and datasets nomenclature.

Step 2 - Segregate data on gender - The next function, which uses the same Ideology that is using datasets nomenclature, to labels each sample with its gender.

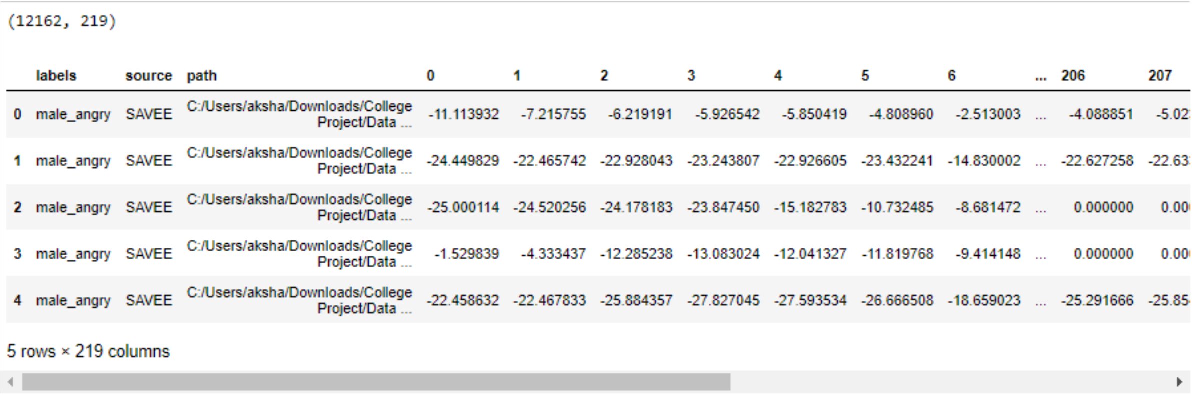
Step 3 - Feature Extraction - The features will subsequently be extracted using the Librosa 3 package. Each of the retrieved characteristics is a matrix in and of itself. This would result in a three-dimensional vector, which would be incompatible with some models.

4.3 Outlier Identification

Outlier values are also termed data modifiers since they frequently lead the prediction system to be misdirected. Outliers also affect general data statistics like mean, variance, and standard deviation. The percentage of outliers in the entire dataset can be utilised to make judgments about how to deal with them. There will be no need for repairs if the outliers are within a short range of difference and contribute to a very tiny fraction. Replacing outlier data with boundary, mean, median, or mode values is one way for dealing with high numbers of outliers.

4.4 Null Value Handling

The null value error is a typical issue that can occur in a dataset. It occurs when the terms 'null' or 'NA' are used as fillers in place of missing data. Null values are frequently treated and handled in the same way as missing values are. We have replaced null value with zero first and in normalization we have changed the value so that it cannot affect the model accuracy.



```
(12162, 219)
```

	labels	source	path	0	1	2	3	4	5	6	...	206	207
0	male_angry	SAVEE	C:/Users/aksha/Downloads/College Project/Data ...	-11.113932	-7.215755	-6.219191	-5.926542	-5.850419	-4.808960	-2.513003	...	-4.088851	-5.02
1	male_angry	SAVEE	C:/Users/aksha/Downloads/College Project/Data ...	-24.449829	-22.465742	-22.928043	-23.243807	-22.926605	-23.432241	-14.830002	...	-22.627258	-22.63
2	male_angry	SAVEE	C:/Users/aksha/Downloads/College Project/Data ...	-25.000114	-24.520256	-24.178183	-23.847450	-15.182783	-10.732485	-8.681472	...	0.000000	0.00
3	male_angry	SAVEE	C:/Users/aksha/Downloads/College Project/Data ...	-1.529839	-4.333437	-12.285238	-13.083024	-12.041327	-11.819768	-9.414148	...	0.000000	0.00
4	male_angry	SAVEE	C:/Users/aksha/Downloads/College Project/Data ...	-22.458632	-22.467833	-25.884357	-27.827045	-27.593534	-26.666508	-18.659023	...	-25.291666	-25.85

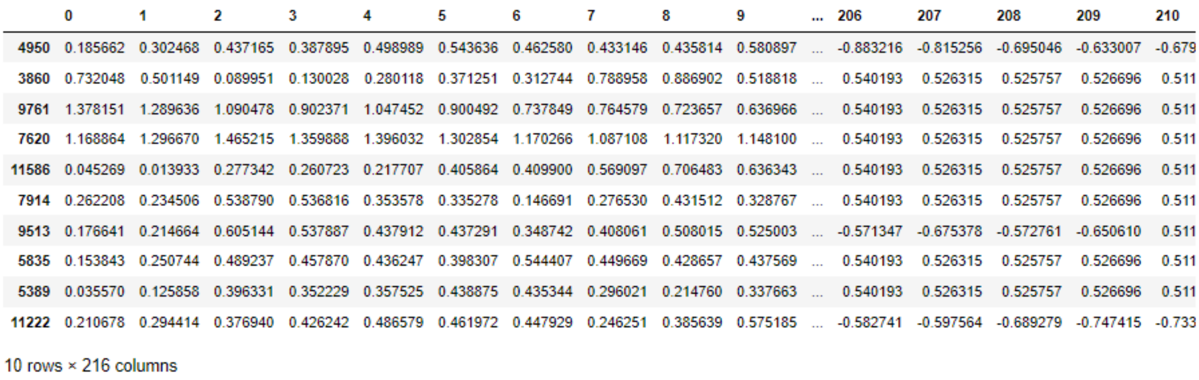
5 rows × 219 columns

Figure 4.9-Null Value Handling Output

4.5 Normalization

Normalization is a data preparation method used often in machine learning. Normalization is the process of converting the values of numeric columns in a dataset to a similar scale without distorting the ranges of values or losing information. Some algorithms require normalisation in order to appropriately model the data.

Distinct units or scales are used to compute different aspects of the audio signal, which are represented by its features. Calculations will be more accurate if the numbers are rescaled to a homogeneous range. For the calculation of several algorithms, distance metrics are used. As a result, all of the values in the dataset must be standardised.



	0	1	2	3	4	5	6	7	8	9	...	206	207	208	209	210
4950	0.185662	0.302468	0.437165	0.387895	0.498989	0.543636	0.462580	0.433146	0.435814	0.580897	...	-0.883216	-0.815256	-0.695046	-0.633007	-0.679
3860	0.732048	0.501149	0.089951	0.130028	0.280118	0.371251	0.312744	0.788958	0.886902	0.518818	...	0.540193	0.526315	0.525757	0.526696	0.511
9761	1.378151	1.289636	1.090478	0.902371	1.047452	0.900492	0.737849	0.764579	0.723657	0.636966	...	0.540193	0.526315	0.525757	0.526696	0.511
7620	1.168864	1.296670	1.465215	1.359888	1.396032	1.302854	1.170266	1.087108	1.117320	1.148100	...	0.540193	0.526315	0.525757	0.526696	0.511
11586	0.045269	0.013933	0.277342	0.260723	0.217707	0.405864	0.409900	0.569097	0.706483	0.636343	...	0.540193	0.526315	0.525757	0.526696	0.511
7914	0.262208	0.234506	0.538790	0.536816	0.353578	0.335278	0.146691	0.276530	0.431512	0.328767	...	0.540193	0.526315	0.525757	0.526696	0.511
9513	0.176641	0.214664	0.605144	0.537887	0.437912	0.437291	0.348742	0.408061	0.508015	0.525003	...	-0.571347	-0.675378	-0.572761	-0.650610	0.511
5835	0.153843	0.250744	0.489237	0.457870	0.436247	0.398307	0.544407	0.449669	0.428657	0.437569	...	0.540193	0.526315	0.525757	0.526696	0.511
5389	0.035570	0.125858	0.396331	0.352229	0.357525	0.438875	0.435344	0.296021	0.214760	0.337663	...	0.540193	0.526315	0.525757	0.526696	0.511
11222	0.210678	0.294414	0.376940	0.426242	0.486579	0.461972	0.447929	0.246251	0.385639	0.575185	...	-0.582741	-0.597564	-0.689279	-0.747415	-0.733

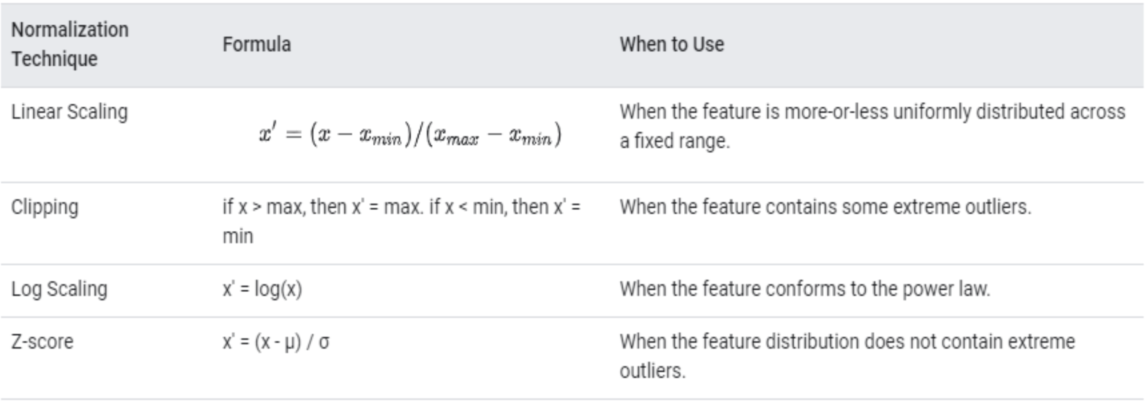
10 rows × 216 columns

Figure 4.10 Normalization Output

Linear scaling normalization or Mean normalization - In machine learning, Linear scaling is the process of transforming data into the range [0, 1] (or any other range) or simply onto the unit sphere.

Z-sore Normalization - The number of standard deviations from from the mean is represented by the Z-score, which is a form of scaling. To confirm that your feature distributions have mean = 0 and std = 1, we can apply z-score. It's handy when there are a few outliers but not so many that trimming is required.

We have used Linear scaling normalization or Mean normalization and Z-sore Normalization on our dataset used in CNN model.



Normalization Technique	Formula	When to Use
Linear Scaling	$x' = (x - x_{min})/(x_{max} - x_{min})$	When the feature is more-or-less uniformly distributed across a fixed range.
Clipping	if x > max, then x' = max. if x < min, then x' = min	When the feature contains some extreme outliers.
Log Scaling	x' = log(x)	When the feature conforms to the power law.
Z-score	x' = (x - μ) / σ	When the feature distribution does not contain extreme outliers.

Figure 4.11

Normalization Methods [31]

4.6 Loss Curve Graph

A learning curve is a graph measuring model learning performance against time or experience. Learning curves are a common machine learning diagnostic tool for algorithms that learn progressively from a training dataset. After each update during training, the model may be tested on the training dataset and a holdout validation dataset, and graphs of the measured performance can be constructed to display learning curves. Examining model learning curves during training may help detect learning issues, such as an underfit or overfit model, as well as whether the training and validation datasets are sufficiently representative.

Chapter 5

RESULTS

The implementations of speech emotion recognition with different methods yielded a number of observations and conclusions from the results and between the numerous methods or approaches, the performance and accuracy have improved overall.

5.1 Observation

Observation 1 - We have observed that our 1D CNN model with only MFCC had the accuracy of 49% but when we apply data augmentation technique to it then we are able to increase the accuracy by 5%.

Observation 2 – We have also observed that in 2D CNN model with MFCC mean but without data augmentation had 65% accuracy and 2D CNN with MFCC mean and data augmentation had the accuracy 60% was able to get higher accuracy when compare with the 2D CNN with data augmentation with difference of 4%.

Observation 3 – We have also seen that in 2D CNN with augmentation and without augmentation perform at same accuracy.

Table 5.1

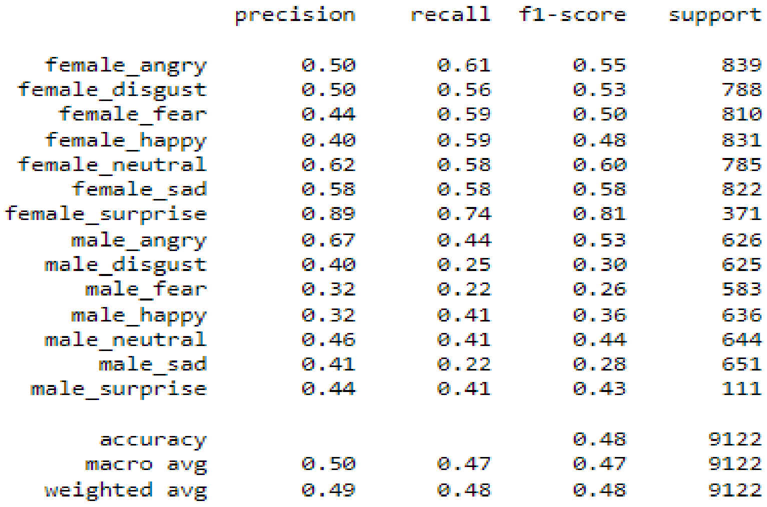
Result Summary

Model	**Feature Used**	**Accuracy**
1D CNN	MFCC	43%
1D CNN	MFCC mean with data augmentation	48%
2D CNN	MFCC mean without augmentation	65%
2D CNN	MFCC with Augmentation	61%
2D CNN	Log-melspectogram without augmentation	63%
2D CNN	Log-melspectogram with augmentation	63%

5.2 1D CNN With MFCC Mean

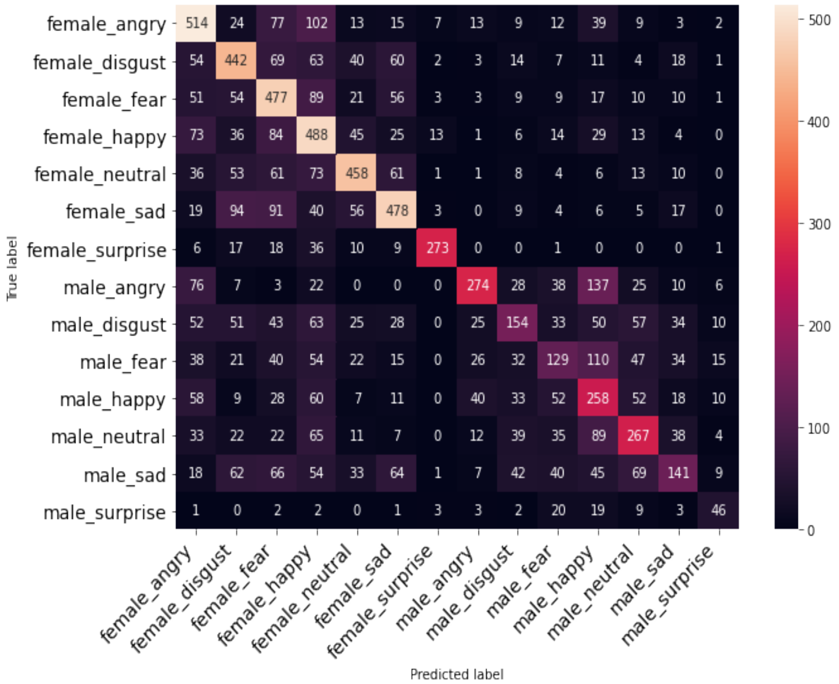
Emotion by gender accuracy of 1D CNN model with MFCC

This is the Classification report for Emotion by gender accuracy of 1D CNN model with MFCC. Where we got the weight accuracy of 49%. In this we are Using our model to identify emotion and gender both together.



```
                  precision    recall  f1-score   support

    female_angry       0.50      0.61      0.55       839
  female_disgust       0.50      0.56      0.53       788
     female_fear       0.44      0.59      0.50       810
    female_happy       0.40      0.59      0.48       831
  female_neutral       0.62      0.58      0.60       785
      female_sad       0.58      0.58      0.58       822
 female_surprise       0.89      0.74      0.81       371
      male_angry       0.67      0.44      0.53       626
    male_disgust       0.40      0.25      0.30       625
       male_fear       0.32      0.22      0.26       583
      male_happy       0.32      0.41      0.36       636
    male_neutral       0.46      0.41      0.44       644
        male_sad       0.41      0.22      0.28       651
   male_surprise       0.44      0.41      0.43       111

        accuracy                           0.48      9122
       macro avg       0.50      0.47      0.47      9122
    weighted avg       0.49      0.48      0.48      9122
```

Figure 5.1 Figure – Classification report of Emotion by gender accuracy of 1D CNN model with MFCC



	female_angry	female_disgust	female_fear	female_happy	female_neutral	female_sad	female_surprise	male_angry	male_disgust	male_fear	male_happy	male_neutral	male_sad	male_surprise
female_angry	514	24	77	102	13	15	7	13	9	12	39	9	3	2
female_disgust	54	442	69	63	40	60	2	3	14	7	11	4	18	1
female_fear	51	54	477	89	21	56	3	3	9	9	17	10	10	1
female_happy	73	36	84	488	45	25	13	1	6	14	29	13	4	0
female_neutral	36	53	61	73	458	61	1	1	8	4	6	13	10	0
female_sad	19	94	91	40	56	478	3	0	9	4	6	5	17	0
female_surprise	6	17	18	36	10	9	273	0	0	1	0	0	0	1
male_angry	76	7	3	22	0	0	0	274	28	38	137	25	10	6
male_disgust	52	51	43	63	25	28	0	25	154	33	50	57	34	10
male_fear	38	21	40	54	22	15	0	26	32	129	110	47	34	15
male_happy	58	9	28	60	7	11	0	40	33	52	258	52	18	10
male_neutral	33	22	22	65	11	7	0	12	39	35	89	267	38	4
male_sad	18	62	66	54	33	64	1	7	42	40	45	69	141	9
male_surprise	1	0	2	2	0	1	3	3	2	20	19	9	3	46

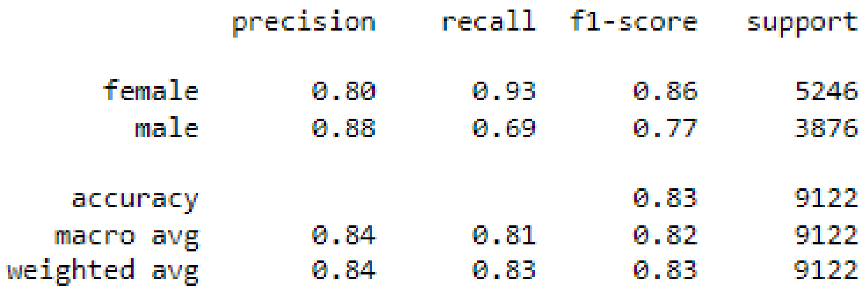
Figure 5.2

Confusion matrix Emotion by gender accuracy of 1D CNN model with MFCC

Emotion by gender accuracy of 1D CNN model with MFCCFor same we also have confusion matrix of Emotion by gender accuracy of 1D CNN model with MFCC.

Gender Accuracy Result of 1D CNN model with MFCC

This is the Classification report for Emotion by Gender Accuracy Result of 1D CNN model with MFCC. Where we got the weight accuracy of 84%. In this we are Using our model to identify gender accuracy only without any emotion label attach to it.



```
              precision    recall  f1-score   support

      female       0.80      0.93      0.86      5246
        male       0.88      0.69      0.77      3876

    accuracy                           0.83      9122
   macro avg       0.84      0.81      0.82      9122
weighted avg       0.84      0.83      0.83      9122
```

Figure 5.3

Classification report Gender Accuracy Result of 1D CNN model with MFCC

For the same we also have confusion matrix of Gender Accuracy Result of 1D CNN model with MFCC.

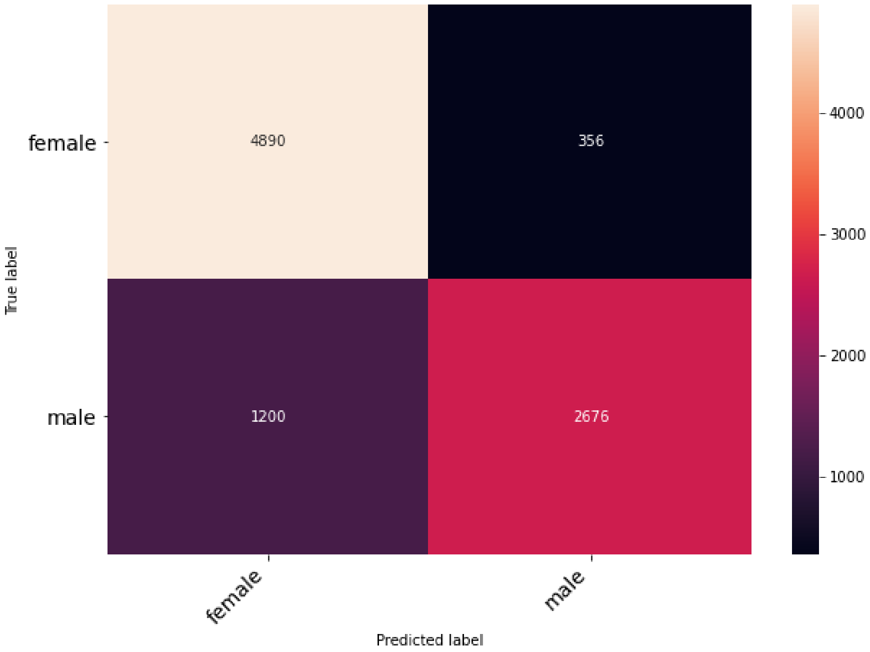
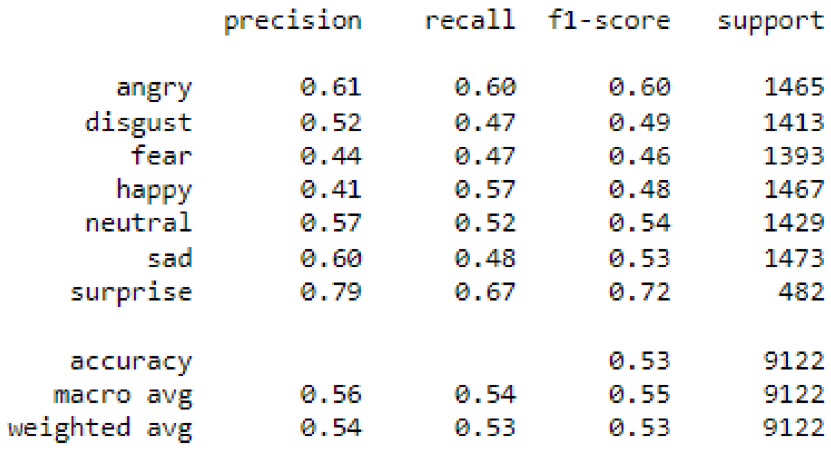


Figure 5.4

Confusion matrix Gender accuracy of 1D CNN model with MFCC

Emotion Accuracy of 1D CNN model with MFCC

This is the Classification report for Emotion Accuracy of 1D CNN model with MFCC. Where we got the weight accuracy of 54%. In this we are Using our model to identify emotion accuracy only without any gender label added to it.

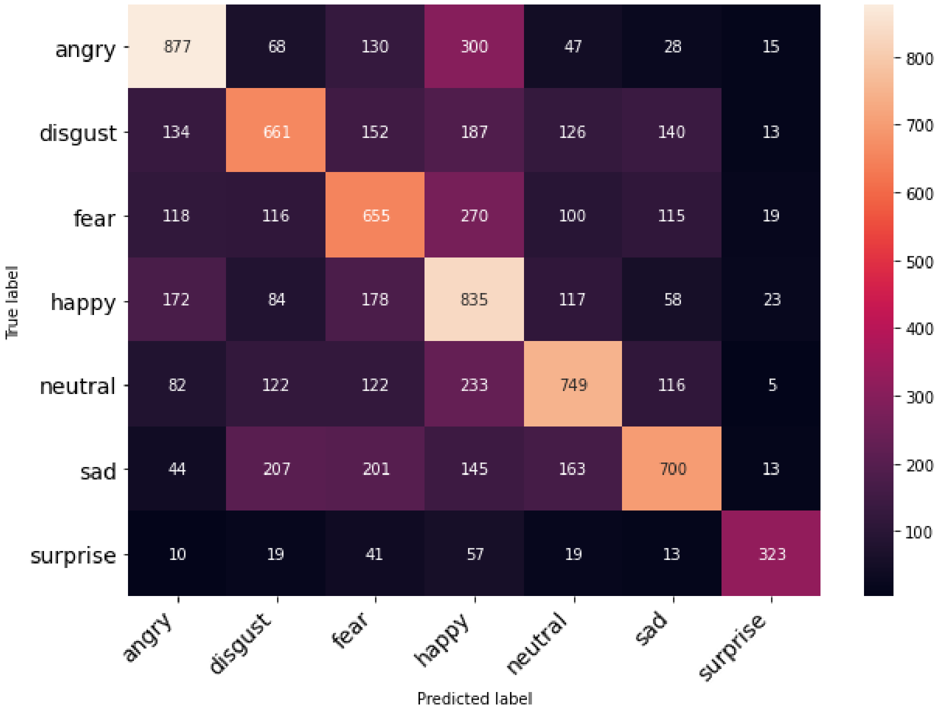


```
              precision    recall  f1-score   support

       angry       0.61      0.60      0.60      1465
     disgust       0.52      0.47      0.49      1413
        fear       0.44      0.47      0.46      1393
       happy       0.41      0.57      0.48      1467
     neutral       0.57      0.52      0.54      1429
         sad       0.60      0.48      0.53      1473
    surprise       0.79      0.67      0.72       482

    accuracy                           0.53      9122
   macro avg       0.56      0.54      0.55      9122
weighted avg       0.54      0.53      0.53      9122
```

Figure 5.5

Classification report Emotion Accuracy Result of 1D CNN model with MFCC

For same we also have confusion matrix of Emotion Accuracy of 1D CNN model with MFCC.



True label	angry	disgust	fear	happy	neutral	sad	surprise
angry	877	68	130	300	47	28	15
disgust	134	661	152	187	126	140	13
fear	118	116	655	270	100	115	19
happy	172	84	178	835	117	58	23
neutral	82	122	122	233	749	116	5
sad	44	207	201	145	163	700	13
surprise	10	19	41	57	19	13	323

Figure 5.6

Confusion matrix Emotion accuracy of 1D CNN model with MFCC

This is the Loss and train graph for our 1D CNN With MFCC Mean over 100 Epoch.

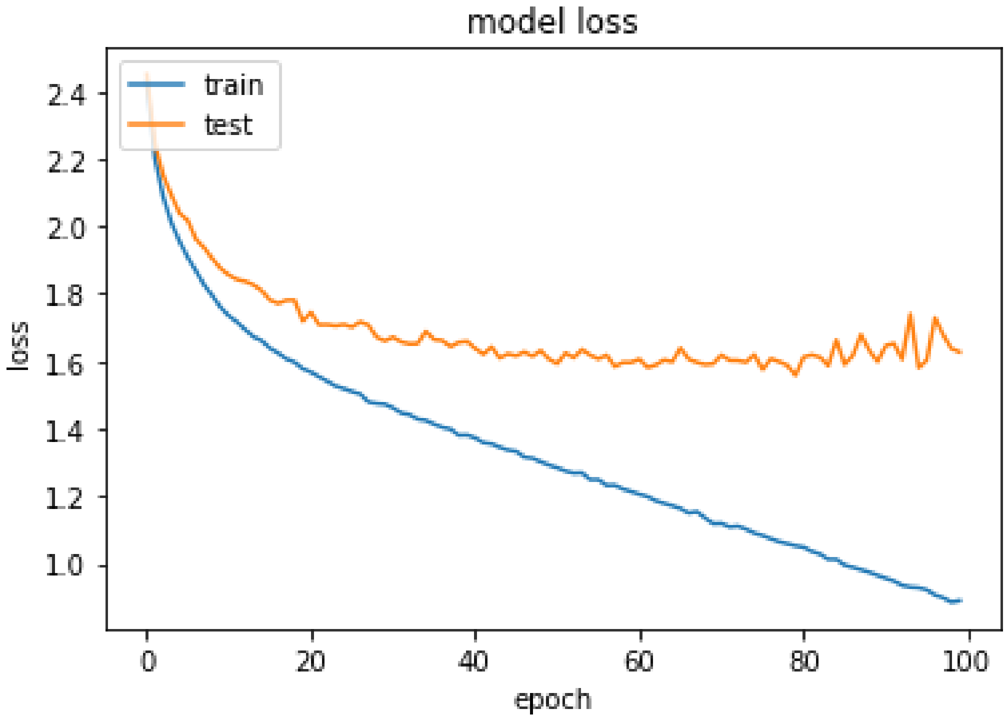


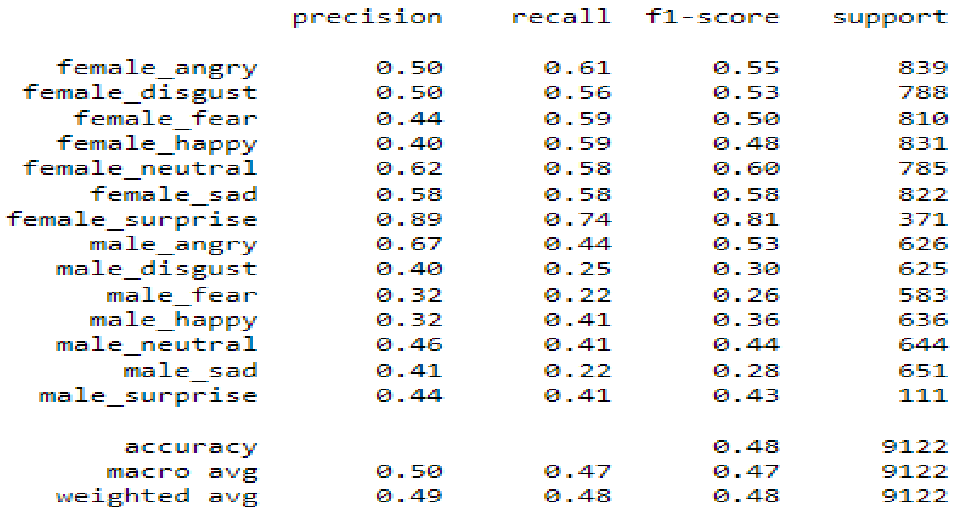
Figure 5.7

Loss curves for 1D CNN model with MFCC feature on combine dataset

5.3 1D CNN model with MFCC mean and Data augmentation

Emotion by gender accuracy for1D CNN model with MFCC mean and Data augmentation

This is the Classification report for Emotion by gender accuracy for1D CNN model with MFCC mean and Data augmentation. Where we got the weight accuracy of 49%. In this we are Using our model to identify emotion and gender both together.

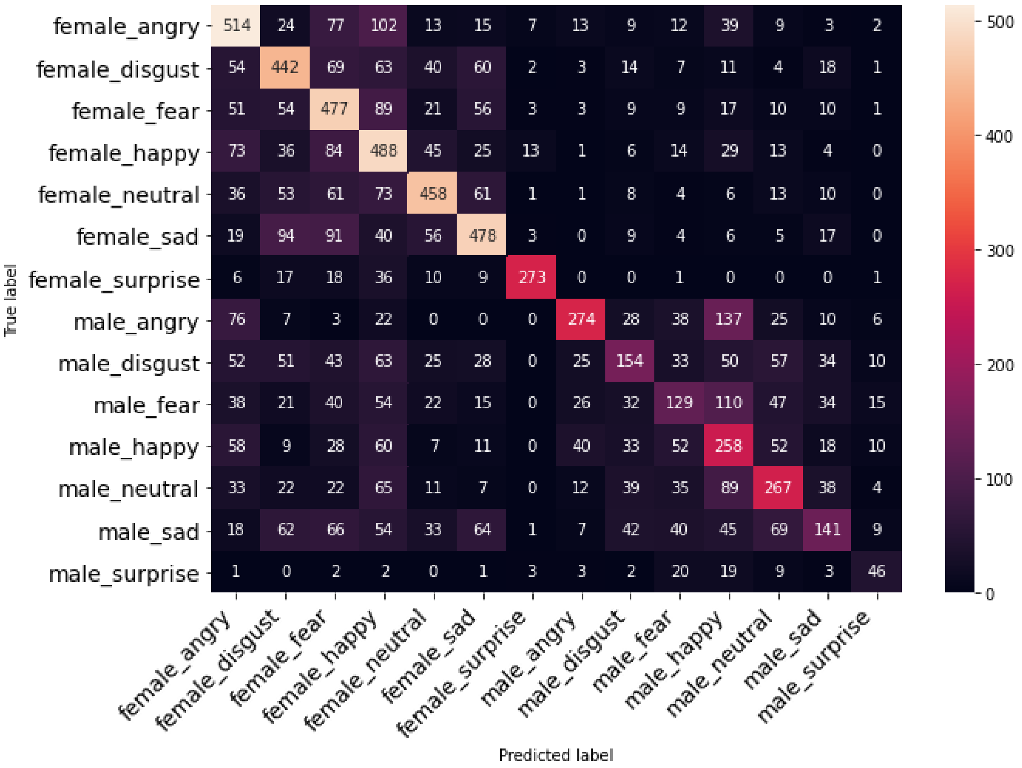


```
                 precision    recall  f1-score   support

   female_angry       0.50      0.61      0.55       839
 female_disgust       0.50      0.56      0.53       788
    female_fear       0.44      0.59      0.50       810
   female_happy       0.40      0.59      0.48       831
 female_neutral       0.62      0.58      0.60       785
     female_sad       0.58      0.58      0.58       822
female_surprise       0.89      0.74      0.81       371
     male_angry       0.67      0.44      0.53       626
   male_disgust       0.40      0.25      0.30       625
      male_fear       0.32      0.22      0.26       583
     male_happy       0.32      0.41      0.36       636
   male_neutral       0.46      0.41      0.44       644
       male_sad       0.41      0.22      0.28       651
  male_surprise       0.44      0.41      0.43       111

       accuracy                           0.48      9122
      macro avg       0.50      0.47      0.47      9122
   weighted avg       0.49      0.48      0.48      9122
```

Figure 5.8

Classification report of Emotion by gender accuracy of 1D CNN model with MFCC mean and Data augmentation

For same we also have confusion matrix of Emotion by gender accuracy for1D CNN model with MFCC mean and Data augmentation.



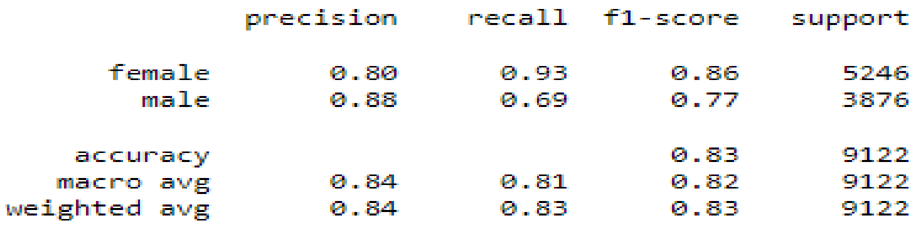
	female_angry	female_disgust	female_fear	female_happy	female_neutral	female_sad	female_surprise	male_angry	male_disgust	male_fear	male_happy	male_neutral	male_sad	male_surprise
female_angry	514	24	77	102	13	15	7	13	9	12	39	9	3	2
female_disgust	54	442	69	63	40	60	2	3	14	7	11	4	18	1
female_fear	51	54	477	89	21	56	3	3	9	9	17	10	10	1
female_happy	73	36	84	488	45	25	13	1	6	14	29	13	4	0
female_neutral	36	53	61	73	458	61	1	1	8	4	6	13	10	0
female_sad	19	94	91	40	56	478	3	0	9	4	6	5	17	0
female_surprise	6	17	18	36	10	9	273	0	0	1	0	0	0	1
male_angry	76	7	3	22	0	0	0	274	28	38	137	25	10	6
male_disgust	52	51	43	63	25	28	0	25	154	33	50	57	34	10
male_fear	38	21	40	54	22	15	0	26	32	129	110	47	34	15
male_happy	58	9	28	60	7	11	0	40	33	52	258	52	18	10
male_neutral	33	22	22	65	11	7	0	12	39	35	89	267	38	4
male_sad	18	62	66	54	33	64	1	7	42	40	45	69	141	9
male_surprise	1	0	2	2	0	1	3	3	2	20	19	9	3	46

Figure 5.9

Confusion matrix Emotion by gender accuracy of 1D CNN model with MFCC and Data augmentation

Gender accuracy result for 1D CNN model with MFCC mean and Data augmentation

This is the Classification report Gender accuracy result for 1D CNN model with MFCC mean and Data augmentation. Where we got the weight accuracy of 84%. In this we are Using our model to identify gender only without any emotion label attach to it.



```
              precision    recall  f1-score   support

      female       0.80      0.93      0.86      5246
        male       0.88      0.69      0.77      3876

    accuracy                           0.83      9122
   macro avg       0.84      0.81      0.82      9122
weighted avg       0.84      0.83      0.83      9122
```

Figure 5.10

Classification report Gender Accuracy Result of 1D CNN model with MFCC and Data augmentation

For same we also have confusion matrix of Gender accuracy result for 1D CNN model with MFCC mean and Data augmentation.

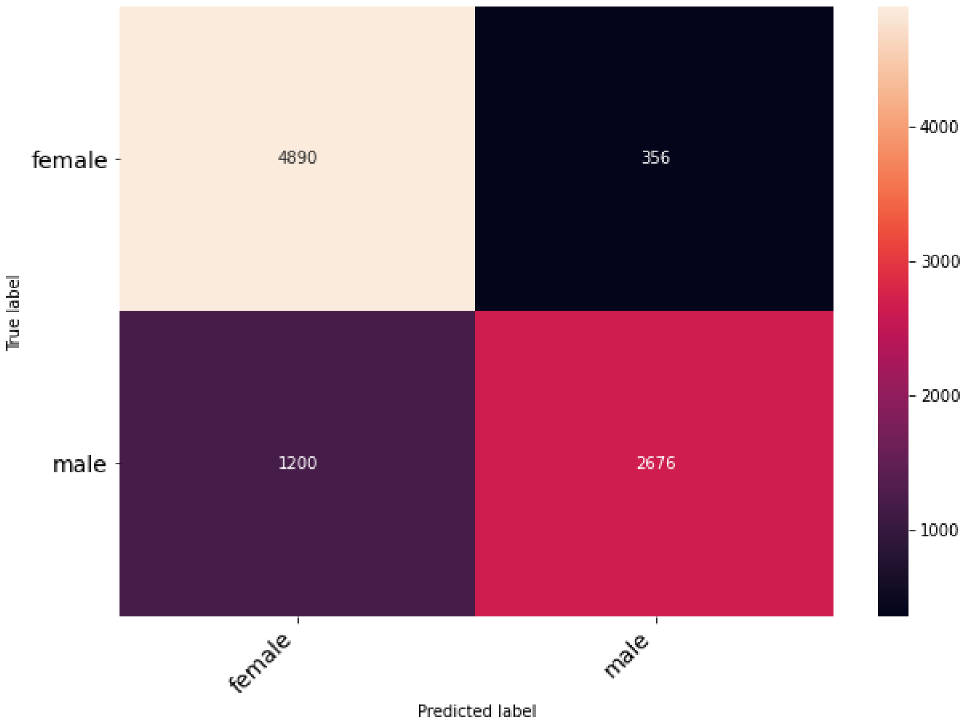
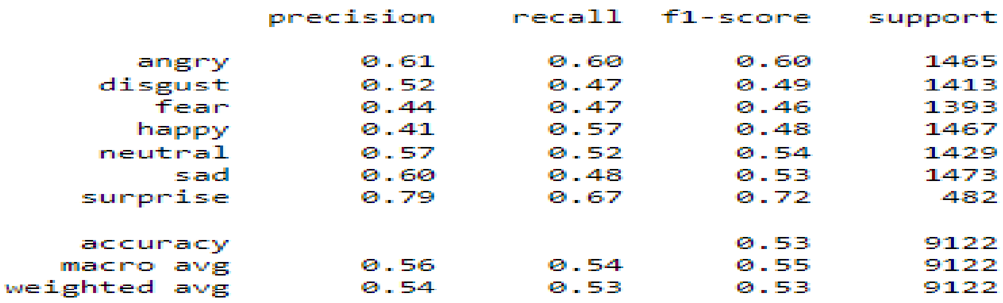


Figure 5.11

Confusion matrix Gender accuracy of 1D CNN model with MFCC Data augmentation

Emotion accuracy for 1D CNN model with MFCC mean and Data augmentation

This is the Classification report Gender accuracy result for Emotion accuracy for 1D CNN model with MFCC mean and Data augmentation. Where we got the weight accuracy of 54%. In this we are Using our model to identify Emotion only without any gender label attach to it.

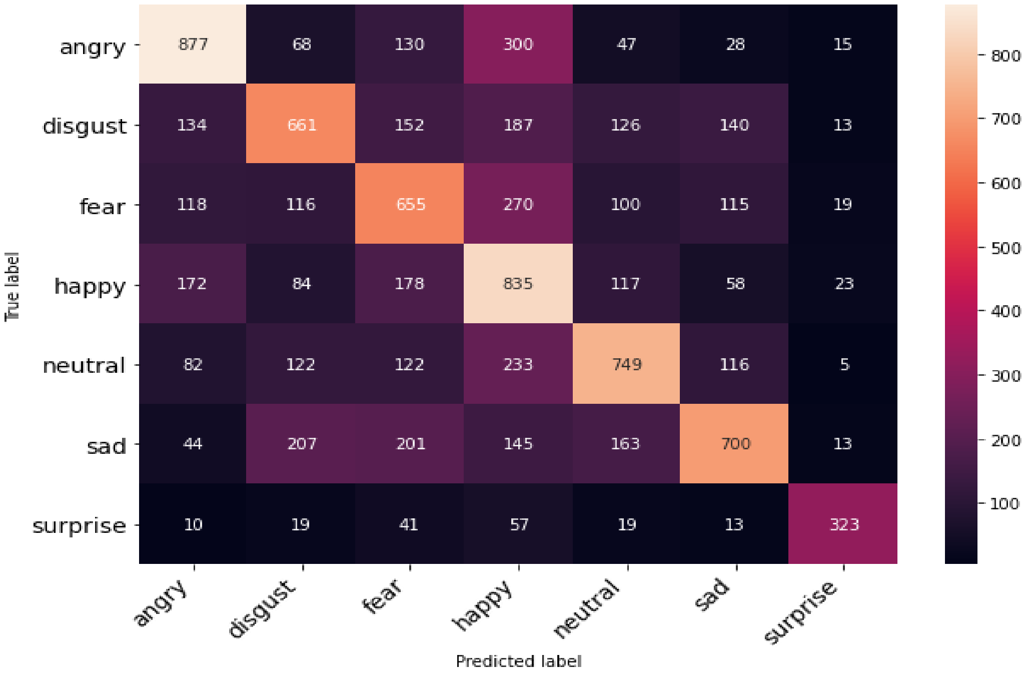


```
              precision    recall  f1-score   support

       angry       0.61      0.60      0.60      1465
     disgust       0.52      0.47      0.49      1413
        fear       0.44      0.47      0.46      1393
       happy       0.41      0.57      0.48      1467
     neutral       0.57      0.52      0.54      1429
         sad       0.60      0.48      0.53      1473
    surprise       0.79      0.67      0.72       482

    accuracy                           0.53      9122
   macro avg       0.56      0.54      0.55      9122
weighted avg       0.54      0.53      0.53      9122
```

Figure 5.12

Classification report Emotion accuracy of 1D CNN model with MFCC mean and Data augmentation

For same we also have confusion matrix of Emotion accuracy for 1D CNN model with MFCC mean and Data augmentation.



True label	angry	disgust	fear	happy	neutral	sad	surprise
angry	877	68	130	300	47	28	15
disgust	134	661	152	187	126	140	13
fear	118	116	655	270	100	115	19
happy	172	84	178	835	117	58	23
neutral	82	122	122	233	749	116	5
sad	44	207	201	145	163	700	13
surprise	10	19	41	57	19	13	323

Figure 5.13

Confusion matrix Emotion accuracy of 1D CNN model with MFCC and Data augmentation

This is the Loss and train graph for our 1D CNN model with MFCC mean and Data augmentation over 150 Epoch.

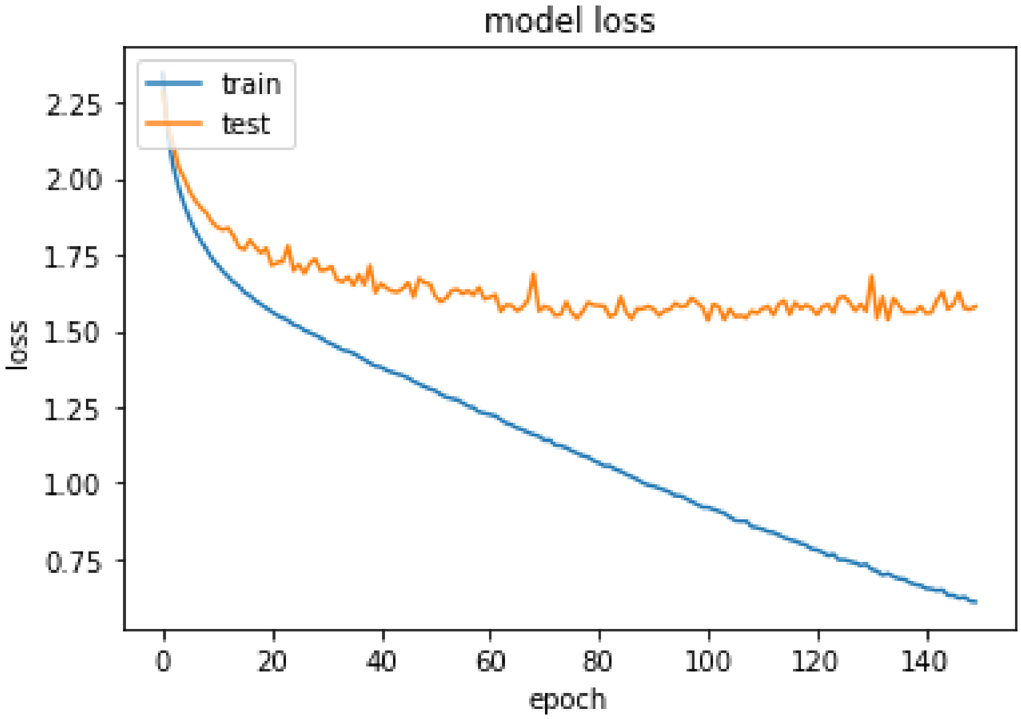


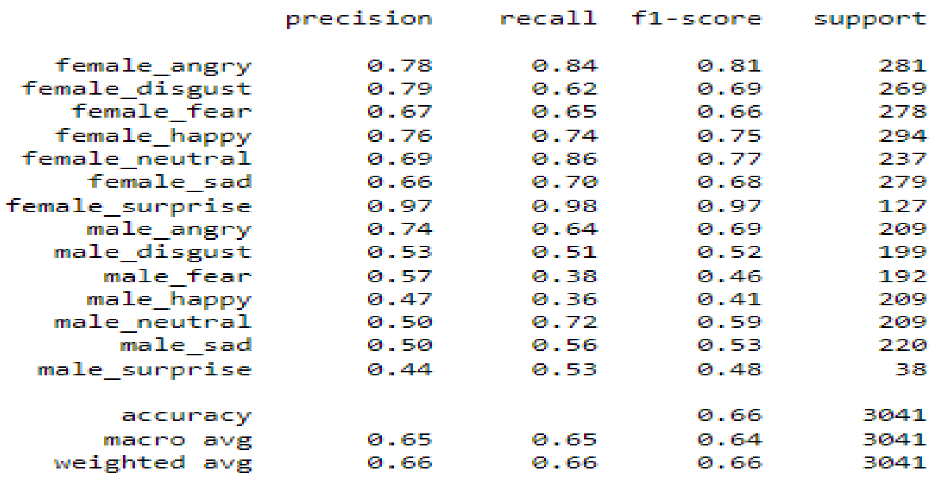
Figure 5.14

Loss curves for 1D CNN model with MFCC and data augmentation feature on combine dataset

5.4 2D CNN with MFCC without data augmentation

Emotion by gender accuracy for 2D CNN with MFCC without data augmentation

This is the Classification report for Emotion by gender accuracy for 2D CNN with MFCC without data augmentation. Where we got the weight accuracy of 65%. In this we are Using our model to identify emotion and gender both together.

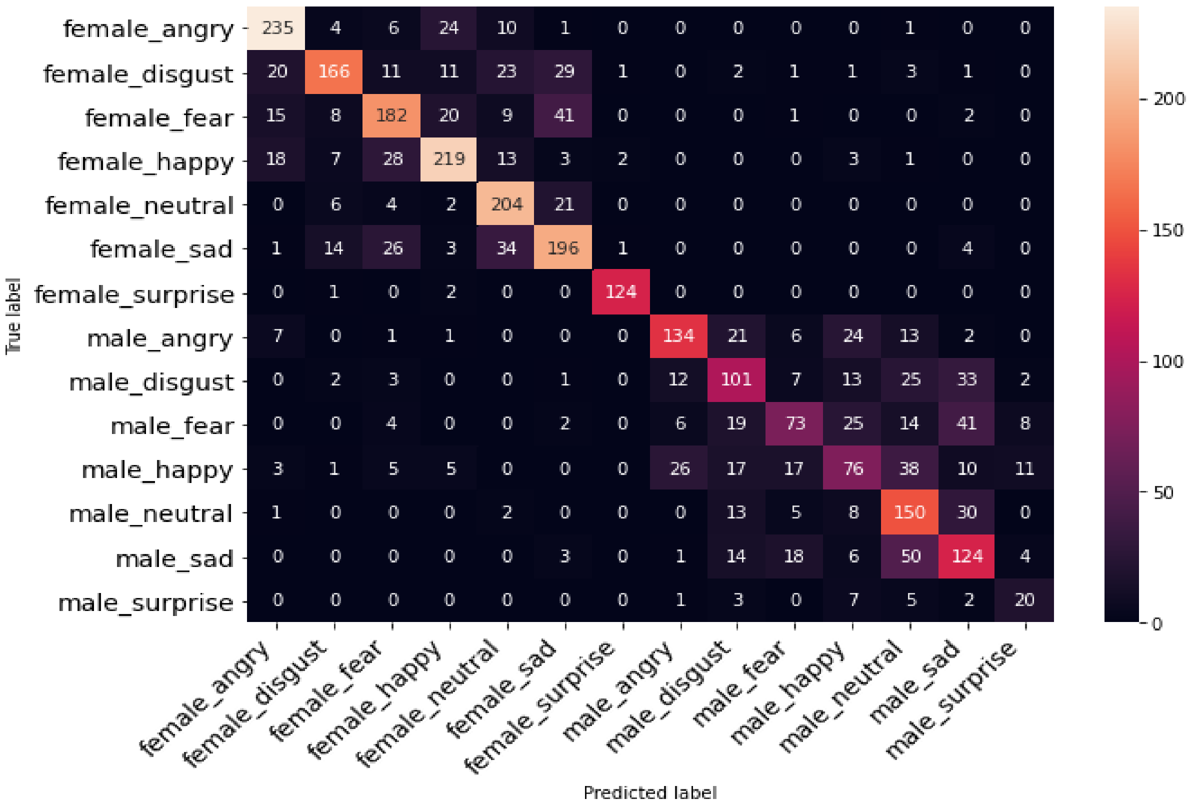


```
                  precision    recall  f1-score   support

   female_angry       0.78      0.84      0.81       281
 female_disgust       0.79      0.62      0.69       269
    female_fear       0.67      0.65      0.66       278
   female_happy       0.76      0.74      0.75       294
 female_neutral       0.69      0.86      0.77       237
     female_sad       0.66      0.70      0.68       279
female_surprise       0.97      0.98      0.97       127
     male_angry       0.74      0.64      0.69       209
   male_disgust       0.53      0.51      0.52       199
      male_fear       0.57      0.38      0.46       192
     male_happy       0.47      0.36      0.41       209
   male_neutral       0.50      0.72      0.59       209
       male_sad       0.50      0.56      0.53       220
  male_surprise       0.44      0.53      0.48        38

       accuracy                           0.66      3041
      macro avg       0.65      0.65      0.64      3041
   weighted avg       0.66      0.66      0.66      3041
```

Figure 5.15

Classification report of Emotion by gender accuracy of 2D CNN model with MFCC

For same we also have confusion matrix of Emotion accuracy for for 2D CNN with MFCC without data augmentation.



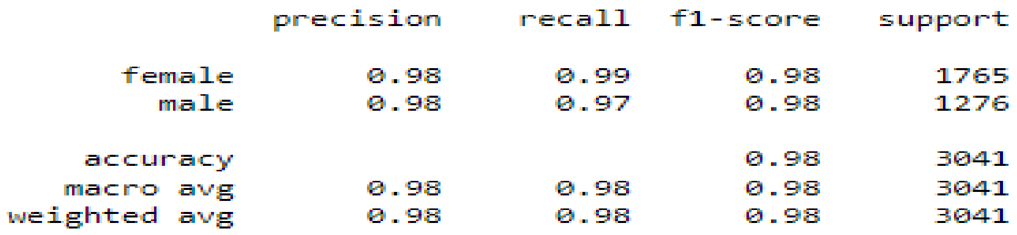
	female_angry	female_disgust	female_fear	female_happy	female_neutral	female_sad	female_surprise	male_angry	male_disgust	male_fear	male_happy	male_neutral	male_sad	male_surprise
female_angry	235	4	6	24	10	1	0	0	0	0	0	1	0	0
female_disgust	20	166	11	11	23	29	1	0	2	1	1	3	1	0
female_fear	15	8	182	20	9	41	0	0	0	1	0	0	2	0
female_happy	18	7	28	219	13	3	2	0	0	0	3	1	0	0
female_neutral	0	6	4	2	204	21	0	0	0	0	0	0	0	0
female_sad	1	14	26	3	34	196	1	0	0	0	0	0	4	0
female_surprise	0	1	0	2	0	0	124	0	0	0	0	0	0	0
male_angry	7	0	1	1	0	0	0	134	21	6	24	13	2	0
male_disgust	0	2	3	0	0	1	0	12	101	7	13	25	33	2
male_fear	0	0	4	0	0	2	0	6	19	73	25	14	41	8
male_happy	3	1	5	5	0	0	0	26	17	17	76	38	10	11
male_neutral	1	0	0	0	2	0	0	0	13	5	8	150	30	0
male_sad	0	0	0	0	0	3	0	1	14	18	6	50	124	4
male_surprise	0	0	0	0	0	0	0	1	3	0	7	5	2	20

Figure 5.16

Confusion matrix Emotion by gender accuracy of 2D CNN model with MFCC

Gender Accuracy Result for 2D CNN with MFCC without data augmentation

This is the Classification report for Gender Accuracy Result for 2D CNN with MFCC without data augmentation. Where we got the weight accuracy of 98%. In this we are Using our model to identify gender only without any emotion label added to it.



```
              precision    recall  f1-score   support

      female       0.98      0.99      0.98      1765
        male       0.98      0.97      0.98      1276

    accuracy                           0.98      3041
   macro avg       0.98      0.98      0.98      3041
weighted avg       0.98      0.98      0.98      3041
```

Figure 5.17

Classification report Gender Accuracy Result of 2D CNN model with MFCC

For same we also have confusion matrix Gender Accuracy Result for 2D CNN with MFCC without data augmentation.

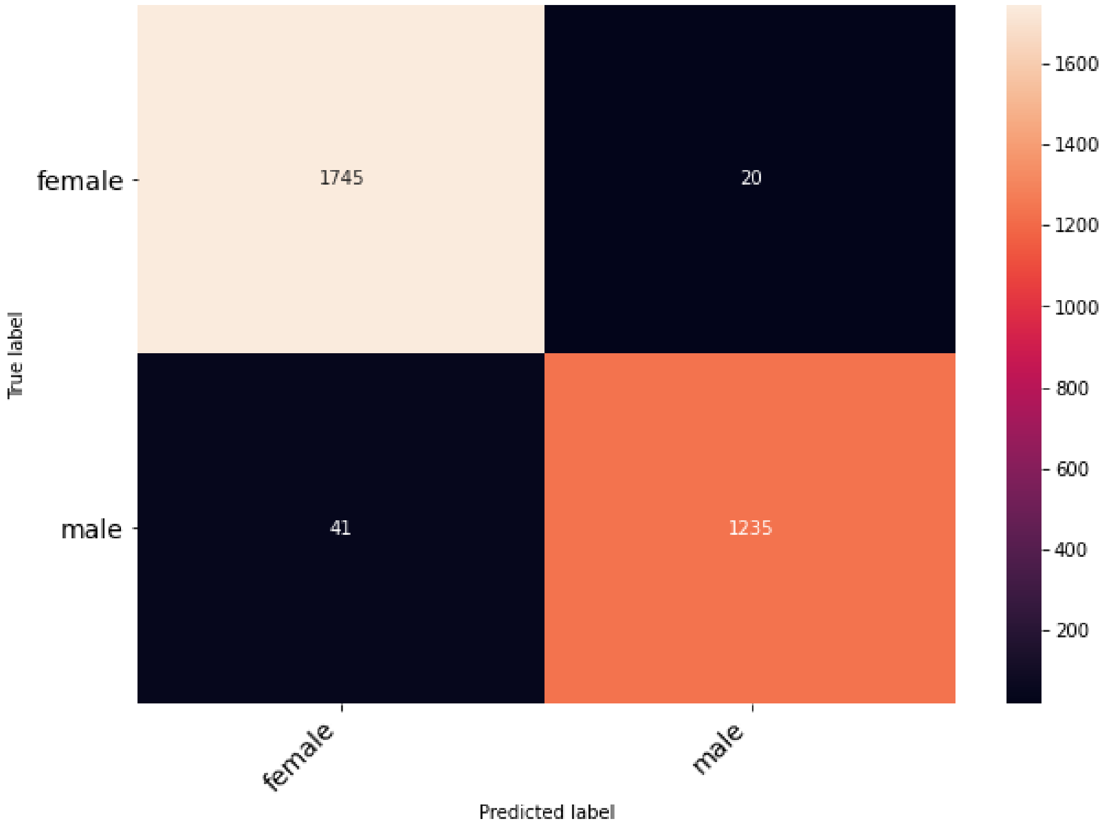


Figure 5.18

Confusion matrix Gender accuracy of 2D CNN model with MFCC

Emotion Accuracy for 2D CNN with MFCC without data augmentation

For same we also have confusion matrix Gender Accuracy Result for 2D CNN with MFCC without data augmentation.

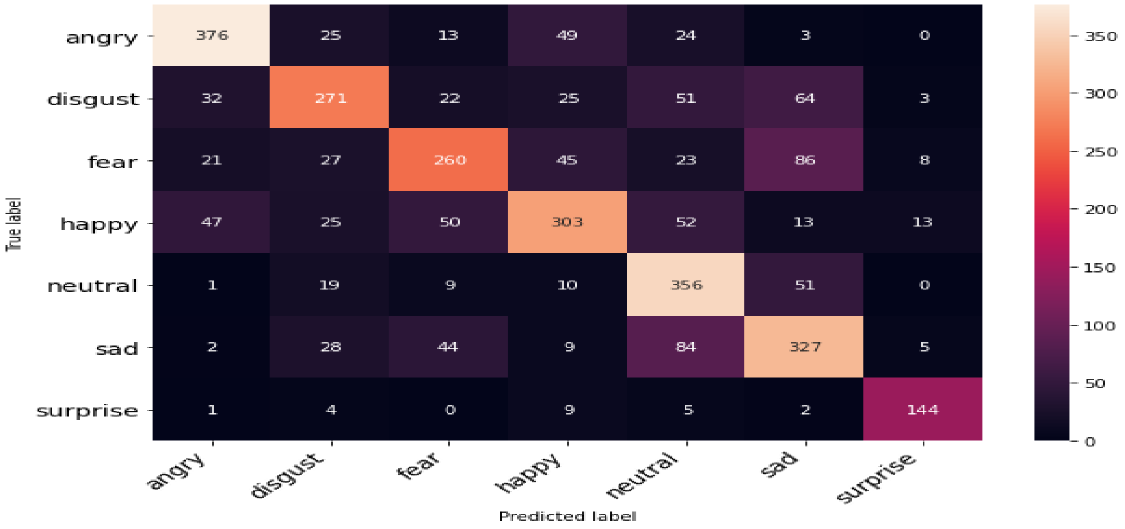


Figure 5.19

Classification report Emotion Accuracy Result of 2D CNN model with MFCC

This is the Loss and train graph for our 2D CNN with MFCC without augmentation over 20 Epoch.

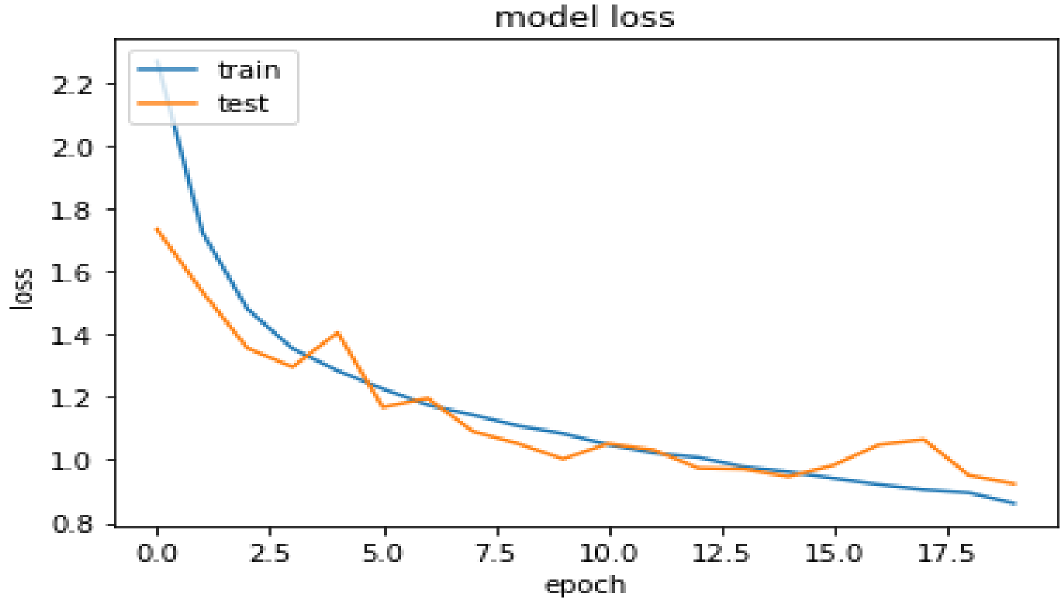


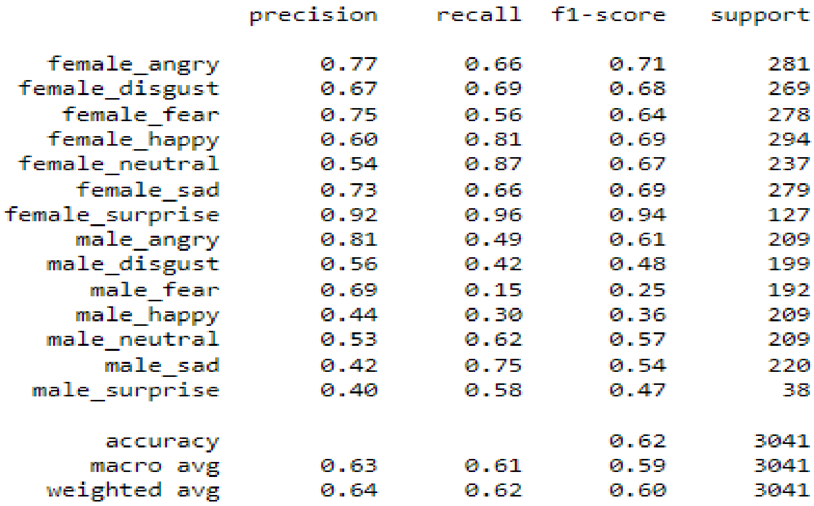
Figure 5.20

Loss curves for 2D CNN model with MFCC feature on combine dataset

5.5 2D CNN with MFCC and data Augmentation

Emotion by gender accuracy for 2D CNN with MFCC and data Augmentation

This is the Classification report for Emotion by gender accuracy for 2D CNN with MFCC and data Augmentation. Where we got the weight accuracy of 64%. In this we are Using our model to identify gender and emotion both.

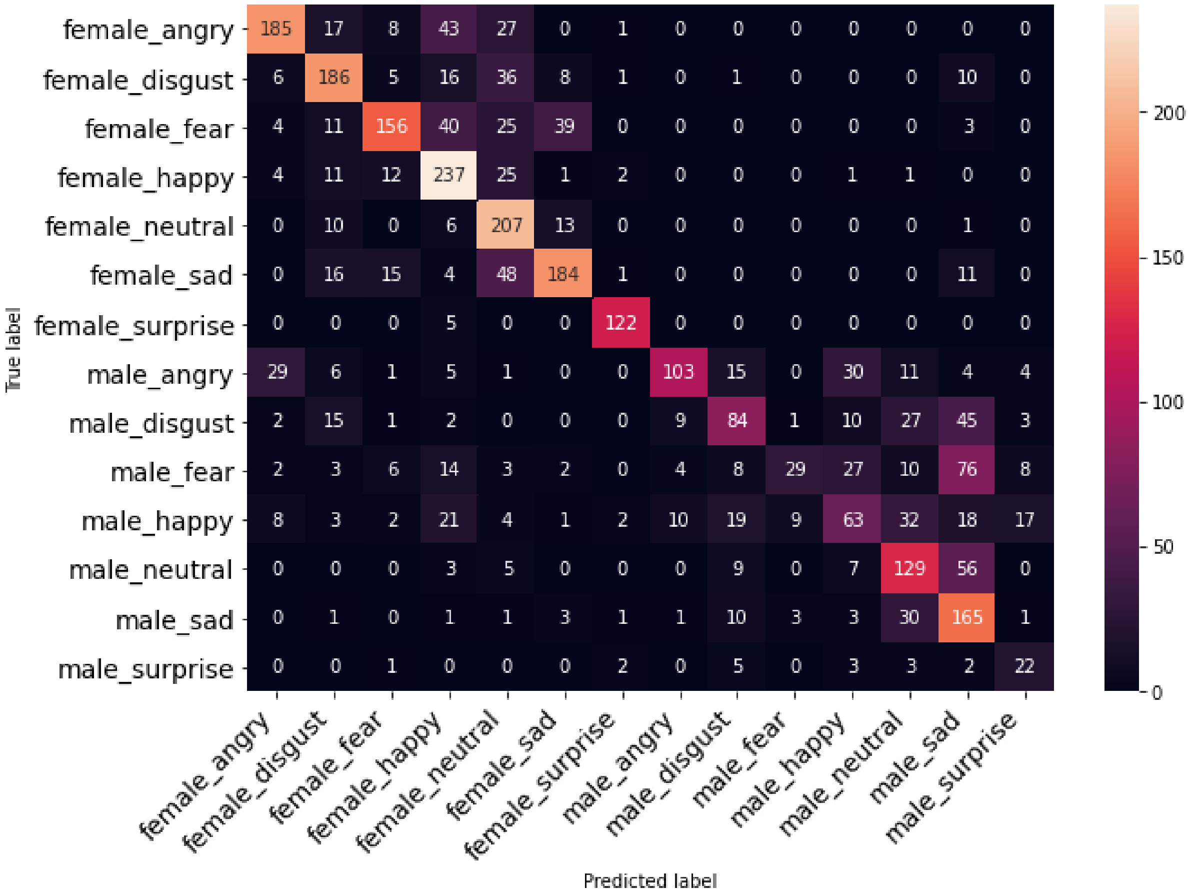


```
                 precision    recall  f1-score   support

   female_angry       0.77      0.66      0.71       281
 female_disgust       0.67      0.69      0.68       269
    female_fear       0.75      0.56      0.64       278
   female_happy       0.60      0.81      0.69       294
 female_neutral       0.54      0.87      0.67       237
     female_sad       0.73      0.66      0.69       279
female_surprise       0.92      0.96      0.94       127
     male_angry       0.81      0.49      0.61       209
   male_disgust       0.56      0.42      0.48       199
      male_fear       0.69      0.15      0.25       192
     male_happy       0.44      0.30      0.36       209
   male_neutral       0.53      0.62      0.57       209
       male_sad       0.42      0.75      0.54       220
  male_surprise       0.40      0.58      0.47        38

       accuracy                           0.62      3041
      macro avg       0.63      0.61      0.59      3041
   weighted avg       0.64      0.62      0.60      3041
```

Figure 5.21

Classification report of Emotion by gender accuracy of 2D CNN model with MFCC mean and Data augmentation

For same we also have confusion matrix for Emotion by gender accuracy for 2D CNN with MFCC and data Augmentation.



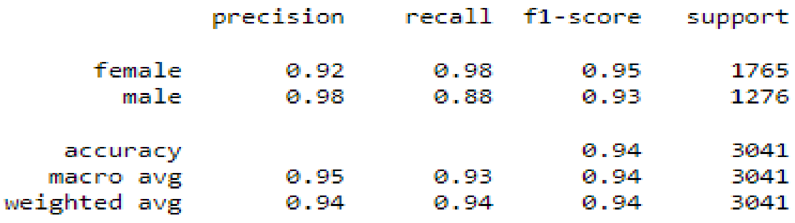
True label \ Predicted label	female_angry	female_disgust	female_fear	female_happy	female_neutral	female_sad	female_surprise	male_angry	male_disgust	male_fear	male_happy	male_neutral	male_sad	male_surprise
female_angry	185	17	8	43	27	0	1	0	0	0	0	0	0	0
female_disgust	6	186	5	16	36	8	1	0	1	0	0	0	10	0
female_fear	4	11	156	40	25	39	0	0	0	0	0	0	3	0
female_happy	4	11	12	237	25	1	2	0	0	0	1	1	0	0
female_neutral	0	10	0	6	207	13	0	0	0	0	0	0	1	0
female_sad	0	16	15	4	48	184	1	0	0	0	0	0	11	0
female_surprise	0	0	0	5	0	0	122	0	0	0	0	0	0	0
male_angry	29	6	1	5	1	0	0	103	15	0	30	11	4	4
male_disgust	2	15	1	2	0	0	0	9	84	1	10	27	45	3
male_fear	2	3	6	14	3	2	0	4	8	29	27	10	76	8
male_happy	8	3	2	21	4	1	2	10	19	9	63	32	18	17
male_neutral	0	0	0	3	5	0	0	0	9	0	7	129	56	0
male_sad	0	1	0	1	1	3	1	1	10	3	3	30	165	1
male_surprise	0	0	1	0	0	0	2	0	5	0	3	3	2	22

Figure 5.22

Confusion matrix of Emotion by gender accuracy of 2D CNN model with MFCC mean and Data augmentation

Gender Accuracy Result for 2D CNN with MFCC and data Augmentation

This is the Classification report for Gender Accuracy Result for 2D CNN with MFCC without data augmentation. Where we got the weight accuracy of 94%. In this we are Using our model to identify gender only without any emotion label added to it.



```
              precision    recall  f1-score   support

      female       0.92      0.98      0.95      1765
        male       0.98      0.88      0.93      1276

    accuracy                           0.94      3041
   macro avg       0.95      0.93      0.94      3041
weighted avg       0.94      0.94      0.94      3041
```

Figure 5.22

Classification report Gender Accuracy Result of 2D CNN model with MFCC and Data augmentation

For same we also have confusion matrix for Gender Accuracy Result for 2D CNN with MFCC and data Augmentation.

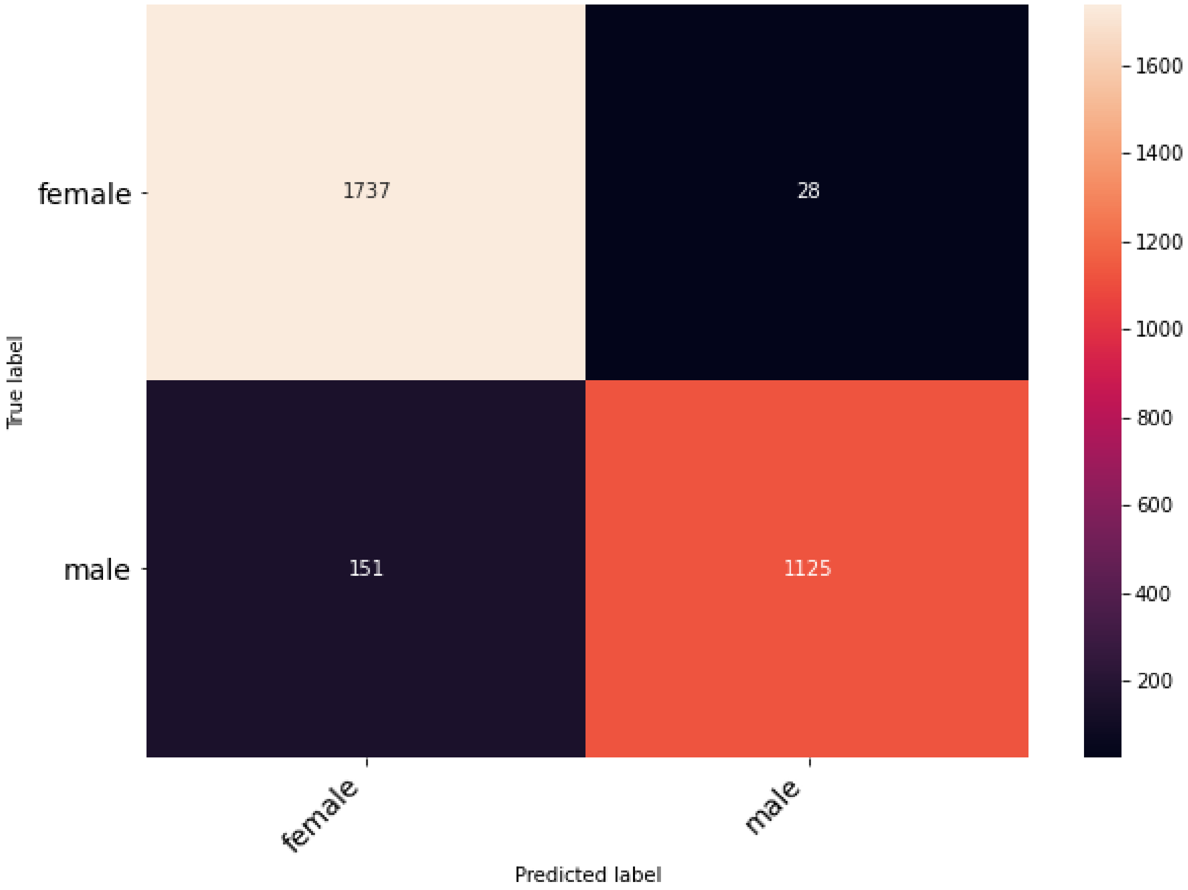
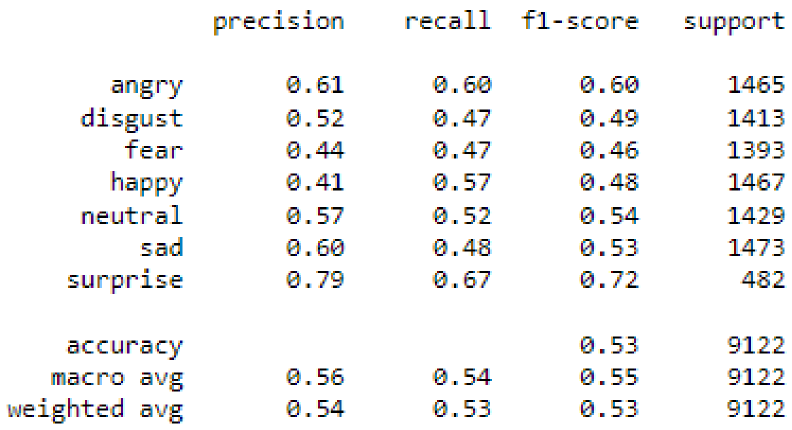


Figure 5.22

Confusion matrix Gender accuracy of 1D CNN model with MFCC and Data augmentation

Emotion Accuracy for 2D CNN with MFCC and data Augmentation

This is the Classification report for Emotion Accuracy for 2D CNN with MFCC and data Augmentation. Where we got the weight accuracy of 54%. In this we are Using our model to identify emotion only without any gender label added to it.



```
              precision    recall  f1-score   support

       angry       0.61      0.60      0.60      1465
     disgust       0.52      0.47      0.49      1413
        fear       0.44      0.47      0.46      1393
       happy       0.41      0.57      0.48      1467
     neutral       0.57      0.52      0.54      1429
         sad       0.60      0.48      0.53      1473
    surprise       0.79      0.67      0.72       482

    accuracy                           0.53      9122
   macro avg       0.56      0.54      0.55      9122
weighted avg       0.54      0.53      0.53      9122
```

Figure 5.23

Classification report Emotion accuracy of 2D CNN model with MFCC mean and Data augmentation

For same we also have confusion matrix for Emotion Accuracy for 2D CNN with MFCC and data Augmentation.

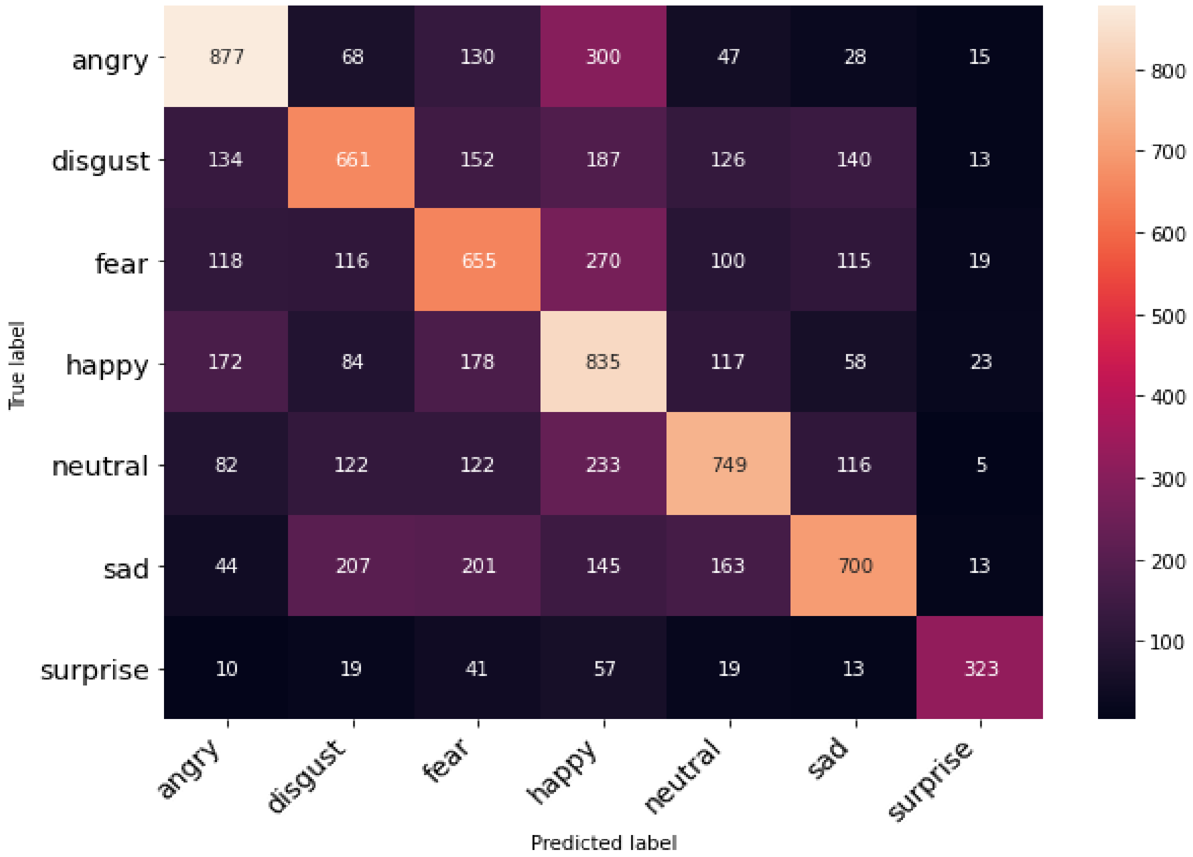


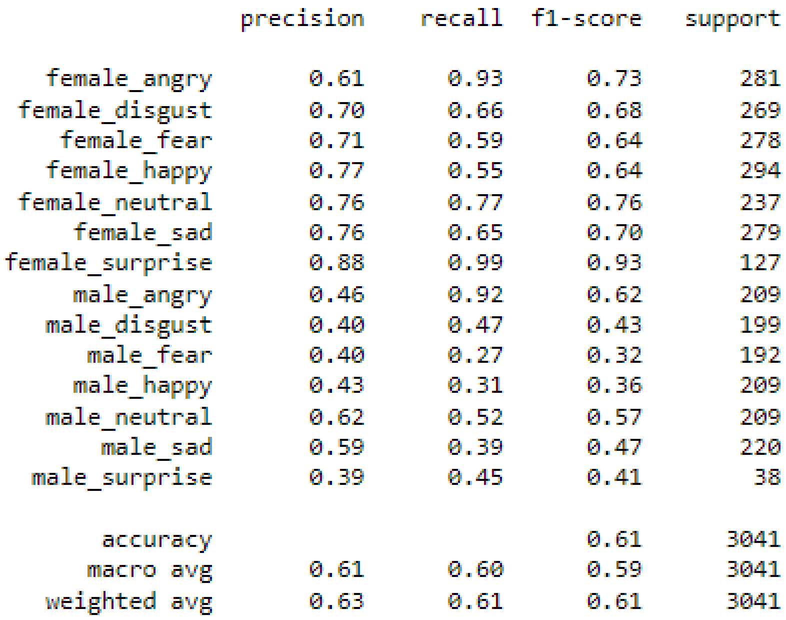
Figure 5.24

Confusion matrix Emotion accuracy of 2D CNN model with MFCC and Data augmentation

5.6 2D CNN Log-melspectogram without Data augmentation

Emotion by gender accuracy for 2D CNN Log-melspectogram without Data augmentation

This is the Classification report for Emotion by gender accuracy for 2D CNN Log-melspectogram without Data augmentation. Where we got the weight accuracy of 63%. In this we are Using our model to identify gender and emotion both.

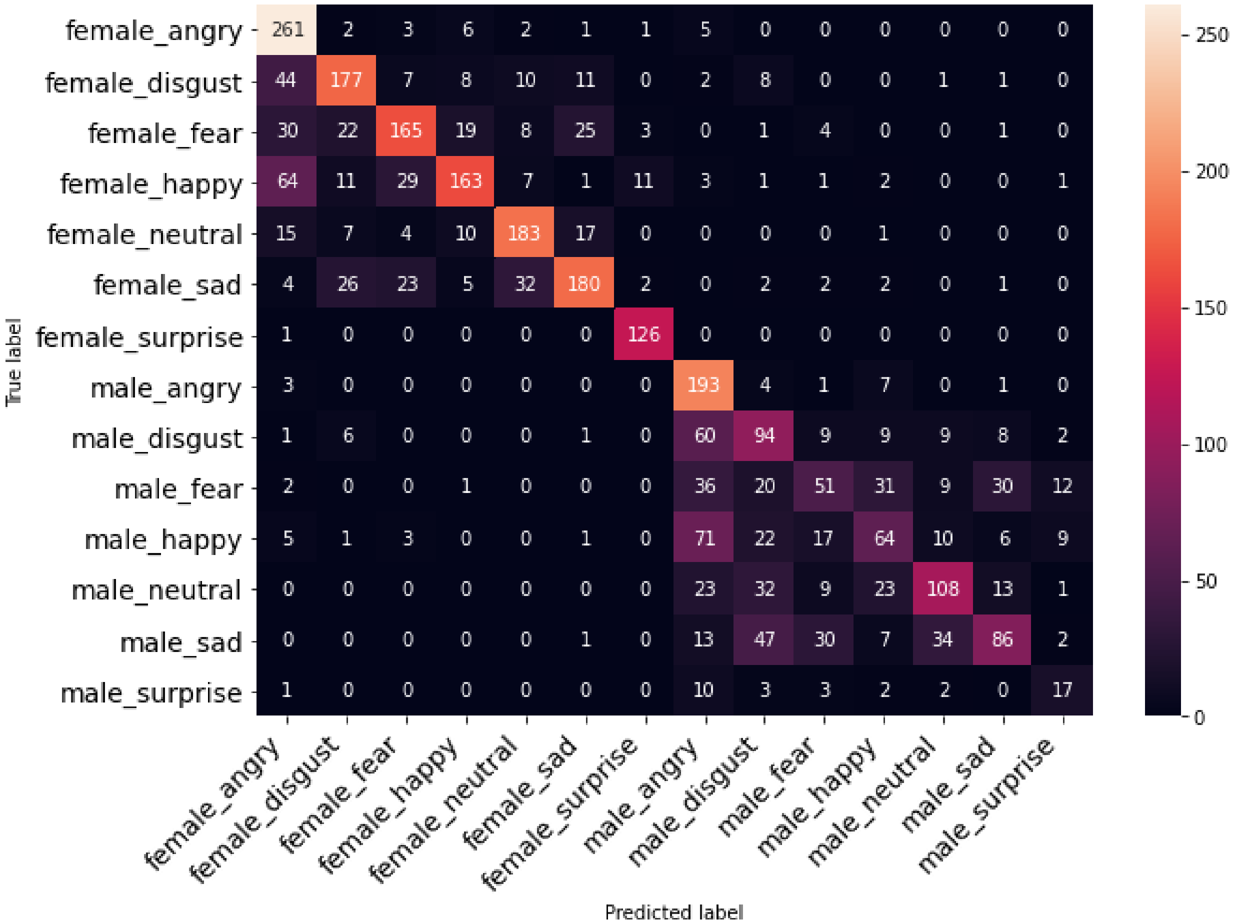


```
                 precision    recall  f1-score   support

   female_angry       0.61      0.93      0.73       281
 female_disgust       0.70      0.66      0.68       269
    female_fear       0.71      0.59      0.64       278
   female_happy       0.77      0.55      0.64       294
 female_neutral       0.76      0.77      0.76       237
     female_sad       0.76      0.65      0.70       279
female_surprise       0.88      0.99      0.93       127
     male_angry       0.46      0.92      0.62       209
   male_disgust       0.40      0.47      0.43       199
      male_fear       0.40      0.27      0.32       192
     male_happy       0.43      0.31      0.36       209
   male_neutral       0.62      0.52      0.57       209
       male_sad       0.59      0.39      0.47       220
  male_surprise       0.39      0.45      0.41        38

       accuracy                           0.61      3041
      macro avg       0.61      0.60      0.59      3041
   weighted avg       0.63      0.61      0.61      3041
```

Figure 5.25

Classification report Emotion by gender accuracy for 2D CNN Log-melspectogram without Data augmentation

For same we also have confusion matrix for Emotion by gender accuracy for 2D CNN Log-melspectogram without Data augmentation.



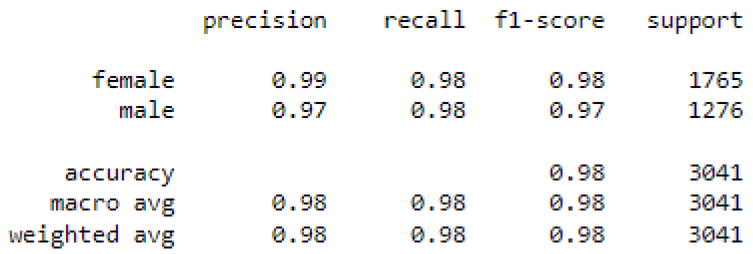
	female_angry	female_disgust	female_fear	female_happy	female_neutral	female_sad	female_surprise	male_angry	male_disgust	male_fear	male_happy	male_neutral	male_sad	male_surprise
female_angry	261	2	3	6	2	1	1	5	0	0	0	0	0	0
female_disgust	44	177	7	8	10	11	0	2	8	0	0	1	1	0
female_fear	30	22	165	19	8	25	3	0	1	4	0	0	1	0
female_happy	64	11	29	163	7	1	11	3	1	1	2	0	0	1
female_neutral	15	7	4	10	183	17	0	0	0	0	1	0	0	0
female_sad	4	26	23	5	32	180	2	0	2	2	2	0	1	0
female_surprise	1	0	0	0	0	0	126	0	0	0	0	0	0	0
male_angry	3	0	0	0	0	0	0	193	4	1	7	0	1	0
male_disgust	1	6	0	0	0	1	0	60	94	9	9	9	8	2
male_fear	2	0	0	1	0	0	0	36	20	51	31	9	30	12
male_happy	5	1	3	0	0	1	0	71	22	17	64	10	6	9
male_neutral	0	0	0	0	0	0	0	23	32	9	23	108	13	1
male_sad	0	0	0	0	0	1	0	13	47	30	7	34	86	2
male_surprise	1	0	0	0	0	0	0	10	3	3	2	2	0	17

Figure 5.26

Confusion matrix for Emotion by gender accuracy for 2D CNN Log-melspectogram without Data augmentation

Gender Accuracy Result for 2D CNN Log-melspectogram without Data augmentation

This is the Classification report for Gender Accuracy Result for 2D CNN Log-melspectogram without Data augmentation. Where we got the weight accuracy of 98%. In this we are Using our model to identify gender only without any emotion label added to it.



```
              precision    recall  f1-score   support

      female       0.99      0.98      0.98      1765
        male       0.97      0.98      0.97      1276

    accuracy                           0.98      3041
   macro avg       0.98      0.98      0.98      3041
weighted avg       0.98      0.98      0.98      3041
```

Figure 5.27

Gender Accuracy Result for 2D CNN Log-melspectogram without Data augmentation

For same we also have confusion matrix for Gender Accuracy Result for 2D CNN Log-melspectogram without Data augmentation.

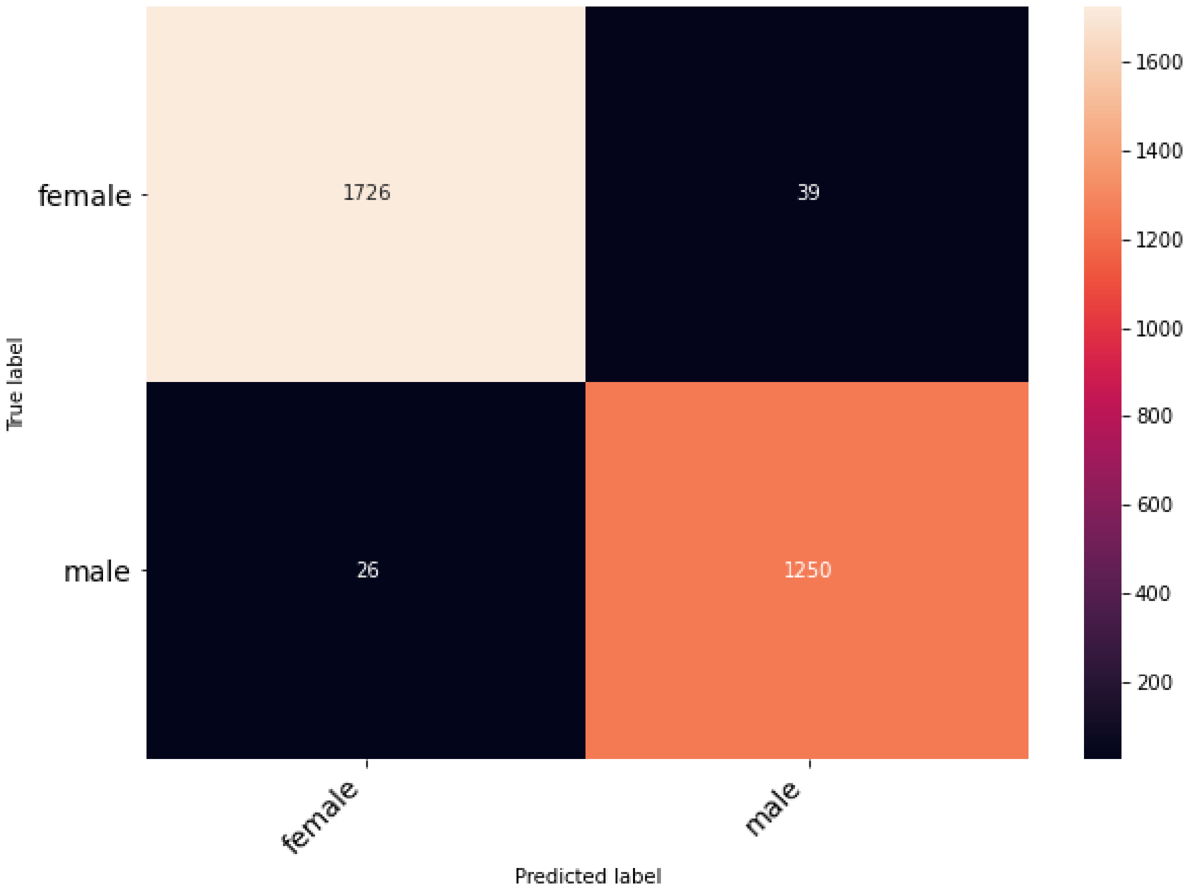
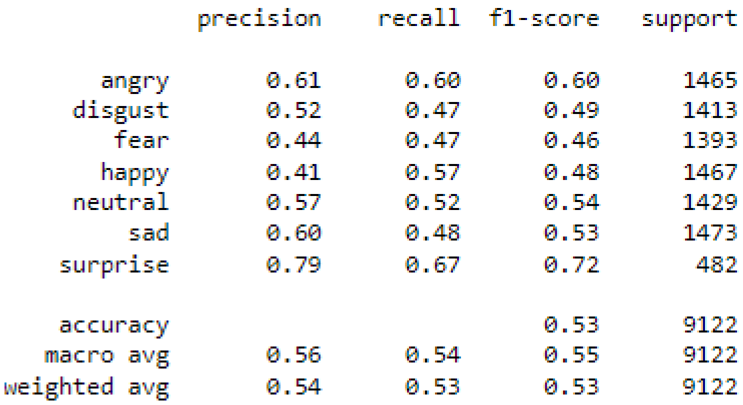


Figure 5.28

confusion matrix for Gender Accuracy Result for 2D CNN Log-melspectogram without Data augmentation

Emotion Accuracy for 2D CNN Log-melspectogram without Data augmentation

This is the Classification report for Emotion Accuracy for 2D CNN Log-melspectogram without Data augmentation. Where we got the weight accuracy of 54%. In this we are Using our model to identify emotion only without any gender label added to it.



```
              precision    recall  f1-score   support

       angry       0.61      0.60      0.60      1465
     disgust       0.52      0.47      0.49      1413
        fear       0.44      0.47      0.46      1393
       happy       0.41      0.57      0.48      1467
     neutral       0.57      0.52      0.54      1429
         sad       0.60      0.48      0.53      1473
    surprise       0.79      0.67      0.72       482

    accuracy                           0.53      9122
   macro avg       0.56      0.54      0.55      9122
weighted avg       0.54      0.53      0.53      9122
```

Figure 5.29

Classification report for Emotion Accuracy for 2D CNN Log-melspectogram without Data augmentation

For same we also have confusion matrix for Emotion Accuracy for 2D CNN Log-melspectogram without Data augmentation.

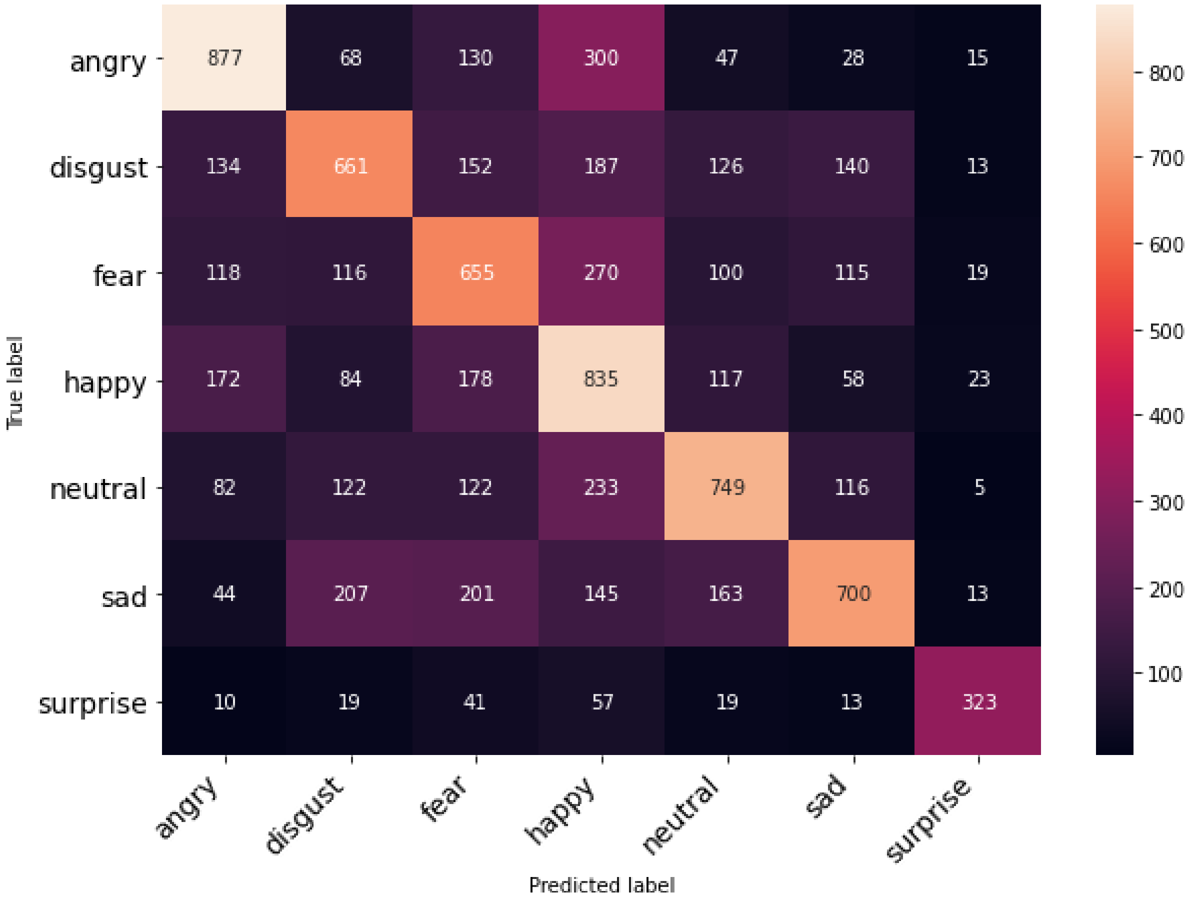


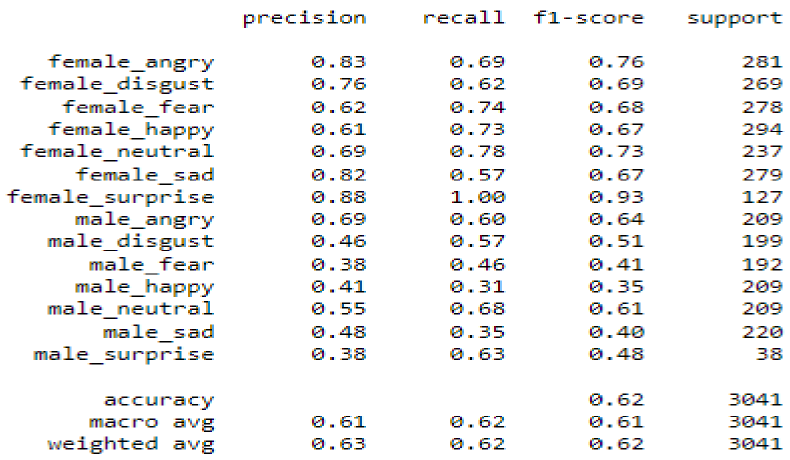
Figure 5.30

confusion matrix for Emotion Accuracy for 2D CNN Log-melspectogram without Data augmentation

5.7 2D CNN Log-melspectogram with Data augmentation

Emotion by gender accuracy for 2D CNN Log-melspectogram with Data augmentation

This is the Classification report for Emotion Accuracy for 2D CNN with MFCC and data Augmentation. Where we got the weight accuracy of 65%. In this we are Using our model to identify gender only without any emotion label added to it.

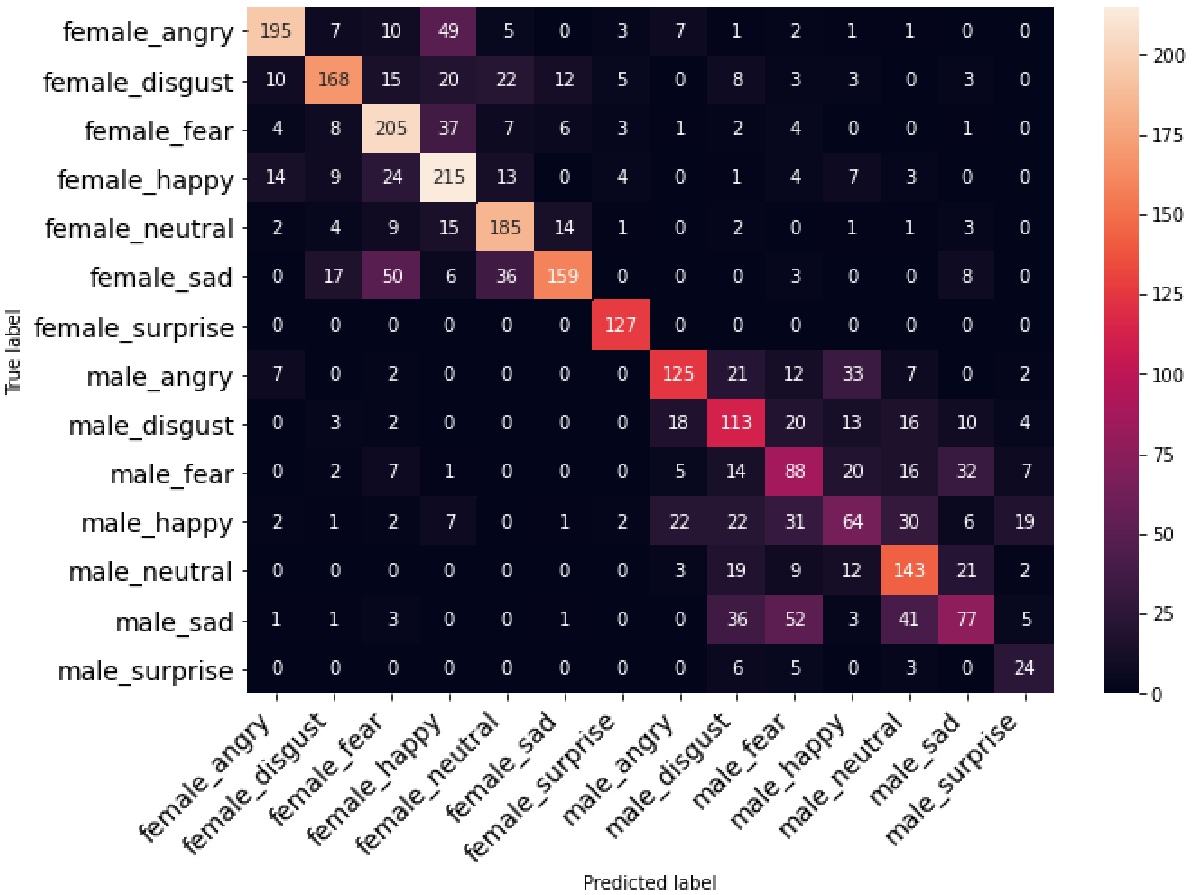


```
                 precision    recall  f1-score   support

   female_angry       0.83      0.69      0.76       281
 female_disgust       0.76      0.62      0.69       269
    female_fear       0.62      0.74      0.68       278
   female_happy       0.61      0.73      0.67       294
 female_neutral       0.69      0.78      0.73       237
     female_sad       0.82      0.57      0.67       279
female_surprise       0.88      1.00      0.93       127
     male_angry       0.69      0.60      0.64       209
   male_disgust       0.46      0.57      0.51       199
      male_fear       0.38      0.46      0.41       192
     male_happy       0.41      0.31      0.35       209
   male_neutral       0.55      0.68      0.61       209
       male_sad       0.48      0.35      0.40       220
  male_surprise       0.38      0.63      0.48        38

       accuracy                           0.62      3041
      macro avg       0.61      0.62      0.61      3041
   weighted avg       0.63      0.62      0.62      3041
```

Figure 5.31

Emotion Accuracy for 2D CNN with MFCC and data Augmentation

For same we also have confusion matrix for Emotion Accuracy for 2D CNN with MFCC and data Augmentation.



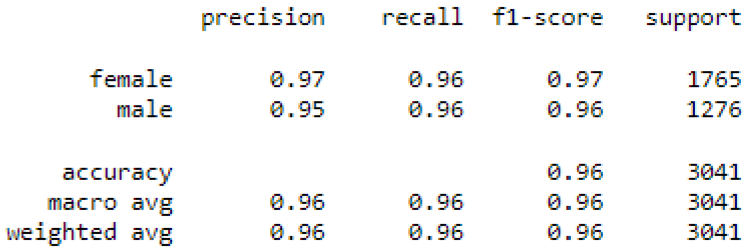
True label \ Predicted label	female_angry	female_disgust	female_fear	female_happy	female_neutral	female_sad	female_surprise	male_angry	male_disgust	male_fear	male_happy	male_neutral	male_sad	male_surprise
female_angry	195	7	10	49	5	0	3	7	1	2	1	1	0	0
female_disgust	10	168	15	20	22	12	5	0	8	3	3	0	3	0
female_fear	4	8	205	37	7	6	3	1	2	4	0	0	1	0
female_happy	14	9	24	215	13	0	4	0	1	4	7	3	0	0
female_neutral	2	4	9	15	185	14	1	0	2	0	1	1	3	0
female_sad	0	17	50	6	36	159	0	0	0	3	0	0	8	0
female_surprise	0	0	0	0	0	0	127	0	0	0	0	0	0	0
male_angry	7	0	2	0	0	0	0	125	21	12	33	7	0	2
male_disgust	0	3	2	0	0	0	0	18	113	20	13	16	10	4
male_fear	0	2	7	1	0	0	0	5	14	88	20	16	32	7
male_happy	2	1	2	7	0	1	2	22	22	31	64	30	6	19
male_neutral	0	0	0	0	0	0	0	3	19	9	12	143	21	2
male_sad	1	1	3	0	0	1	0	0	36	52	3	41	77	5
male_surprise	0	0	0	0	0	0	0	0	6	5	0	3	0	24

Figure 5.32

confusion matrix for Emotion Accuracy for 2D CNN with MFCC and data Augmentation

Gender Accuracy Result for 2D CNN Log-melspectogram with Data augmentation

This is the Classification report for Emotion Accuracy for 2D CNN with MFCC and data Augmentation. Where we got the weight accuracy of 65%. In this we are Using our model to identify gender only without any emotion label added to it.



```
              precision    recall  f1-score   support

      female       0.97      0.96      0.97      1765
        male       0.95      0.96      0.96      1276

    accuracy                           0.96      3041
   macro avg       0.96      0.96      0.96      3041
weighted avg       0.96      0.96      0.96      3041
```

Figure 5.33

Classification report for Emotion Accuracy for 2D CNN with MFCC and data Augmentation

For same we also have confusion matrix for Emotion Accuracy for 2D CNN with MFCC and data Augmentation.

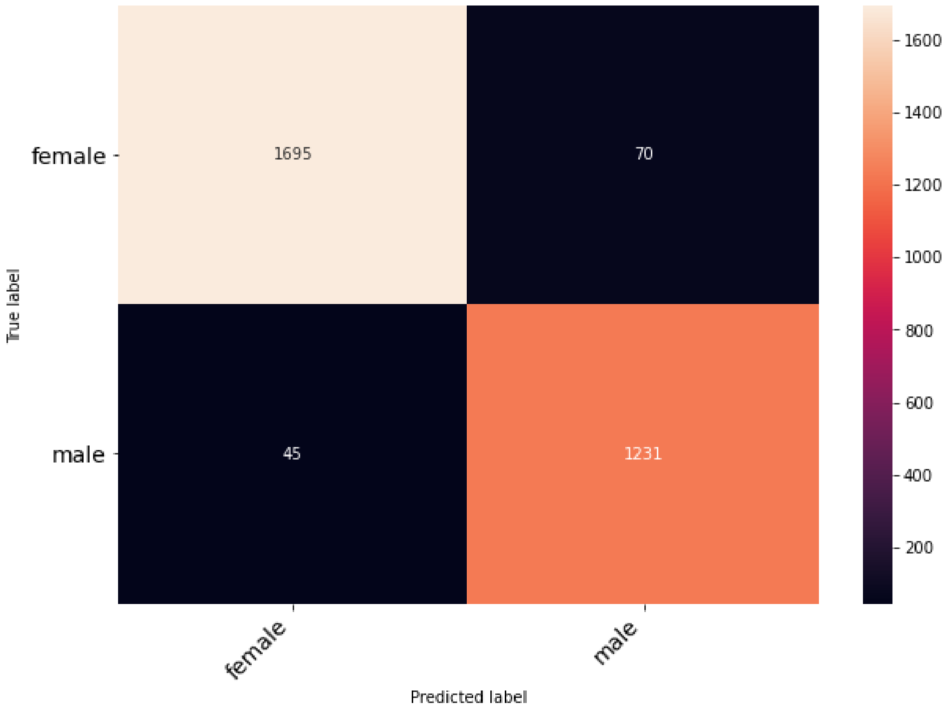
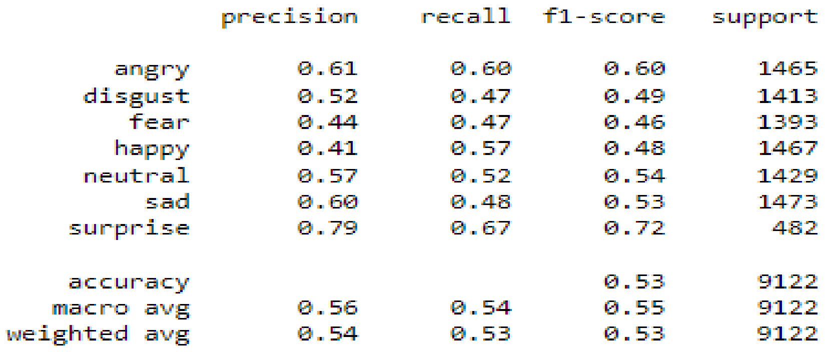


Figure 5.34

confusion matrix for Emotion Accuracy for 2D CNN with MFCC and data Augmentation

Emotion Accuracy for 2D CNN Log-melspectogram with Data augmentation

This is the Classification report for Emotion Accuracy for 2D CNN with MFCC and data Augmentation. Where we got the weight accuracy of 65%. In this we are Using our model to identify gender only without any emotion label added to it.



```
              precision    recall  f1-score   support

       angry       0.61      0.60      0.60      1465
     disgust       0.52      0.47      0.49      1413
        fear       0.44      0.47      0.46      1393
       happy       0.41      0.57      0.48      1467
     neutral       0.57      0.52      0.54      1429
         sad       0.60      0.48      0.53      1473
    surprise       0.79      0.67      0.72       482

    accuracy                           0.53      9122
   macro avg       0.56      0.54      0.55      9122
weighted avg       0.54      0.53      0.53      9122
```

Figure 5.35

Classification report for Emotion Accuracy for 2D CNN with MFCC and data Augmentation

For same we also have confusion matrix for Emotion Accuracy for 2D CNN with MFCC and data Augmentation.

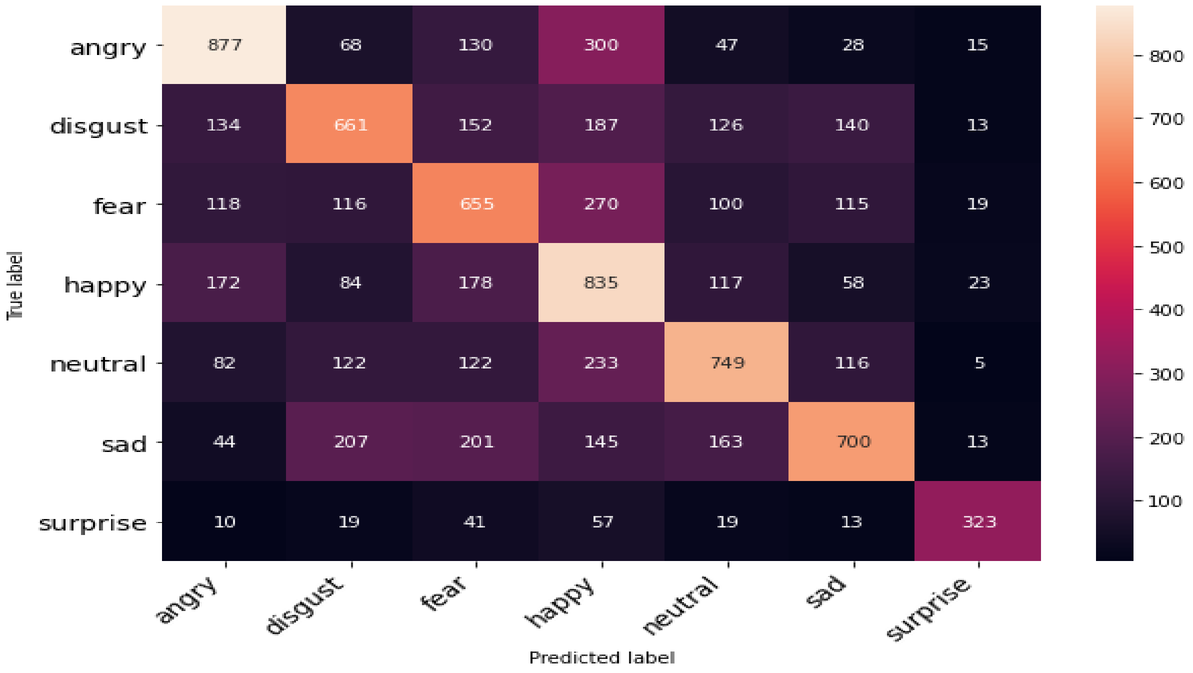


Figure 5.36

Confusion matrix for Emotion Accuracy for 2D CNN with MFCC and data Augmentation

Chapter 6
CONCLUSIONS AND FUTURE ASPECTS

The new era of automation has begun as a result of the increasing growth and development in the fields of AI and machine learning. The majority of these automated gadgets are controlled by the user's vocal commands. Many advantages can be created over present systems if, in addition to identifying words, the machines can interpret the speaker's emotion (user). Computer-based tutorial applications, automated contact centre dialogues, a diagnostic tool for therapy, and an automatic translation system are some of the applications of a speech emotion detection system. The processes for creating a speech emotion recognition system were addressed in detail in this thesis, and various tests were conducted to understand the influence of each step. The low amount of publicly available speech databases made it difficult at first.

The processes for creating a speech emotion recognition system were addressed in detail in this thesis, and various tests were conducted to understand the influence of each step. It was initially difficult to create a well-trained model due to the low amount of publicly available speech databases. Then, in previous research, multiple unique ways to feature extraction had been offered, and identifying the optimal approach required doing numerous experiments. Finally, learning about the strengths and weaknesses of each classifying algorithm in terms of emotion recognition was part of the classifier selection process. When comparing a single feature to an integrated feature space, the results show that an integrated feature space produces a higher recognition rate.

Deep convolutional architectures that can learn from spectrogram representations of speech are becoming increasingly popular. They, along with recurrent networks, are thought to be a solid foundation for SER systems are a type of SER system. Over time, increasingly complicated SER designs have been built. a focus on eliciting emotionally compelling local and global situations.

In future research, a large quantity of data will be necessary to correctly develop deep running-based models. Speech emotion recognition studies employ significantly less data than video or text-based investigations. In order to boost performance and generalisation, high-quality databases are required. To discern emotions from speech, extensive datasets must be used to create an accurate mapping of acoustic parameters and emotional intensity. Despite the fact that several acoustic characteristics have been employed in investigations, researchers still don't know what emotional intensity is displayed when a given feature is identified at a certain level. These issues must be resolved in order to more correctly perceive emotions and their strength.

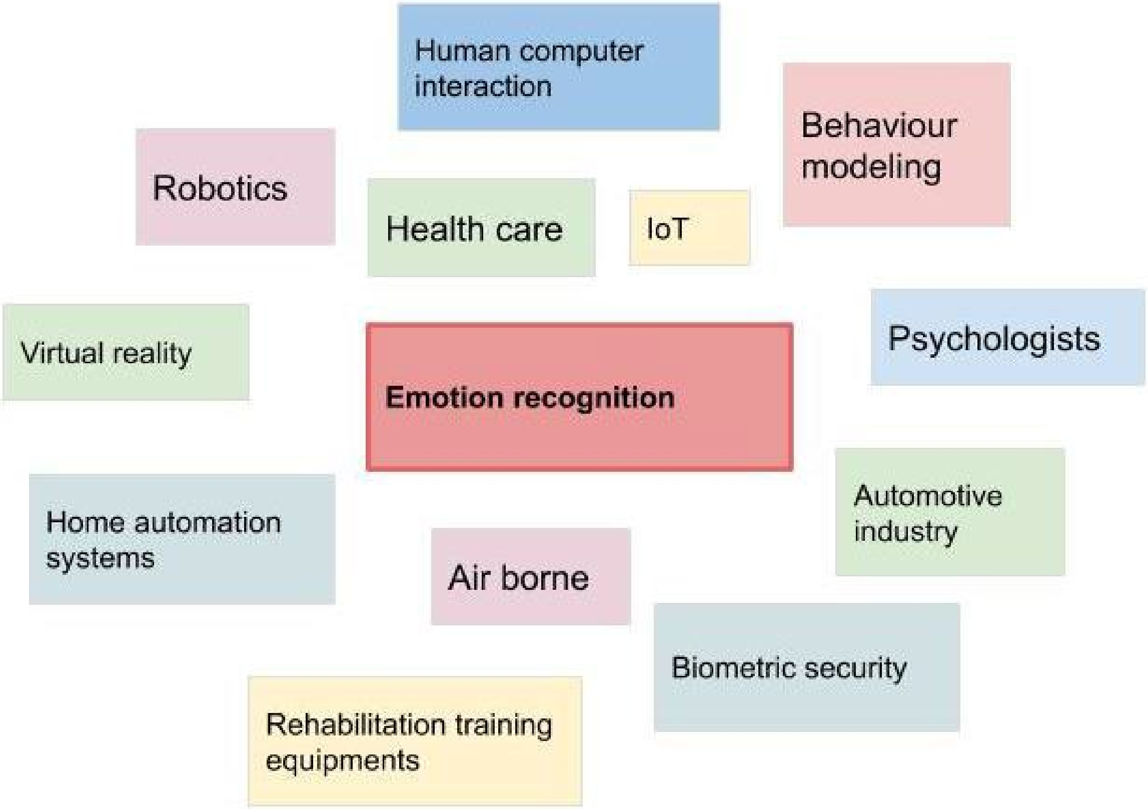


Figure 6.1
Applications of SER

REFERENCES

[1] A. Mujaddidurrahman, F. Ernawan, A. Wibowo, E. A. Sarwoko, A. Sugiharto and M. D. R. Wahyudi, "Speech Emotion Recognition Using 2D-CNN with Data Augmentation," 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), 2021, pp. 685-689, doi: 10.1109/ICSECS52883.2021.00130.

[2] M. Ghai, S. Lal, S. Duggal and S. Manik, "Emotion recognition on speech signals using machine learning," 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), 2017, pp. 34-39, doi: 10.1109/ICBDACI.2017.8070805.

[3] T.Sai Samhith, G.Nishika, M.Prayuktha, M.Bharat Chandra, Dr.Sunil Bhutada, G.Prasadu IJCRT2106108 International Journal of Creative Research Thoughts (IJCRT) www.ijcrt.org a815.

[4] Kerkeni, Leila & Serrestou, Youssef & Mbarki, Mohamed & Raoof, Kosai & Mahjoub, Mohamed. (2018). Speech Emotion Recognition: Methods and Cases Study. 175-182. 10.5220/0006611601750182.

[5] arXiv:1904.06022 [cs.LG] (or arXiv:1904.06022v1 [cs.LG] for this version) https://doi.org/10.48550/arXiv.1904.06022.

[6] Learning Speech Emotion Representations in the Quaternion Domain E. Guizzo, Tillman Weyde, Simone Scardapane, D. Comminiello Computer Science ArXiv 2022 TLDR.

[8] Björn W. Schuller Communications of the ACM, May 2018, Vol. 61 No. 5, Pages 90-99 10.1145/3129340.

[9] P. Thirumal, L. Shylu Dafni Agnus, N. Sanjana, K. Yogajeeva and V. Namitha, "Speech Recognition Using Data Augmentation," 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), 2021, pp. 1-5, doi: 10.1109/ICAECA52838.2021.9675607.

[12] Mingyi Chen, Xuanji He, Jing Yang, Han Zhang, "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition", IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440-1444, 2018.

[13] S. Zhang, S. Zhang, T. Huang and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," in IEEE Transactions on Multimedia, vol. 20, no. 6, pp. 1576-1590, June 2018, doi: 10.1109/TMM.2017.2766843.

[14] National Natural Science Foundation of China (No: 61271309) and Shenzhen Science & Technology Research Program (No: JC201105170727A).

[15] Kun Han1, Dong Yu, Ivan Tashev Department of Computer Science and Engineering, The Ohio State University, Columbus, 43210, OH, USA Microsoft Research.

[16]. The Impact of Attention Mechanisms on Speech Emotion Recognition Shouyan Chen, Mingyan Zhang, Xiaofen Yang, Zhijia Zhao, T. Zou, Xinqi Sun Computer Science Sensors 2021.

[17] A DCRNN-based ensemble classifier for speech emotion recognition in Odia language Monorama Swain, Bubai Maji, P. Kabisatpathy, A. Routray Computer Science Complex &amp; Intelligent Systems 2022.

[18] A. Pon Bharathi, M. Ramachandran, Ramu Kurinjimalar, Sriram Soniya Art Design, Modelling and Fabrication of Advanced Robots 2022.

[19] International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072.

[21] Speech Technology Progress Based on New Machine Learning Paradigm Vlado Delić, Zoran Perić, Milan Sečujski, Nikša Jakovljević, Jelena Nikolić, Dragiša Mišković, Nikola Simić, Siniša Suzić, Tijana Delić Comput Intell Neurosci. 2019; 2019: 4368036. Published online 2019 Jun 25. doi: 10.1155/2019/4368036
PMCID: PMC6614991.

[22] Alif Bin Abdul Qayyum, Asiful Arefeen, C. Shahnaz Computer Science 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON).

[23] Mustaqeem, M. Sajjad, Soonil Kwon Computer Science IEEE Access 2020

[24] T. Wani, T. Gunawan, Syed Asif Ahmad Qadri, H. Mansor, F. Arifin, Y. Ahmad Computer Science 2021 IEEE 7th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA).

[25]. Shujie Zhao, Yan Yang, Jingdong Chen, Lijun Zhang Computer Science 2021 29th European Signal Processing Conference (EUSIPCO) 2021.

[27] Meghna Pandharipande, Rupayan Chakraborty, A. Panda, Sunil Kumar Kopparapu Computer Science 2018 26th European Signal Processing Conference (EUSIPCO) 2018.

[28] Zhentao Liu, A. Rehman, Min Wu, Weihua Cao, Man Hao Computer Science Inf. Sci. 2021.

[29] Minji Seo, Myung-Ho Kim Computer Science Sensors 2020.

[30] Guang-Min Xia, Fan Li, Dong-Di Zhao, Qian Zhang, Song Yang Computer Science 2019 5th International Conference on Big Data Computing and Communications (BIGCOM) 2019.

[31] Website Source – https://developers.google.com/machine-learning/data-prep/transform/normalization.

[32] Website Source – https://flatironschool.com/blog/deep-learning-vs-machine-learning/.

Minor Project Report

On

Emotion Recognition in speech Using Machine Learning

Submitted to

Amity University Uttar Pradesh

In partial fulfilment of the requirements for the award of the degree of

Bachelor of Technology

In

Computer Science & Technology

By

Akshay Kumar (A2305218221)

Under the guidance of

Dr. Sanjay Kumar Dubey

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY

AMITY UNIVERSITY UTTAR PRADESH, NOIDA (U.P.)

2021

DECLARATION

We, Akshay Kumar and Aditya Chandrayan, students of B.Tech (CSE) hereby declare that the project entitled "**Emotion Recognition in speech Using Machine Learning**", which is submitted by me to the Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, in partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Computer Science & Engineering has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition

Place: NOIDA	Akshay Kumar
Date:	(A2305218221)
	Aditya Chandrayan
	(A12405218070)

CERTIFICATE

On the basis of the declaration submitted by Akshay Kumar and Aditya Chandrayan, students of B.Tech (CSE), We hereby certify that the project titled "**Emotion Recognition in speech Using Machine Learning**", which is submitted to the Department of Computer Science & Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, in partial fulfilment of the award of the degree of Bachelor of Technology in Computer Science & Engineering, is an original contribution with existing knowledge and faithful record of work carried out by them under my guidance and supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: NOIDA
Date:

Dr. Sanjay Kumar Dubey
Associate Professor
Dept. of Computer Science & Engineering
Amity School of Engineering and Technology
Amity University Uttar Pradesh

ACKNOWLEDEGEMENT

We take this opportunity to express our profound gratitude and deep regards to my supervisor Dr. Sanjay Kumar Dubey for his exemplary guidance, monitoring, and constant encouragement throughout this project. His invaluable comments on system modelling have always inspired me to refine my work. The blessing, help and guidance by him time to time shall carry us a long way in the journey of life on which us about to embark.

We would also like to thank to our Head of Institute Prof. (Dr.) Abhay Bansal, and all the staff of department of Computer Science & Engineering for their regular encouragement and endeavour's.

We are grateful to Amity University, Noida, Uttar Pradesh, for its time and support, which helped us in this project through various stages. Lastly, We would like to thank almighty, our parents, friends for their constant encouragement without which this assignment would not be possible.

.

Akshay Kumar
A2305218221
Aditya
Chandrayan
A12405218070

CONSENT FORM

This is to certify that I, **Akshay Kumar** student of B. Tech **Computer Science** (Stream/branch) of **2018-2022** (year-batch) presently in the VII Semester at Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida give my consent to include all my personal details,

Akshay Kumar, A2305218221 (Name, Enrolment ID) for all accreditation purposes.

Akshay Kumar
Signature of the Student
Enrolment

Number A2305218221

Abstract

There is an emotion related everything a person does, and recognizing is a much complex task. These emotions play a significant role in understanding how a person feels about something. In this world where humans themselves are the product and they are fed what others want, then analysing these different emotions is of major importance. In this paper we will try different paper, compare algorithms and dataset, and formulate a research methodology. Different people have used techniques ranging from CNN to Fourier parameter. We have studied and concluded a method or ourselves. Different people have used different dataset and the dataset play a significant role in results. Datasets which have voice with loudspeakers along with some background noise can decrease the accuracy drastically. Language and geography also matter. Emotions and reaction pitches and energies are different for people of different culture, nation, and perspective. This makes this task even more complex on a global scale.

TABLE OF CONTENTS

TABLE OF FIGURES

INDEX OF TABLES

1. Introduction

In Current world of technology, Human computer interaction are increasing day by day and becoming more common. As human we interact with computer either physically or through speech for initializing any computation. Humans are emotional being and emotion plays an important role in our life and to express our emotion or feeling we use various medium as a form of communication. Speech has been most common and natural way for humans to exchange, express or interact with each other's. We all know that computer is smarter but dumb at the same time as computer cannot understand emotion between human and computer. The demand for emotion recognition in the speech which is gaining more popularity.

Emotion recognitions have wide range of application such as it can be used at hospital for sociological and mental health purposes, at call center to shield the employees from angry and toxic customers, in gaming industry or social sites to stop heat speech and decreasing toxic behavior and many more.

Speech emotion recognition (SER) aims to identify the emotion in words spoken by the speaker which enabling the computer to understand the human words more accurately and make the experience more human friendly. Speech emotion recognition can also be describes as discovering the emotions of speech spoken by the person. In speech, Emotion Recognition is a great challenge as for understanding the emotions of a particular speaker as emotion is very subjective to an individual and can vary from person to person, also it is needed to extract intrinsic data from his/her speech. Then it is converted into proper data which is later processed.

Pitches and energies of emotion is different for every person. Along with this different nationalities and cultures have completely unique Pitches and energies to their emotion also they have their own way of reacting to a certain thing. Same person can react to same thing with same emotions but with unique styles. Background noises can certainly be a great problem in the dataset. Tracking emotions of elderly people is even more difficult. Spontaneous reactions can be extremely hard to track. Circumstances and many other things invisible and untrickable factors also impact the emotions. Also, a lot of resources are consumed in this process. Both human costs along with the time cost are required. For labelling we need humans to sit and listen to each audio and then label them. This is very costly and error prone process. As emotions are very subjective so human labelling other person's emotions can lead to error. In recent years, the SER has been talking of show and many researchers are working day and night to advance in this field. Newer technologies and methodologies are coming every day to make task easier. Noises are automatically removed. Bigger and bigger datasets are being and too are being specified too much, as different nationalities, cultures, mood, ages etc. are being categorized. All these details will help for bringing this technology on a global scale by removing the barriers which restrict it from being implemented globally. Different sectors are contributing in this as they are realizing its importance and increased research is being done. Huge Datasets are being created for this purpose. This can also improve the real time translations applications, as they do not take in account the emotions of the speakers and give the direct translation of even a sarcasm. Linguistic barriers will be removed by this. Many other previously existing phenomenal applications would also be improved a lot by SER. In this paper we will study different research papers. All of them have used some different approach and datasets. Algorithms ranging from CNN, FP, to SVM are used. Many distinctive features including energy, pitch, MFCC, LPCC, MEDC etc. are also used. EMODB, CASIA and EESDB are the major databases which are used by majority of researchers. The reason could be their huge size. Accuracy and results of all these are given. We will study all these papers and study them and then compare them. Algorithms, databases, accuracy along with time elapsed will also be studies. Then after comparison we will give our own research methodology with the way which we find suitable for us.

2. Literature Review / Related Work

Author name	Paper Name	Dataset	Methodology	Conclusion / Result	Publisher and Year
Wisha Zehra, Abdul Rehman Javed, Zunera Jalil,Habib Ullah Khan and Thippa Reddy Gadekallu	Cross corpus multi-lingual speech emotion recognition using ensemble learning [1]	SAVEE (English), URDU (Urdu), EMO-DB (German) and EMOVO (Italian)	Support Vector Machines (SVM), Sequential Minimal Optimization (SMO), Random Forest with 10 trees and Decision Tree (J48)	The Result show that ensemble learning approach promises better accuracy then other state-of-the-art techniques	Springer Received: 26 Sep 2020 Accepted: 3 Dec 2020 Published: 11 Jan 2021
Mustaqeem, Soonil Kwon	MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach [2]	IEMOCAP and EMO-DB datasets	One-dimensional dilated convolutional neural network (DCNN) with multi-learning strategy and long-term contextual dependencies	IEMOCAP and EMO-DB datasets with high recognition accuracy, 73% and 90%, respectively	Science Direct Received:10 July 2020, revised: 26 Oct 2020, Accepted 26 Oct 2020, Available 31 Oct 2020

Mingke Xu, Fan Zhang, Xiaodong Cui, Wei Zhang	Speech Emotion Recognition with Multiscale Area Attention and Data Augmentation [3]	IEMOCAP (Interactive Emotional dyadic Motion Capture database)	Multiscale area attention in a Deep Convolutional Neural Network and data augmentation with Vocal Tract Length Perturbation (VTLP)	The result of the experiment was 79.34% weighted accuracy (WA) and 77.54% unweighted accuracy (UA)	arXiv.org Accepted by ICASSP 2021 Submitted on 3 Feb 2021
Patrick Meyer, Ziyi Xu and Tim Fingscheidt	Improving Convolutional Recurrent Neural Networks for Speech Emotion Recognition [4]	IEMOCAP and Emo-DB (Berlin database of emotional speech)	Convolutional Neural Network (CNN) followed by OS/SVM (open SMILE toolkit with an SVM), ST-CLDNN and FM-CDNN	The result of their modification to CNN gave them unweighted average recall of 64.5% on average	Published in 2021 IEEE Spoken Language Technology Workshop (SLT) Conference Date Jan 2021
Lin Jiang, Ping Tan, Junfeng Yang, Xingbao Liu and ChaoWag	Speech emotion recognition using emotion perception spectral feature [5]	CASIA, EMODB, FAU AIBO	Emotion perception spectrum is Feature Extraction which is base psychoacoustic model and to test it they used SVM	The result show that proposed feature is better than traditional features like MFCC, TEO, and Breath	Science Direct Received: 16 April 2019 Revised: 10 May 2019 Accepted: 3 June 2019

Sharif Noor Zisad, Mohammad Shahadat Hossain and Karl Andersson.	Speech Emotion Recognition in Neurological Disorders Using Convolutional Neural Network [6]	RAVDESS and custom local dataset	Convolutional neural network (CNN) with the data augmentation method	The results demonstrate that the CNN. model performed better than the mentioned. machine learning techniques.	Springer September 2020
Mao Li, Bo Yangy, Joshua Levyy , Andreas Stolcke , Viktor Rozgic, Spyros Matsoukas, Constantinos Papayiannisy, Daniel Bone, Chao Wang	Contrastive Unsupervised Learning for Speech Emotion Recognition [7]	IEMOCAP and MSPPodcast dataset	Contrastive Predictive Coding (CPC) Method	The result show that CPC can learn useful features from unlabelled speech corpora that benefit emotion recognition.	Publisher: IEEE Submitted on 12 Feb 2021

Bagus Tris Atmaja, Masato Akagi	Evaluation of Error and Correlation-Based Loss Functions for Multitask Learning Dimensional Speech Emotion Recognition [8]	IEMOCAP and MSP-IMPROV datasets	Two Loss functions: Error-based and Correlation-based loss functions with mean squared error (MSE) and mean absolute error (MAE)	CCC loss is better than MSE and MAE for multitask learning dimensional emotion recognition	Journal of Physics: Conference Series Submitted on 24 Mar 2020, last revised 18 Nov 2020
Anusha Koduru, Hima Bindu Valiveti & Anil Kumar Budati	Feature extraction algorithms to improve the speech emotion recognition rate [9]	RAVDESS	Three Algorithm used are SVM, Decision tree and LDA are used	Decision tree gives 85%, SVM gives 70% and LDA gives 65% accuracy	Springer Received: 16 Sep 2019 / Accepted: 3 Jan 2020 / Published online: 14 Jan 2020
Premjeet Singh, Goutam Saha, Md Sahidullah	Deep scattering network for speech emotion recognition [10]	EmoDB, IEMOCAP and RAVDESS	They used radial basis function (RBF) kernel-based support vector machine (SVM)	The low value of wavelets per octave generates coarse level frequency domain information improving SER performance	European Signal Processing Conference Submitted on 11 May 2021

3. Proposed Method

There are countless number of ways to communicate with each other in the society, but Speech is one of the fastest and most common way to communicate with each other and to exchange emotion with them. For us to empower the computer to better understand human communication is to start with identifying the real meaning behind what is said and to do so we need to identify the emotion buried under what is said for more natural communication between human and computer.

We can identify the emotion in speech by identifying one or more of the classes of feature namely the lexical features (the vocabulary used), the visual features (the expressions the speaker makes) and the acoustic features (sound properties like pitch, tone, jitter, etc.).

There are many factors which can affect the output and accuracy of the Emotion detection in speech. A typical set of emotions contains 300 emotional states which are decomposed into six primary emotions like anger, happiness, sadness, surprise, fear, neutral. Success of speech emotion recognition depends on naturalness of database.[11]

Steps for detecting Emotions from Speech data: -

1. Emotional Speech Input
2. Feature Extraction and Selection
3. Classifier
4. Emotional Speech Output

3.1 Databases

Database is most important and necessary part in any machine learning model or research as it's the fundamental on which outcome depends greatly there are many great datasets available for speech emotion dataset. Datasets are also important for training, testing or validation and analysis of techniques or model. We have used four dataset and created one new dataset by combining them so that we could have large amount of data for our model.

The fore dataset used are Crowd-sourced Emotional Mutimodal Actors Dataset (Crema-D), Ryerson Audio-Visual Database of Emotional Speech and Song (Ravdess), Surrey Audio-Visual Expressed Emotion (Savee) and Toronto emotional speech set (Tess).

3.1.1 Surrey Audio-Visual Expressed Emotion (SAVEE)

Database

The SAVEE database has been made by recording of four native English male speakers aged between 27 and 31 years. The dataset contains files in four folders where each folder belongs to one person which are identified as DC, JE, JK, KL. Emotion was created psychologically under six category that is anger, disgust, fear, happiness, sadness, and surprise. The naming of the file is such that the prefix letters tell the emotion which are as ‘a’ for anger, 'd' for 'disgust', 'f' for 'fear’, 'h' for 'happiness’, 'n' for 'neutral’,'sa' for 'sadness' and 'su' for 'surprise'.

3.1.2 Ryerson Audio-Visual Database of Emotional Speech and Song (Ravdess)

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song): with 24 professional actors ,12 male speakers and 12 Female speakers, the lexical features (vocabulary) of the utterances are kept constant by speaking only 2 statements of equal lengths in 8 different emotions by all speakers. We got 7356 audio samples ((Total size: 24.8 GB) which were in the wav format. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. We chose this database for the verity of speech available which is produced by trained professional under the suitable condition. The naming of the file are done in such a way that it identify as first digit tells us about Modality (01 = full-AV, 02 = video-only, 03 = audio-only), second one tells us about Vocal channel (01 = speech, 02 = song), third one tells us about Emotion (01 for neutral, 02 for calm, 03 for happy, 04 for sad, 05 for angry, 06 for fearful, 07 for disgust, 08 for surprised), forth one tells us about Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion, fifth one tells us about Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door"), sixth one tall about Repetition (01 = 1st repetition, 02 = 2nd repetition) and the last one seventh tells us about Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

3.1.3 Crowd-sourced Emotional Mutimodal Actors Dataset (Crema-D)

CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) contain 7442 original audio clips from 91 actors in which 48 were male and 43 were female actors aged between 20 to 74 belonging to various culture, race, places, and ethnicities.

The actors for datasets spoke from a selected sentence of 12 sentences or lines. These 12 sentences were categorized under 6 different emotions i.e. Anger, Disgust, Fear, Happy, Neutral and Sad with 4 different emotion level to the emotion i.e., Low, Medium, High, and Unspecified.

CREMA-D is one of the best datasets as it has large number of speakers which help with many things like the model don't overfit and there is large no of files to train model too.

The naming of the file is done in such a way that first 4 digits tell the actor ID, second 3 alphabet tells which sentence is spoke by actor from the 12 selected sentences, third 3 alphabet tells the emotion the audio file i.e., 'ANG' for Anger, 'DIS' for Disgust, 'FEA' for Fear, 'HAP' for Happy/Joy, 'NEU' for Neutral, 'SAD' for Sad and last 2 alphabet tell about the emotion level of the audio file in the dataset i.e. 'LO' for Low, 'MD' for Medium , 'HI' for High ,'XX' for Unspecified.

3.1.4 Toronto emotional speech set (Tess)

TESS (Toronto emotional speech set) dataset contains 2800 files which are set of 200 targeted words spoken in the phrases of "Say the word ____" by two native speaking English actresses aged 26 and 64 years old and they were also university educated with musical trading. Phrase in dataset was said to stimulate seven emotion that was anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral while keeping the threshold under the normal range.

3.1.5 Combining the four Datasets

For Dataset in this paper, we have Combining the four Datasets into one dataset. If we don't combine these four datasets into one, then there are chances of overfitting the model. The other reason why we have decided to use four combined dataset is so that we have large variety and different circumstances for our dataset and when our model is used on unseen data it will be able to perform better than the model trained on only one dataset.

Emotion	**Gender**	**No. of audio file**
happy	male	827

angry	male	827
sad	male	827
disgust	male	827
fear	male	827
angry	female	1096
disgust	female	1096
sad	female	1096
happy	female	1096
fear	female	1096
neutral	male	837
neutral	female	1056

Table 1. Combined Dataset Audio file count under emotion and gender label.

3.2 Feature Extraction

Speech signals contain many parameters which indicates emotion components of speech. Changes in these parameters Means changing the emotion component of the speech changes. Therefore, appropriate selection of characters or parameter is one of the most important tasks. There are multiple ways of achieving speech emotion recognition by using different features present in speech. We need to extract useful features for our classifier from the dataset.

Then we load the database in python where we will extract the audio features from the file such as power, pitch, and vocal tract configuration with the help of python library like librosa, pyAudio and many more available to us.

Most widely and simplest features used by many researchers: Energy and Energy related features: Energy is one of the most fundamental and necessary features in speech signal that we record when a person. We can extract multiple energy related characteristics in the speech sample that is available to us by calculating the energy, such as mean value, max value, variance, variation range, contour of energy [12]
There are mainly three type of signal domain feature which contain most descriptive features for audio file: -

i. Time domain: - These are easily extracted from waveforms of the raw audio. Zero crossing rate, amplitude envelope, and RMS energy are examples.

ii. Frequency domain: - These focus on the frequency part of the audio. Signals are generally converted from the time domain to the frequency domain using the Fourier Transform. Band energy ratio, spectral centroid, and spectral flux are examples.

iii. Time-frequency representation: - These features combine both the time and frequency components of the audio signal. The time-frequency representation is obtained by applying the Short-Time Fourier Transform (STFT) on the time domain waveform. Spectrogram, mel-spectrogram, and constant-Q transform are examples.[14]

A spectrogram is a chart or pictural portrayal of the range of frequencies of a sound sign as it differs with time. Spectrogram can show both time and frequency part of the sound signs which is gotten by Appling Fast Fourier Change on locally small sections of the sound file.[14]

Conversion from frequency (f) to mel scale (m) is given by

$$m=2595\cdot\log(1+f/500) \qquad (1)$$

A mel-spectrogram is a therefore a spectrogram where the frequencies are converted to the mel scale. [14]

The data of the pace of progress in ghastly groups of a sign is given through its cepstrum. A cepstrum is essentially a range of the log of the range of the time signal. The following range is neither inside the frequency space nor inside the time area and accordingly, it changed into named the quefrency (a re-arranged word of the expression frequency) area. The Mel-Frequency Cepstral Coefficients (MFCCs) aren't anything anyway the coefficients that make up the mel-frequency cepstrum.[14]

The cepstrum passes on the various qualities that develop the formants (a trademark part of the nature of a discourse sound) and tone of a sound. MFCCs accordingly are valuable for profound learning models. [14]

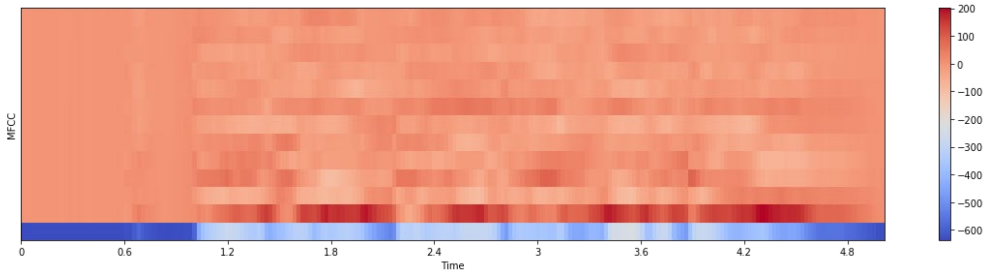


Fig1. Mel-Frequency Cepstrum Coefficients (MFCC) of the Speech signal

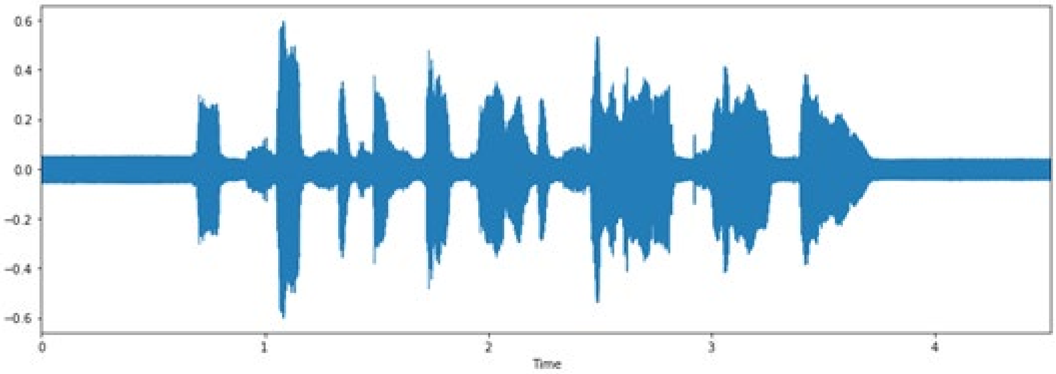


Fig2. Time Domain Plot of the Speech signal

Pitch and the features related to it: In every single speech frame the value of pitch frequency can be calculated. Also, it's statistics is derived in full speech frame. These values display the global properties of characteristic parameters. There is same 19 dimensions of energy in each pitch feature vector. Qualitative features: Of any emotional content, its Utterance is very greatly related to the voice quality. We can even represent it numerically by parameters which are estimated directly from speech signal. The acoustic parameters related to speech quality are: (1) Voice level: There are many measures on which we can rely upon like amplitude, energy, and duration; (2) Voice Pitch; (3) phrase, feature boundary, phoneme, and word; (4) Temporal structures [12]. Linear prediction cepstrum coefficients (LPCC): The channel of the speech can be embodied by it. Everyone who has some varying emotional speech will have different channel characteristics, then to identify the emotions which are present in speech we extract these feature coefficients. The computational method of LPCC is generally a recurrence of computing the LPC. It is in accordance with the all-pole model. Mel-Frequency Cepstrum Coefficients (MFCC): It is like, or we can say that it is based in human ear's hearing. It uses a nonlinear frequency unit, which is simulate the hearing system of us humans. Mel frequency scale: it is a greatly used feature of speech. It has simple calculation, ability of distinction anti-noise and other advantages.[13]

In this paper we have used MFCC (Mel-Frequency Cepstrum Coefficients) as our main feature to classify the speech on the bases of emotion.

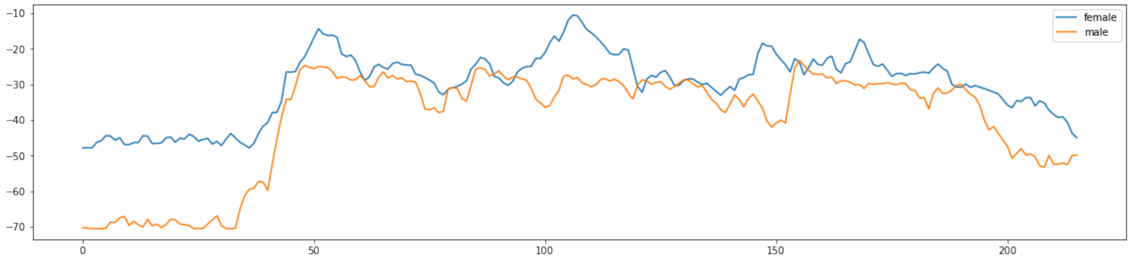


Fig3. Feature Comparison of Gender - Male and Female; Emotion - Angry

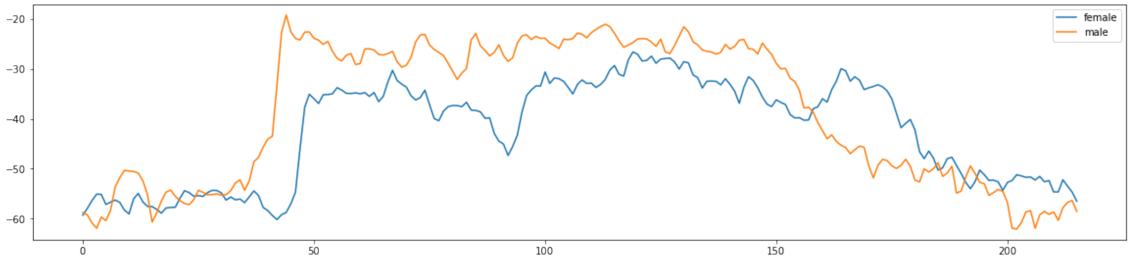


Fig4. Feature Comparison of Gender - Male and Female; Emotion - happy

To maximize the efficiency of space and memory we read each audio file and extract the mean across all the MFCC bands by time and only keeping the extracted feature found in audio file then dropping the entire audio file also replacing the values whose mean is NA to 0. Then we perform normalization on dataset and save the dataset. So that we can split the dataset into the train and test for our 1D CNN model.

I have also explored many ways that can have effect on the outcome of the model such as Data Augmentation on the dataset before using classification or feature extraction method.

3.3 Classification

For this paper we have implemented CNN or ConvNet (Convolutional Neural Networks) which come under the umbrella of artificial neural network and its best for processing on sequential data set and image. CNN model is mostly used for image classification problem, in which input image is converted into two-dimensional input for the reference to feature learning. The major benefit of using CNN is that it can train itself from the sequence of data where CNN do not require domain specific knowledge or manually create input features for CNN. The major drawback of CNN is that the dataset must be fitted to input layers so that it can pass through the input layers and sometime due to fitting the data we can affect the output of CNN model. There are one or more than one convolutional layer, which is followed by a pooling layer and finally a fully connected layer can be repeated as much as the requirement needed or desired to train the model for recognition of pattern over the network in CNN The High layer of CNN use the data generated by the lower layers to recognize complex pattern. We have used keras library of python to implement 1D CNN and 2D CNN.

There 1D, 2D or 3D CNNs which shares the same properties and approach between them and the main difference between them is the input data and the way the filter the input data and input data type which is also commonly called feature detection used on the data.

4. Finding and Performance Analysis

To train and evaluate the model, we started with importing the path of all 4 datasets (Crema-D, Ravdess, Savee and Tess) into the environment and then combined them into one dataset where path of all the audio file, source dataset name and label them under male and female with emotion is saved in into new combined dataset.

For feature selection we used MFCC ((Mel-Frequency Cepstrum Coefficients)) and for calcification 1D CNN is used to classify the MFCC array data into different Emotion category that is gender and emotion or only emotion or only gender. We were able to see that our model works with the help of CNN model loss graph as the train curve is giving a smooth decreasing slope. We have used our both gender dependent and gender independent classification of emotion.

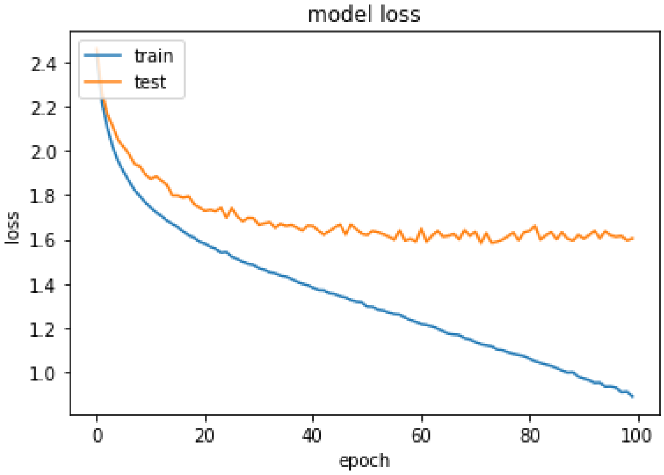


Fig5. 1D CNN Model Loss Graph

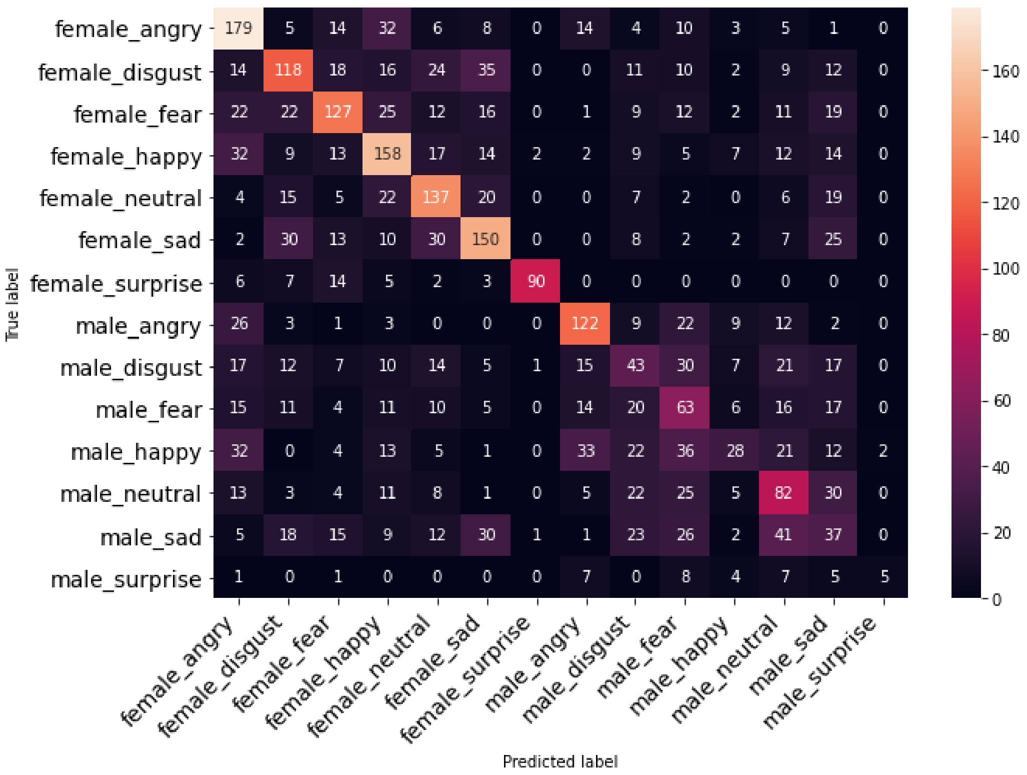


Fig6. Emotion by gender accuracy table

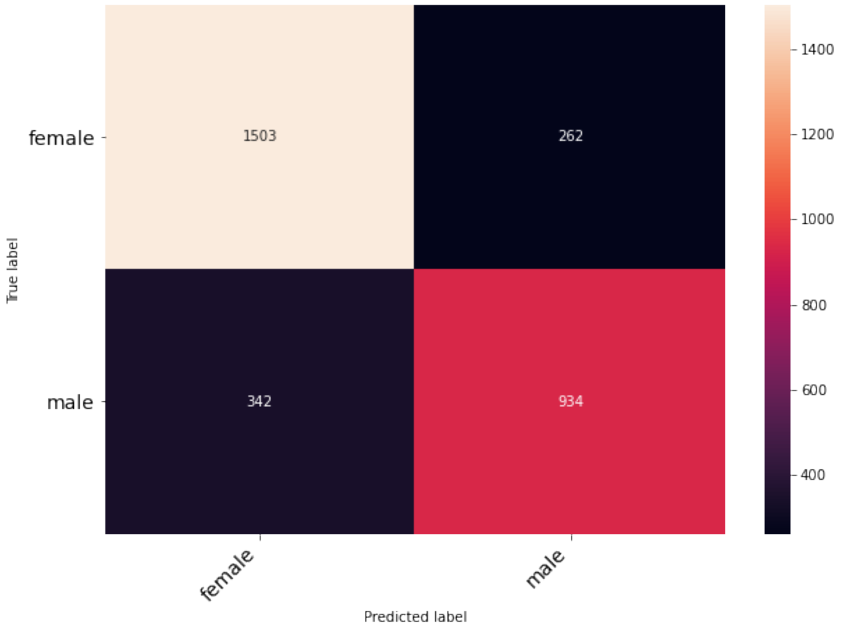


Fig7. Gender Accuracy Result

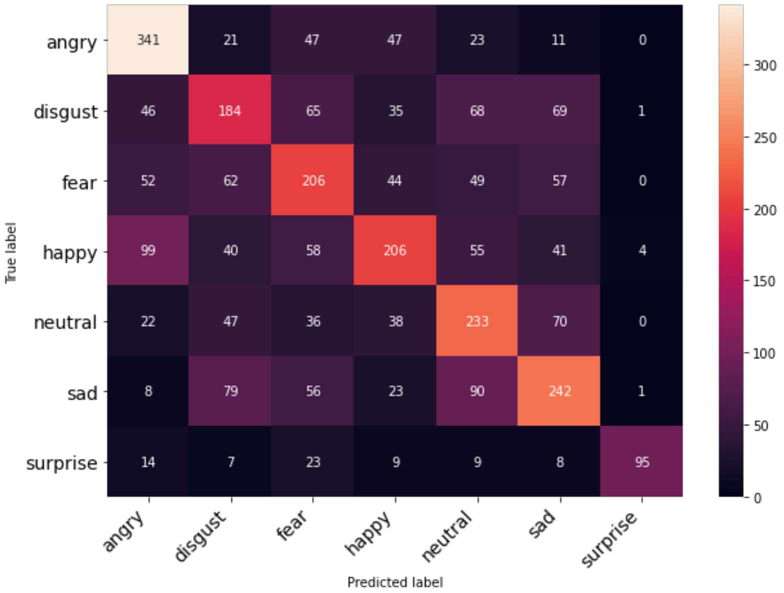


Fig8. Emotion Accuracy

The Gender vs Emotion model accuracy score was 0.44 and the weighted average score is 0.45. The Gender accuracy result of our 1D CNN model is with accuracy score of 0.80 and weighted average accuracy of 0.80. The Emotion Accuracy result of our 1D CNN model with accuracy Score of 0.49 and weighted average accuracy of 0.50.

5. Conclusion

In this paper We were able to implement 1D CNN model to classify the Emotion in speech which was trained on the dataset which is combination for 4 different datasets and for feature extraction where MFCC features are extracted from a speech audio file which are in the format of .wav having the weighted average accuracy of 50%. The primary objective of the paper was to create the base level emotion recognition system with the wide data set to train so that it can perform on new unseen data more effectively, so we trained the model on multiple combined datasets. We will say there is still chances to get better accuracy if we explore new methodology to classify the data or can manipulate the MFCC to give better and clear graph without changing the emotion.

The Emotion Recognition is still had not been widely used in our world for consumer and much new research are continually happening as emotion is affected and based on multiple factors and understanding them with high Accuracy is still hard. Though many company are deriving the use emotion recognition into the new world and we hope it to benefit many people.

For further Enhancement can be done by introducing new and effective emotion recognition model and combining it with facial Expression recognition System to the error can be minimized.

References

[1] Zehra, W., Javed, A.R., Jalil, Z. et al. Cross corpus multi-lingual speech emotion recognition using ensemble learning. Complex Intell. Syst. 7, 1845–1854 (2021). https://doi.org/10.1007/s40747-020-00250-4

[2] Mustaqeem & Kwon, Soonil. (2020). MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. Expert Systems with Applications. 167. 10.1016/j.eswa.2020.114177.

[3] M. Xu, F. Zhang, X. Cui and W. Zhang, "Speech Emotion Recognition with Multiscale Area Attention and Data Augmentation," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6319-6323, doi: 10.1109/ICASSP39728.2021.9414635.

[4] P. Meyer, Z. Xu and T. Fingscheidt, "Improving Convolutional Recurrent Neural Networks for Speech Emotion Recognition," 2021 IEEE Spoken Language Technology Workshop (SLT), 2021, pp. 365-372, doi: 10.1109/SLT48900.2021.9383513.

[5] Jiang L, Tan P, Yang J, Liu X, Wang C. Speech emotion recognition using emotion perception spectral feature. Concurrency Computat Pract Exper.2019; e5427. https:/ /doi.org/10.1002/cpe.5427

[6] Zisad, Sharif & Hossain, Mohammad & Andersson, Karl. (2020). Speech Emotion Recognition in Neurological Disorders Using Convolutional Neural Network. 10.1007/978-3-030-59277-6_26.

[7] M. Li et al., "Contrastive Unsupervised Learning for Speech Emotion Recognition," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6329-6333, doi: 10.1109/ICASSP39728.2021.9413910.

[8] Tris Atmaja, Bagus & Akagi, Masato. (2021). Evaluation of error- and correlation-based loss functions for multitask learning dimensional speech emotion recognition. Journal of Physics: Conference Series. 1896. 012004. 10.1088/1742-6596/1896/1/012004.

[9] Koduru, A., Valiveti, H.B. & Budati, A.K. Feature extraction algorithms to improve the speech emotion recognition rate. Int J Speech Technol 23, 45–55 (2020). https://doi.org/10.1007/s10772-020-09672-4

[10] Singh, Premjeet, Goutam Saha and Md. Sahidullah. “Deep scattering network for speech emotion recognition.” ArXiv abs/2105.04806 (2021): n. pag.

[11] M. E. Ayadi, M. S. Kamel, F. Karray, “Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases,” Pattern Recognition, 2011.

[12] Jia Rong, Gang Li, Yi-Ping Phoebe Chen “Acoustic feature selection for automatic emotion recognition from speech”, Elsevier, Information Processing and Management volume 45 (2009)

[13] P.Shen, Z. Changjun, X. Chen, “Automatic Speech Emotion Recognition Using Support Vector Machine”, International Conference On Electronic And Mechanical Engineering And Information Technology, 2011.

[14] website “https://devopedia.org/audio-feature-extraction”