Bimodal Emotion Classification Using Deep Learning

Ashutosh Kumar Singh

D18128839

A dissertation submitted in partial fulfilment of the requirements of
Technological University Dublin for the degree of
M.Sc. in Computer Science (Data Science)

2020

DECLARATION

I certify that this dissertation which I now submit for examination for the award of MSc in Computer Science (Data Science), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: _Ashutosh Kumar Singh_

Date: September 1, 2020

ABSTRACT

Multimodal Emotion Recognition is an emerging associative field in the area of Human Computer Interaction and Sentiment Analysis. It extracts information from each modality to predict the emotions accurately. In this research, Bimodal Emotion Recognition framework is developed with the decision-level fusion of Audio and Video modality using RAVDES dataset. Designing such frameworks are computationally expensive and require more time to train the network. Thus, a relatively small dataset has been used for the scope of this research. The conducted research is inspired by the use of neural networks for emotion classification from multimodal data. The developed framework further confirmed the fact that merging modality can enhance the accuracy in classifying emotions. Later, decision-level fusion is further explored with changes in the architecture of the Unimodal networks. The research showed that the Bimodal framework formed with the fusion of unimodal networks having wide layer with more nodes outperformed the framework designed with the fusion of narrow unimodal networks having lesser nodes.

Key words: *Multimodal Sentiment Analysis, Bimodal Emotion Recognition, Decision-level fusion, Unimodal Networks, CNN*, MFCC**

***CNN** – Convolutional Neural Network*

***MFCC** – Mel-Frequency Cepstral Coefficient*

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor **Jack O' Neill** for his constant motivation, constructive suggestions, recommendations and tremendous support throughout the dissertation process. It was an honour to work and study under his supervision.

I would also like to extend my thanks to **Dr. Luca Longo**, M.Sc. theses coordinator, for providing valuable inputs in giving direction to the formulation of the proposed research.

Finally, I would like to acknowledge the love and constant support that I got from my family and friends in completing the thesis. Again, special thanks to everyone who believed in me !

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 CHAPTER 1- INTRODUCTION

1.1 Background

Sentiment Analysis has remained one of the hottest topics for all the researchers across the globe. From a long time, the researches in sentiment analysis have majorly revolved around the study of single modality. But now with the advent of platforms like Facebook, YouTube, Twitter etc., people have been expressing their emotions on various topics varying from politics, external affairs, technology, and other related issues. These emotions/sentiments regardless of topic can be in any form of modality, it could be either text, either audio or visuals.

Recent advances in the field of Human-Computer Interaction has led to the scope of sentiment analysis using multiple modalities. In such cases, developing a model using single modality might result in the loss of information and will not be the correct way to justify the emotion. Thus, a combination of different unimodal networks would aid in improving the generalization performance for emotion classification and could also increase the accuracy. Even after multiple researches on the same, effective emotion recognition with the help of deep learning networks remain still challenging. The major scope of the research is to develop an efficient Bimodal deep learning model by extracting the input features from audio and visual modalities which can classify human emotions accurately and further generalize the concept that Bimodal emotion recognition systems have edge over Unimodal emotion recognition system. In doing so, it also aims to develop a simple Bimodal framework with changes in architecture of Unimodal model before fusion which could be well-suited for the smaller data and can be run even utilizing less computing power.

1.2 Research Problem

In the recent years, multiple researches have already been conducted on analyzing and extracting information using all the three modalities - texts, visuals, and audio. However, to develop a powerful emotion recognition system which can accept all

modalities and give excellent accuracy, is still challenging because of the loss of some of the important features while extracting them from each modality. This feature loss cannot correspond to good accuracy and the model cannot be generalized because of this shortcoming. The major drawback of building such system is the computational capacity. Classification using single modality itself takes lot of time but fusing information from each modality and then training the model requires high end GPU powered systems. This research aims to provide some cues regarding the feasibility of implementing such frameworks on normal systems which have less computational capacity. Along with proving the superiority of Bimodal (Audio and Visual) frameworks over Unimodal frameworks, the research will focus on experimenting with the Decision Level fusion of both the individual deep learning models. In order to do so, Early-stage and Late-stage Decision Level fusion will be experimented. In Early-stage Decision Level fusion, each of the Unimodal models will have wide layer made up of large number of nodes just before their individual output layers. While in Late-stage Decision Level fusion, each of the Unimodal models will have narrow layer consisting of lesser number of nodes just before their individual output layers. This would give more insights about decision-level fusion which could be used in developing an efficient final Bimodal system that can classify human emotions

Thus, focus of this work can be formalized by the research question:
"Can early-stage decision-level fusion improve the accuracy of a Bimodal model in emotion-classification problems?"

The proposed experiment will be used to investigate the below two sub-questions derived from the earlier stated Research question.
Sub-question 1: Do Bimodal models give better accuracy than the Unimodal models alone?
Sub-question 2: Should the individual models output a large number of nodes (a "wide" final layer) or a small number of nodes (additional layers leading to a "narrow" node) at the time of decision-level fusion?

1.3 Research Objectives

The aim of this work can be outlined from the hypothesis:

Null Hypothesis

The performance of Bimodal deep learning model in terms of Accuracy, Precision and F1-score, designed to classify human emotions with the decision-level fusion of Audio model built extracting Mel Frequency Cepstral Coefficient features and Video model developed extracting features from the video frames, is affected by the number of nodes in the final layer of each individual models.

Alternate Hypothesis

The performance of Bimodal deep learning model in terms of Accuracy, Precision and F1-score, designed to classify human emotions with the decision-level fusion of Audio model built extracting Mel Frequency Cepstral Coefficient features and Video model developed extracting features from the video frames, is affected by the number of nodes in the final layer of each individual models.

The prime objective of the proposed research is to conduct experiments that could answer the research question and help in developing an efficient Bimodal deep learning framework. The Research objectives corresponding to the two sub-questions derived from the research question are as follows:

Objective 1: Compare the Accuracy of the individual audio and video models with the final Bimodal model. Compare all the final Bimodal models by measuring the accuracy, precision and F1-score.

Objective 2: Make changes in the architecture of best performing audio and video models to analyze the difference in the performance of the Bimodal model.

1.4 Research Methodologies

The research methodology used in order to conduct the experiment is quantitative as the results of the classification from the model will be quantified and evaluated. By type, the proposed research is Primary in nature as it is collected from the online source and pre-processed to develop a Bimodal framework. The research follows

Mixed Methods approach – as is uses a combination of Qualitative and Quantitative methods for data collection(Audio and Video) and its analysis using the metrics like Accuracy and F1 score. This research is Empirical in form as the Accuracy of Bimodal frameworks will be evaluated and compared with the Unimodal frameworks developed using neural networks. By reasoning, it is Deductive since it follows the top down approach from formulating the research question to the evaluation of results achieved from the conducted experiment.

The design of this research is driven by the popular Cross-Industry Standard Process For Data Mining(CRISP-DM) methodology which involves various phases. Chapter 2 covers the first stage i.e., Business Understanding for the proposed research while the other phases are explained in Section 3.3. This section details out all the CRISP-DM phases which were used as part of this research.

1.5 Scope and Limitations

The scope of the proposed research is limited to form a Bimodal deep learning framework which can show improvement in performance over both the Unimodal models. Further, an experiment is conducted to observe any change in its performance after making changes in the number of nodes in the final layer of the unimodal networks, which could provide useful insights how fusion behaves with wide and narrow networks. The results obtained could also be generalized in case of smaller datasets.

The major limitation of the proposed research is lack of access to high-powered computing resources which could help in faster and effective data pre-processing and training of all the models. The high GPU powered systems could also help in the further optimization of those models. Another limitation is the sample size of the data which is relatively small. The model could overfit the training samples and thus, could result in poor performance on the unseen data.

1.6 Document Outline

The steps involved in conducting the proposed research has been explained Chapter-wise. There are total 6 Chapters including Introduction, the other chapters are:

Chapter 2 covers the detailed literature review done in order to conduct the proposed research. It briefs out some of the previous researches which have already been conducted for each of the modality and also, lists out the work in regard to Multimodal emotion classification. This section also covers the Gaps which were identified as part of the research.

Chapter 3 discusses the Design and Methodology which is followed to conduct all the experiments. It explains the approach of the project in detail and majorly, it covers the important stages from data preparation to its pre-processing, from the modeling technique to the evaluation which has been followed for implementing the Bimodal framework.

Chapter 4 covers the detailed implementation of all the models. It explains the technical aspects - from model design to its compilation and execution. It also discusses the Results obtained from each of the implemented models. The comparative analysis based on the accuracy is done in this section.

Chapter 5 builds upon the Chapter 4. It covers the detailed evaluation and analysis of each of the models. The models are compared based on the different performance metrics outlined in Chapter 3. Average Precision and F1 Score with Confusion Matrix was also used to evaluate the overall performance for the developed models.

Chapter 6 concludes the research work with the discussion in terms of findings. It summarizes the experiments and evaluation done for this research. It also discusses the future work and recommendations which could be later put to use in order to do this research.

2 CHAPTER 2 - LITERATURE REVIEW AND RELATED WORK

2.1 Background

Sentiment Analysis started evolving with the introduction of web-based applications like blogs, forums, social networking, e-commerce websites, etc. where people were free to express their emotions positively or negatively about almost everything. The opinions expressed involuntarily by the users can certainly help the stakeholders and the general public in making correct decisions. This study of sentiments behind the presented opinions was called as Sentiment Analysis. It deals with classifying the users' opinions and feedbacks into different categories like happy, sad, angry, etc. The strength of these opinions can also be sub-divided like strongly positive, strongly negative, weakly positive, etc.

The field of extracting emotions from text is a well-researched area and is used by multiple industries across different platforms ranging from business to manufacturing, government sector to politics, entertainment to healthcare etc. However, there is comparatively less research in extracting sentiments from modalities other than text, like Speech and Visual. Today we are observing that the content on social media is drifting towards multiple modalities. Multimodal content is widely popular and is easily accessible on various famous platforms like YouTube, Instagram, Twitter etc. Thus, to generate insights from these diverse modalities is a challenge in the field of emotion classification. Since the aim of this research is to extract emotion from Audio and Visuals, thus primary focus will be on building a Bimodal framework utilizing the information extracted from each of the modalities. The other sections cover the major approaches used by the researchers to analyze emotions from different types of data and progress achieved to date in the field of Bimodal Emotion Recognition System.

2.2 Related Work

This section highlights some of the prominent work done in the field of each modality concerning to our research and also covers the fusion techniques which have been used by most of the researchers in the field of Multimodal Sentiment Analysis.

2.2.1 Emotion Recognition Using Text

Even though texts modality has not been used to conduct the experiments for the proposed research but since it has been widely used in developing a Multimodal deep learning framework which utilizes all the three modalities, some of the research related to this was done to form a basic understanding. The paper by (Tocoglu, Ozturkmenoglu & Alpkocak, 2019) focuses on Turkish tweets from twitter for analysis and classification purposes. For this study, the performance of different neural networks is examined on around 205000 Turkish tweets which are collected raw using Tweepy python library. A lexicon-based approach is used for categorizing the data into different classes. The experiment is conducted by implementing ANN, CNN and RNN. In performance evaluation, CNN outperforms other neural networks with an accuracy of 87%. This work provides a study of high dimensional textual data which is successfully categorized using an optimum CNN model.

In another study (Batbaatar, Li & Ryu, 2019), the experiment is conducted on emotion-annotated datasets which are gathered from 10 different domains like dialogues, fairy tales, tweets, blogs and news headlines. Unlike (Tocoglu, Ozturkmenoglu & Alpkocak, 2019), the pre-processing of textual data is not carried out in the traditional format. The noisy inputs are kept for learning-based approaches and only the numbers, special characters and twitter ids are eliminated from data. The models are trained using two sub-neural networks specifically, making use of the Bi-LSTM neural network for extracting semantic features and a CNN network with word embedding vectors for extracting the emotional relationship. The LSTM model is built using Word2Vec, GloVe and FastText word embedding techniques. It is observed that the FastText technique outperformed other embedding methods. The one associated shortcoming of this study is long training periods.

The experiment (Zhang, Chen, Huang & Cai, 2019) is a text classification study carried out on public feedback posted on government released policies and amendments to correctly classify them and allot them to the respective government authorities. The dataset is comprised of around 4500 Chinese short texts categorized into 22 classes. The high dimensionality is reduced using a word embedding vectorization with a distributed representation method. This is given as an input to the CNN model with 256 different kernel and padding operations. In this study, a differential evolution (DE) algorithm is used as a parameter optimization which comprises four crucial steps for optimization viz. initialization, mutation, crossover and selection. Thus, a CNN-DE model is built for text classification, and the performance evaluation of the model is calculated using precision, recall, accuracy and F1-score.

(Zheng & Zheng, 2019) presents a model to deal with problem of overfitting which is a combination of the LSTM and CNN model and have subsequently achieved better results in text classifications. This study makes use of four datasets including Yahoo responses, Sogou news, Yelp reviews and top 250 short show reviews on Douban movies for sentiment analysis. Standardization and data pre-processing are carried out by converting these Chinese texts and randomly assigning 10% of the total dataset for validation purposes. Deep neural network is designed with the combination of CNNs, and LSTMs with the attention layer for extracting the relevant features. The model gave an overall accuracy of 94%.

2.2.2 Emotion Recognition Using Audio

This section covers the studies specific to speech/audio emotion recognition. The paper (Huang, Wu, Hong, Su & Chen, 2019) proposed a work that uses deep neural networks with SVM for Audio emotion recognition. SVM based model is used for classifying the verbal and non-verbal sounds which are then applied as an input to the Bi-LSTM model with the attention layer. The model is evaluated using accuracy as a performance metric and the results show that the model produced an accuracy of 63% .

(Atmaja & Akaji, 2019) conducted a similar work on IEMOCAP dataset using only audio modality. Here, the speech segments are subjected to silence removal filter using

a threshold value and a minimum number of samples. Audio features are then extracted from this using 13 MFCCs and 13 chroma embedding where, the output is in the form of vectors. In order to preserve the information required for mapping the past and future features; the Bi-LSTM network is applied on the top of the feature extracted layer. Similarly, to deal with the irrelevant extracted features; the attention layer is implemented in combination with the LSTM layer. The system gave an accuracy of 76%, where accuracy is used as the only performance metrics.

In (Zheng & Yang, 2019), Deep Belief Network (DBN) is used to study the audio signals using multiple Restricted Boltzmann Machines (RBM). In this study, 24 MFCC features are extracted from Audio data. These extracted features are given as an input to an RBM network consisting of hidden layers using Bernoulli distribution. The experimented results produced an overall accuracy of 86% on training data while, 78% accuracy on testing data.

Another study (Orjesek, Jarina, Chmulik & Cuba, 2019) focussed on recognizing the musical/song emotions using a deep neural network. The dataset is obtained from MediaEval Emotion from the music dataset consisting of raw audio files. As the study makes use of raw speech segments without any pre-processing, a powerful CNN model is designed. Accuracy and RMSE were used as performance metrics. It is built by stacking a CNN layer on a time-distributed layer and a fully connected layer yielding an accuracy of 48%. This accuracy however appeared to be seemingly less when compared with another benchmark models.

2.2.3 Emotion Recognition Using Videos

This section gives an overview of research majorly done for facial emotion recognition using a video/ image modality. (Saravanan, Perichetla & Gayathri, 2019) proposed a study that makes use of image modality in detecting facial expressions by implementing deep convolutional networks. A dataset introduced in ICML 2013 Challenges in Representation Learning titled as FER-2013 comprising seven basic emotions is used for this study. A basic CNN model is built using two 2D convolutional layers and two max-pooling layers. This simple CNN model produced an accuracy of 65%. Later, six sets of 2D convolutional layer are added which is

followed by two max-pooling layers and two fully connected layers. The model is evaluated by tuning the hyperparameters like the batch size, optimizer and the number epochs. The highest accuracy of about 78% is obtained using an Adam optimizer with a learning rate of 0.0001. The experimented model is tested live using a webcam feed tool, where the results show that the model correctly predicted all the instances of happiness and surprises. As the class distribution of dataset is uneven, with only few instances for the 'disgust' class, the model failed to classify this emotion and in addition the model also performed poorly in categorizing fear and anger emotions.

A similar paper by Abdulsalam et al. (2019) conducts a study in human-computer domain for understanding the human facial expressions from videos. In this experiment, a deep CNN is trained on an Amsterdam Dynamic Facial Expression (ADFES-BIV) dataset has three different levels of intensity with ten emotion labels. Before designing the model, these 1.04s length of videos are preprocessed where video frames are extracted at the rate of 13 frames per second. These frames are then converted to grey-scale level followed by Histogram Equalization (HE) for adjusting the contrast and other lightning conditions for getting a clear image for classification. Similarly, human faces are detected and extracted using a Voila Jones algorithm. To reduce the processing time, the detected faces are then cropped to get an actual face area which is then resized into a uniform size of 70*70 pixels. Finally, a CNN model is implemented on these processed image frames. The experimental results showed an accuracy of 95%, where out of 340 facial records only 21 records were misclassified. Like Saravanan et al. (2019), the video frames with label disgust were misclassified, due to not much availability of the disgust instances in training data.

A study by Lalitha & Thyagharajan (2019) provides an optimal solution in recognition of micro facial expressions based on videos. An optimal convolution network is applied on the dataset which has around 600 video sequences filmed by 123 actors. With the intention of getting a clearer image, the pre-processing of video frames is carried out using an Adaptive Median filter. This filter classifies the noisy pixels by comparing the pixel values with its neighboring pixels and then assigning the median pixel value to the noisy pixels.

This is followed by feature extraction using Geometric Features (GF), Histogram of Oriented Gradient Features (HOF) and Local Binary Pattern (LBP) and a feature selection process using a Modified Lion optimization algorithm. The model is then built on these selected optimal features using a 3D convolutional layer and a max-pooling layer with SoftMax activation function at the output layer. The effectiveness of this model is evaluated using the performance metrics using mean absolute error, precision, recall, F-measures and accuracy. The results of this model yielded an accuracy of 98% which is better when compared with the other existing models. As the feature extraction process is carried out using MATLAB therefore, the system is subjected to more pre-processing time in comparison with other models.

2.2.4 Multimodal Emotion Recognition

This section covers the interesting researche done in the field of Multimodal Emotion Recognition. (Sharma & Mansotra, 2019) explored the area of emotion mapping of students in the classroom. Three different modalities - video, audio and text were utilized for this study. A digital webcam and microphone was used for capturing the live video and audio data whereas, textual data is collected from student's feedback posted on twitter using twitter API. Audio features are extracted using Mel- frequency Cepstral Coefficient (MFCC) methodology while Haar Cascades classifier is used for feature extraction from videos. These extracted features along with the textual features extracted using a TF_IDF approach are then fed to each individual neural network. Transfer learning model VGG19 is used for facial emotion recognition, RNN network with the LSTM layer for speech emotion recognition while SVM machine learning model is used for analyzing text. The model is evaluated using accuracy and precision as a performance metrics. The experimental results showed the overall accuracy to be of 76%, with an individual model accuracy as 81%, 77% and 83% for facial, text and speech, respectively. The system was trained on Facial Emotion Research (FER) 2013 dataset and tested on the live student's database, yielding an optimal accuracy of 81%.

Another unique approach is explored in the paper (Tripathi & Beigi, 2019) which makes use of deep learning techniques applied to the IEMOCAP which is one of the benchmark datasets for Multimodal Sentiment Analysis. The dataset consists of about 12 hours of video-audio data with transcript texts and having 10 emotion labels

recorded by 10 actors. Feature extraction of audio data is extracted using a Fourier Transformation for a total of 34 features. While texts are subjected to pre-trained Glove word embedding vectorization, video features are extracted using Voila Jones algorithm. Deep learning models are then applied to these extracted features for further model building. A Bi-LSTM model with an attention layer is implemented for training the audio features, CNN network for textual data whereas, and Conv2D network with SoftMax activation function built for the visual data. The three models are concatenated using a late fusion methodology which produces an overall accuracy of 87%. This study is associated with the performance limitation, as another study mentioned in this paper yields an accuracy of 95% on the same data.

Verma et al. (2019) addressed a system that combines both common and unique latent information from a Multimodal dataset for detecting human sentiments. This work is applied to CMU-MOSI and POM datasets consisting of tri modalities. This is followed by extracting the combined features of each modality with an intention to capture the latent embeddings. The combined features are then concatenated using a fusion layer to build a final model for emotion recognition. The performance metrics used for model evaluations are MAE, Pearson's Correlations and F1 score. As the DeepCU model is designed using a combination of each modality, the model becomes more complex while fusing these features.

Another study shows an attempt where an improved version of Multimodal sentiment analysis is achieved using a multi-attention LSTM model. In the paper (Xi, Wang, & Yan, 2020), the system architecture is designed using a unimodal feature extraction model followed by the attention layer and Bi-LSTM network for emotion classification using tri-modal data. Firstly, feature extraction is carried out for each modality where, word2vec embedding vector is used for textual data, an open-source OpenSmile software with a frame rate of 30hz and 100ms of window size is used for extracting audio features and 3D CNN network is used for feature extraction of visual data. A network designed with two attention layers for eliminating the redundant features an gave an overall accuracy of 91%.

Gallo et al. (2019) carried out the multi-modality fusion including multi-feature fusion to improve the accuracy of audio-text sentiment analysis on another benchmark dataset

CMU-MOSI. While Nemati et al. (2019) presented a multi-modal approach that incorporates images and text explanations in real-world situations to enhance the efficiency of multi-modal classification. Another research by Chen et al. (2019) introduced a hybrid Multimodal data fusion method where audio and visual modalities are merged earlier and later their features from cross-modal space are combined with textual modality. Likewise, Cornejo and Pedrini (2019) developed a hybrid deep CNN network to extract audio and visual features. The high dimensional features were reduced to low dimensions using Principal Component Analysis (PCA) before performing fusion. The model performed well giving a decent accuracy with the inclusion of PCA feature selection technique. Since the scope of this research is limited to fusion of Audio and Video modality, several techniques relevant to develop Bimodal Emotion Recognition System were employed from the existing research.

2.3 State-of-the-Art

This section discusses the state-of-the-art in the field of Bimodal Sentiment Analysis covering the effective techniques which have been used for extracting information from each modality and further, fusing that information to predict the emotion the combined modalities convey.

2.3.1 Audio Feature Extraction

The below image shows the visual representation of three different forms of visual representations of a sound wave. The first one represents the Audio wave in time domain, which compares change in amplitude over time. The other two represent the frequency domain features – Spectrogram and Mel Frequency Cepstral Coefficients (MFCCs) respectively. Recently, MFCCs have become one of the key features when it comes to Speech Recognition. The use of Mel Frequency Cepstral Coefficients (MFCCs) is considered as one of the standard methods for feature extraction in designing a Multimodal framework. For the scope of this research as well, MFCC features will be extracted for classifying the Audio wave efficiently.

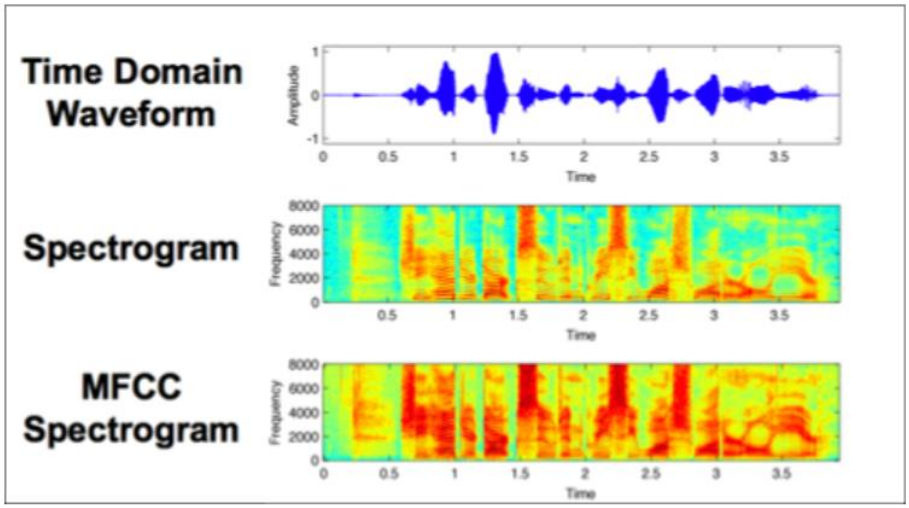


Figure 2.1: Visual Representation Of Audio Wave[1]

2.3.2 Visual Feature Extraction

Extracting features from Videos is related to the domain of Image processing where multiple frames (or images) are extracted from each of the clip and based on the scope of the problem, features are extracted from it. Visual features can be broadly divided into two: Geometric based and Pixel based. Most of the common Geometric based features are Hough Transform, Blob Extraction, Edge Detection etc. Edge Detection is one of the key tool for image processing and computer vision problems (Asghari, Jalali, 2015). It helps in identifying the points in the image where there is discontinuity in brightness, depth, or surface orientation. This research has been conducted on Pixel based feature extraction from Images.

The coloured images are made up of multiple colours which are primarily generated by three colours: Red, Green and Blue. For these three colours, there are 3 matrices or channels. Each of the matrices correspond to particular colour and has values ranging between 0-255 representing the intensity or brightness of the pixel.

[1] https://medium.com/@mikesmales/sound-classification-using-deep-learning-8bc2aa1990b7

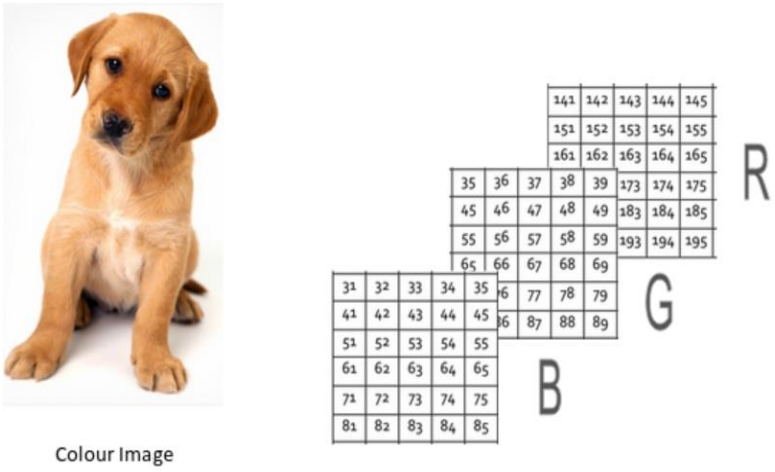


Figure 2.2: Coloured Image in RGB channels[2]

2.3.3 Neural Networks

Neural networks give more edge over other techniques in identifying the hidden patterns from the data. Most of the effective Multimodal Emotion recognition system are developed using only neural networks.

Dense Networks

Dense Neural Networks are the simplest form of Neural networks which are composed of fully connected Dense layers. Each neuron within a layer is connected with each neuron in the next layer. The structure of Dense networks is inspired by the functioning of human brain. When it comes to Linear classification, densely connected layer offers learning functions from all the combinations of the previous layer's features. The figure 2.3 shows the basic representation of Dense Neural Networks consisting of one input layer, 3 hidden layers and final output layer.

[2] https://www.analyticsvidhya.com/blog/2019/08/3-techniques-extract-features-from-image-data-machine-learning-python/

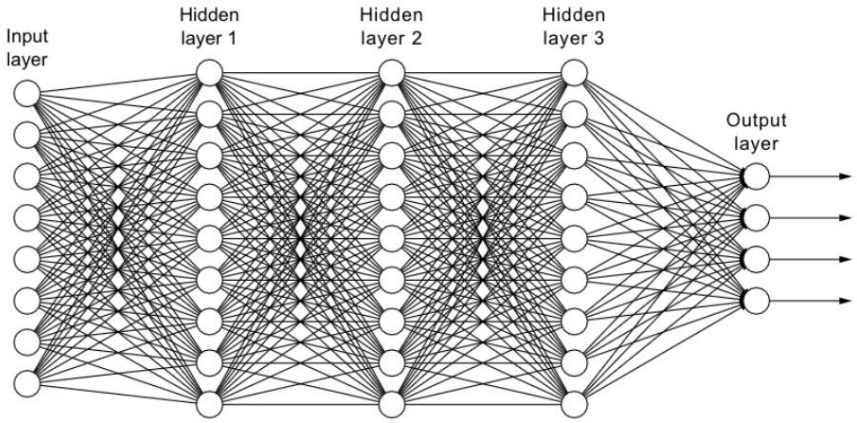


Figure 2.3: Dense Neural Networks[3]

Convolutional Neural Network

Convolutional Neural Networks are often referred as ConvNet or CNN and are widely used in the field of Image processing. Conv2D can take the matrix with pixel values from the colored image as input and convolve through it with the help of filters to extract output features. (Waseem, Davidson, Warmsley, & Weber, 2017) states that the role of the ConvNet is to reduce the images into a form which is easier to process, without losing features which are critical for getting a good prediction. Not only in the field of Image processing, it has been widely used for Audio and text classification as well.

(Lu, Zhang, & Nayak, 2020) recently proposed a new attention-based neural network architecture called Classifier-Attention-Based Convolutional Neural Network (CABCNN). The algorithm used Conv1D for Audio inputs followed by the pooling layers along with the attention mechanism. This design seems to significantly reduce the complexity by reducing the number of features required to train the network.

Based on the reviewed researches, it is clear that CNNs are better when it comes to image classification problem when the training samples are huge. But not only in the field of image, there has been significant contribution of CNNs in the field of Audio classification as well.

[3] https://freecontent.manning.com/neural-network-architectures/

2.3.4 Fusion Methods

The fusion of different modalities is generally performed at two levels: *feature level* and *decision level*. Fusion at the feature level is done before the modeling process by integrating or combining features from all modalities; therefore, it is referred to as early integration (Nefian et al., 2002). The models in such fusion are trained using single input vector which is the shared representation of Audio and Video features. On the other hand, at the decision-level, modeling of each modality is performed separately and then the outputs or decisions of the models are integrated to produce the final decision (Snoek, Worring, & Smeulders, 2005). Thus, it is often referred as Late fusion/integration.

More recently, some of the researches have used Hybrid fusion which is the combination of feature and decision level fusion. In the study by (Nemati et al., 2019), a hybrid Multimodal data fusion method is proposed for multimodal emotion recognition which benefits from both feature and decision-level fusion. Fusion can be characterized in terms of modeling, whether the model is trained using single input vectors for each modality or using separate inputs for each modality. The below figure illustrates the modeling process for Feature and Decision Level fusion.

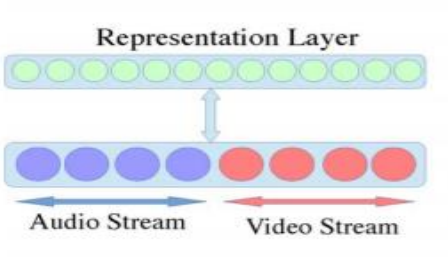


Figure 2.4: Modeling with Feature Level Fusion

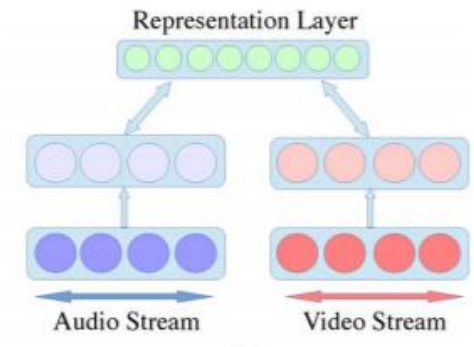


Figure 2.5: Modeling with Decision Level Fusion

For the scope of this research, decision-level fusion is experimented further at two stages and have been termed as 'Early stage' and 'Late Stage'. Early stage will be when the network has a wide layer with increased number of nodes just before the final output layer in each of the model. While for Late stage decision level fusion each of the model will have more layers and thus, less nodes in the final layer of each model.

2.4 Gaps In Research

The various researches in the analysis of human emotions and literature review by (Kaur, Kautish,2019), (Cambria et al.,2017), (Poria et al.,2018), (Soleymania et al., 2017) in the field of Multimodal Sentiment analysis were examined. The major gaps which were observed in the research were either related to loss of important features or the model was so complex that it sometimes suffered performance limitation in the form of underfitting.

Also, all the researches have been done on huge benchmark datasets which would require lot of time to train the model even with the increased computational capacity. Like, the model proposed by (Batbaatar, Li, Ryu, 2019) had a shortcoming of long training periods while (Zhang, Chen, Huang & Cai, 2019) proposed a model which suffered from underfitting while categorizing multiclass data with multiple contextual possibilities. Thus, developing a Bimodal framework which can take less time to train and are not computationally expensive has not been addressed in most of the researches.

Similarly, it was observed in the researches by (Saravanan, Perichetla & Gayathri, 2019), (Abdulsalam, Alhamdani & Abdullah, 2019) , (Roopa S.,2019) used CNNs, transfer learning approach with feature extraction techniques etc. on different datasets for understanding human facial expressions. (Verma, Wang, Zhu & Liu, 2019),(Hammad, Liu & Wang, 2019) also suffered loss of important features giving them relatively lower accuracy. Model developed by (Tripathi & Beigi, 2019) also suffered performance limitation because of the complexity of the model.

Thus, this research aims to address these gaps and develop an efficient Bimodal framework without using complex structures which is trained on smaller dataset and is not computationally expensive. To best of my knowledge, the chosen dataset RAVDES has not been used yet for Multimodal Sentiment Analysis. The research work performed on this dataset will further explore the fact that the Bimodal framework can outclass the individual single modal networks in classifying human emotions.

3 CHAPTER 3 - DESIGN AND METHODOLOGY

This chapter covers the approach which has been followed to test the proposed hypotheses. Also, it covers the steps from Data preprocessing to implementing models for the same which would be evaluated later in the following chapter.

3.1 Project Approach

Extracting sentiments from the multimodal data is relatively new and has become quite a challenge for the researchers. This research aims to develop an effective Bimodal framework using Audio and Video modality to classify human emotions. In order to accomplish the proposed research, the experiments will be conducted in two phases. For the first experiment, the audio and video files will be processed independently, and their corresponding models will be developed. Later, the information extracted from each modality will be fused together at the decision level to build a final model in order to classify emotions. Different deep learning architectures based on the limited availability of GPU, will be experimented and tested for accuracy. This will basically conclude Experiment 1 and help in answering the *Sub-question 1* mentioned in Chapter 1.

Once the best performing model is found, it will be further examined to test *Sub-question 2* which is the basis of Experiment 2:

- Should the individual models output a large number of nodes (a “wide” final layer) or a small number of nodes (additional layers leading to a “narrow” node) ?

Since the scope of the research is limited in terms of availability of the computational power and limited GPU, relatively smaller data was used to design the experimental framework. The performance difference in terms of accuracy, precision and F1 score of each developed model will be further compared and evaluated to answer the research question.

3.2 Design Aspects

The research includes extraction of audio and video features from the data, training the model and evaluating further on test data. Since designing the Multimodal classifier could be a costly affair in terms of computational power, the proposed experiments were conducted using the Google Colaboratory, which is an online cloud based Jupyter Notebook environment and offers a Tesla K80 GPU with 12 GB of RAM that can continuously run for 12 hours. Storage of Google drive was expanded from 12 GB to 100 GB to store the data online for feature extraction and further building models.

3.3 Detailed Design and Methodology

The proposed research follows the popular CRISP-DM methodology which is quite popular across businesses in implementing end-to-end Machine learning models. CRISP-DM stands for Cross Industry Standard Process for Data Mining, which basically segregates the whole process into 6 stages as shown in the Figure 3.1.

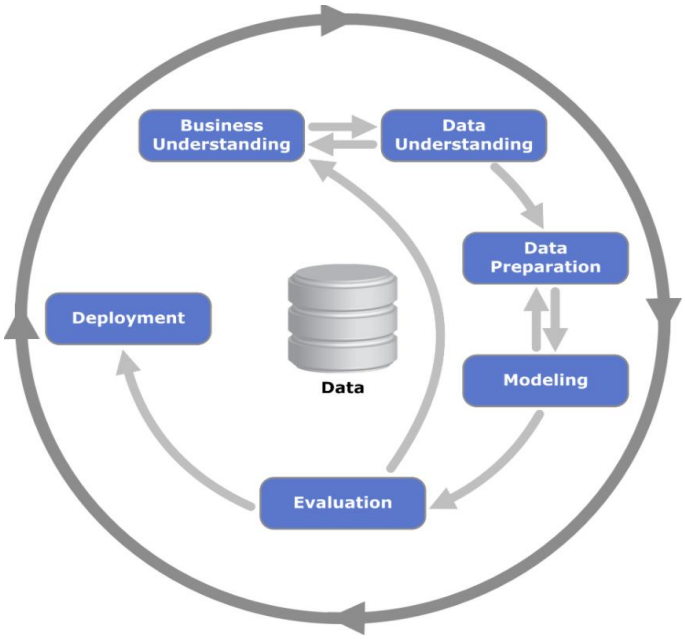


Figure 3. 1: CRISP-DM Cycle[4]

[4] https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome

The model implementation using CRISP-DM goes through six different stages in which some of the stages are inter-connected with each other showing dependencies on one another.

Business Understanding covers the key objectives and approaches which are required to implement the proposed project from a business perspective. This stage basically involves preparing the draft plan for the project implementation which could be reviewed later from the business point of view. Chapter 1 and 2 cover this phase of the project which discuss the major scope, limitation, and objectives of the research.

Data Understanding involves accumulating data from various sources, reviewing data quality or generating basic first insights from the same. This helps in uncovering the pre-requisites which would be essentially needed for the research. Section 3.4 discusses this phase of methodology from the perspective of the proposed research.

Data Preparation, as the name suggests involves construction of data in the format which is understandable. It includes all the activities from collecting raw data from numerous sources to make it suitable for use as inputs for the model implementation. Section 3.5 covers all the steps from feature extraction to pre-processing making it compatible to feed into proposed models.

Modeling stage involves the technical implementation of the analytical models. From choosing suitable modeling techniques to setting the hyperparameters for optimization, all are included in this phase of the CRISP-DM. Section 3.6 gives an overview of the modeling techniques used for the scope of this research and Chapter 4 discusses the brief implementation of models with all the essentials required for designing neural networks.

Evaluation covers evaluating the models designed in the earlier stage with respect to the new and unseen data to check its robustness and performance, which could help in making the generalizable claims. This stage based on the various evaluation metrics, helps in finding out the best performing model out of all the developed models for the research. Section 3.7 provides an overview of the evaluation metrics which will be used to test the developed models while Chapter 5 presents the detailed evaluation and analysis of all the experiments.

Deployment deals with deploying the developed framework into applications as per the business requirements. This stage is out of scope for the proposed research and thus, has been excluded from the design methodology.

3.4 Data Understanding

To design any framework, it becomes extremely crucial to familiarise yourself with the data on hand. This section discusses the dataset and its characteristics and how they were used to develop a Bimodal Classifier.

The raw data in the form of Audio and Video files was downloaded from Zenodo[5] which provides an open access to RAVDESS(Ryerson Audio-Visual Database of Emotional Speech and Song) dataset. The database contains 24 professional actors(12 female, 12 male) speaking the same sentences in a North American accent. The dataset contains two set: Speech and Song for each modality (Video only & Audio only). Speech includes neutral, calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains neutral, calm, happy, sad, angry, and fearful emotions. The song files also contain the same sentences except for the change in pitch. In order to conduct this experiment, both Speech and Song files were considered for the Audio in order to increase the number of samples. While only Speech files were used for feature extraction from Videos. Below are the further details of each filetype:

- *Video Only files*

Speech files collectively contains 2880 files: 60 trials per actor x 2 modalities (AV, VO) x 24 actors = 2880. Out of 2880, 1440 files were used for feature extraction as they only represent Video Only(VO) files.

- *Audio Only files*

Speech file contains 1440 files: 60 trials per actor x 24 actors = 1440 while Song file contains 1012 files: 44 trials per actor x 23 actors = 1012.

Each file has a unique name and consists of 7-part numerical identifier(e.g. *02-01-06-01-02-01-12.mp4)* where the 3rd part represents the emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).

[5] https://zenodo.org/record/1188976

3.5 Data Preparation

Since raw data for both the modalities was considered for developing a multimodal emotion classifier, the feature extraction from Audio and Video becomes the initial step to prepare a data representing the features with the labels which would essentially influence the classification. The two sub-sections below will cover how the useful features were extracted from Audio and Video files.

3.5.1 Audio Feature Extraction

The audio features can be broadly categorized into Time Domain and Frequency Domain Features. Time domain features are essentially represented in the form of Audio waves while Frequency domain features are generated by converting the time-based signal into the frequency domain. The figure 3.2 shows the representation of Audio signals. Frequency domain features are considered to be better for classification problems as they can provide extra information like pitch, rhythm, melody etc.

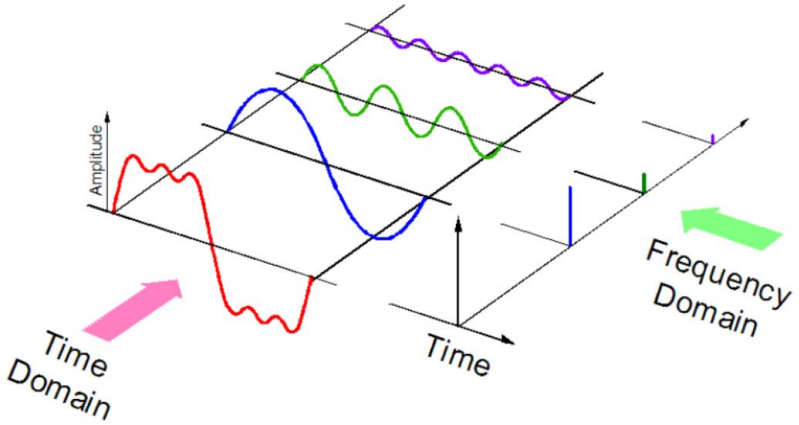


Figure 3.2: Audio Signals in Time and Frequency Domain[6]

Mel Frequency Cepstral Coefficients(MFCCs), Log-Mel-Spectrogram, Chroma etc. are amongst the most commonly used frequency domain features. MFCCs are generally a good representation of the vocal tract(including tongue and teeth) which helps in generating the sound. The use of about 20 MFCC coefficients is common in Automatic Speech Recognition(ASR), although 10-12 coefficients are often considered

[6] https://www.analyticsvidhya.com/blog/2017/08/audio-voice-processing-deep-learning/

to be sufficient for coding speech (Hagen et al., 2004). For the proposed work, python's *librosa* library was used for MFCC extraction. Each audio file was sampled using *Kaiser_fast* and 40 MFCC features were extracted and stored in the form of an array with the labels for each emotion. *Kaiser_fast* is known to reduce the load time when the default sample rate is used and thus it was given preference over '*scipy*'.

3.5.2 Video Feature Extraction

Videos cannot be directly fed to models as Inputs, they essentially need to be in the pixel formats. Thus, the first step to prepare data from the raw videos was to extract frames from each Video. Once the frames are extracted it becomes a typical image classification problem. This was achieved using Python's OpenCV library which provides a lot of flexibility to play around with the extracted frames. In order to conduct this research, two frames were extracted after every half second of the clip and then were resized and concatenated horizontally to form a collage. The intuition behind joining the frames horizontally was to increase the redundancy which could help in training the network better. This gave around 6-7 frames for each clip depending on its duration. Each of the extracted frames are of high resolution and has dimension of 256*512. This is how each frame looked like:

Figure 3.3: Image for Neutral Emotion

Digital images are made of pixels, where each pixel is a discrete value. For colored images, each pixel is represented by 3 channels – Red, Green and Blue where each has a value between 0 and 255. The color of an individual pixel in this case is calculated by combining the RGB values. Thus, here each frame of 256*512 dimensions is represented by three matrices of 256*512 pixels. Using Python's OpenCV library, all

the features and labels were extracted and later stored in the form of array for further training the networks.

3.5.3 Data Pre-processing

Although the features were extracted individually from each audio file and video file, yet in order to merge the two different models the input dimensions should be similar. For each video, there are 6-7 frames corresponding to 1 emotion while there is 1 audio clip for the same emotion. As the number of frames or images were higher in number for each of the Video sample than the Audio clip, it was important to balance the number of samples in both types of data. To ensure this, Video data was sliced to 2452 samples in order to match with the Audio samples(Speech + Song = 2452 samples). Another thing to keep in mind while training the merged model is to ensure one to one mapping between the Audio and Video data, the labels need to be in sync. It was done class wise, so that for each class there are same number of audio recordings and images. The below figure 3.4 represents the number of samples for each of the labels in both type of data after all the preprocessing was done.

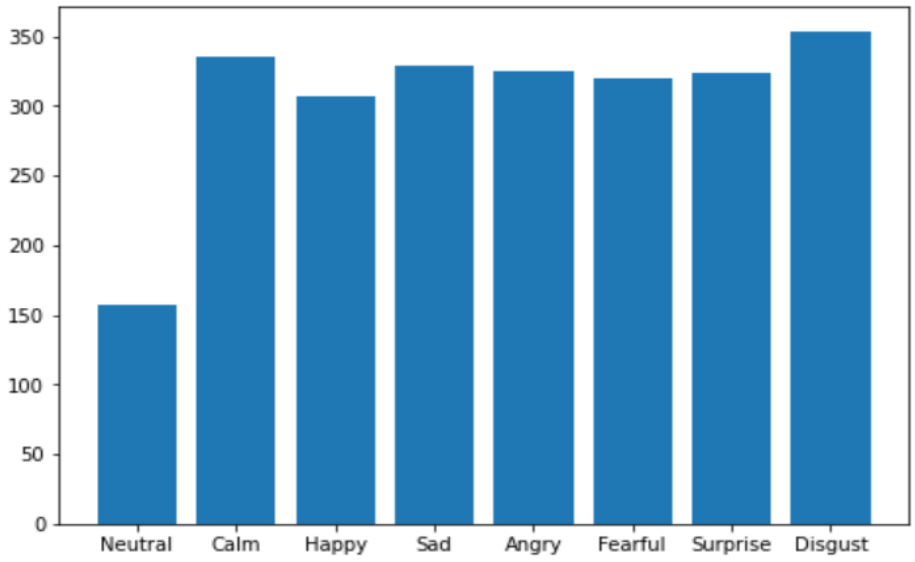


Figure 3.4: Distribution of Classes in Target variable

3.5.4 Data Splitting

After pre-processing Audio and Video data, each set was split into Training and Test set with 80% and 20% split, respectively. Later, the training set was further divided into Training and Validation set with 75% and 25% split in order to fine tune and avoid overfitting or underfitting of the model by monitoring the loss and accuracy in each epoch. So basically, whole data has the 60%-20%-20% Training-Validation-Test split. The same split has been used throughout for implementing all the models.

3.6 Modeling

Researchers often explore multiple models to test the proposed hypothesis and ensure all the processes which need to be followed to enhance the efficiency of the framework. This section follows the same approach and discusses the modeling technique which will be used to find the answer for the research questions mentioned earlier. The figure 3.5 below shows the framework designed for developing a Bimodal emotion classifier for the proposed research.

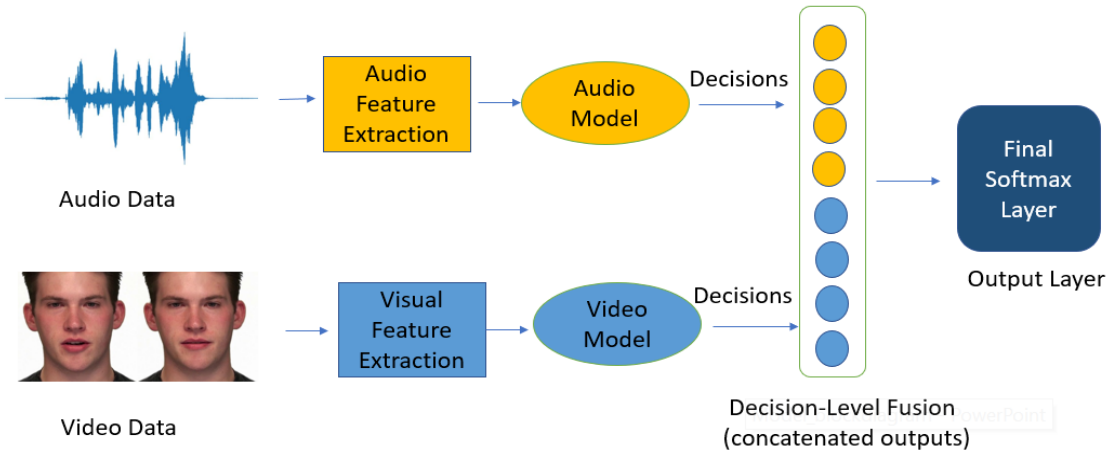


Figure 3.5: Bimodal Emotion Recognition Framework

As discussed in the earlier sections, the research has been divided into two different categories one building upon the other. For Experiment 1, the model building approach has been to try out different combinations of simple Dense and CNN networks for

individual audio and video models. The more complex models while concatenating at decision level, could cause issues in resource(GPU) allocation, thus architecture for each of the model was kept simple.

Dense Only networks consist of only dense layers made up of n number of nodes. As stated in Keras documentation, Keras dense layer provides the flexibility to use 2D(or higher rank data) as inputs. If the input to the layer has a rank greater than 2, then it is flattened prior to the initial dot product with the kernel. Using the same functionality, Audio and Video inputs were passed to the Dense networks. Similarly, Conv1D and Conv2D neural networks have also been used to form the Audio and Video models, respectively. The input shapes were changed for each of the model accordingly. For Audio CNN, input dimensions which are in the form of numpy arrays were expanded along the axis using np.expand(), while inputs for Video CNN were directly fed to Conv2D which accepts four-dimensional data in the form of (Batch Size, Height of Image, Width of Image, Number of channels). Here, number of channels is 3 because of the colored images which are used as inputs. The figure 3.6 represents the conventional convolution layering pattern for colored image classification. The Modeling in terms of implementation of each of the model is covered in detail in Chapter 4.

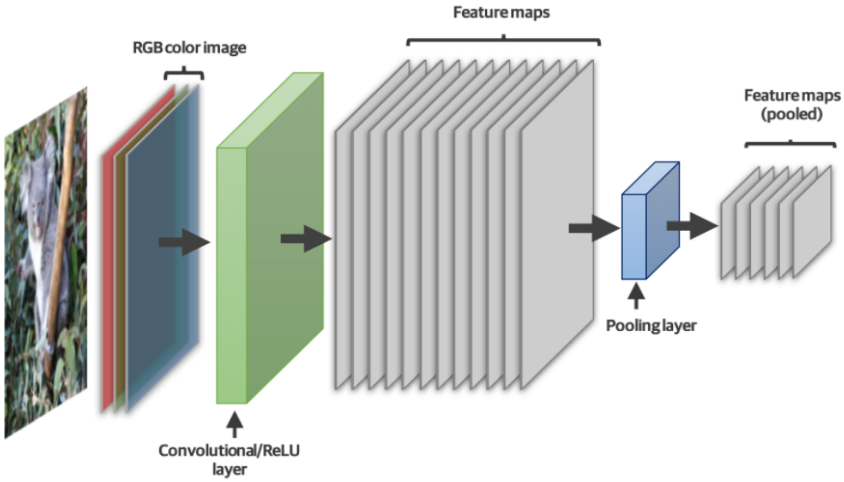


Figure 3.6: Convolution layering pattern[7]

[7] https://www.oreilly.com/library/view/strengthening-deep-neural/9781492044949/ch04.html

In order to learn the complex patterns from the data, it is essential that the neural network has an activation function. Without activation functions, each neuron in the network will only perform a linear transformation on the passed inputs. The Activation functions introduce non-linearity in the networks and make it more powerful to generate insights from the complex data. The most common Activation functions used in the neural networks are shown in Figure 3.7. For this research in each of the dense and convolution layer of the models, ReLU(Rectified Linear Unit) activation function has been used which deactivates the neuron for the negative input values. It is considered to be computationally efficient since it activates only certain number of neurons.

The other activation function Softmax is used in the final output layers of each of the model to classify multiclass target variable. While Sigmoid is often used for Binary classification problems, Softmax is used if there are more than 2 classes. It consists of same number of neurons as the number of classes which need to be classified i.e., 8 in our experiment. Thus, each model's output layer is followed by the Softmax activation function consisting of 8 nodes which predicts the probabilities of the data point belonging to the same class.

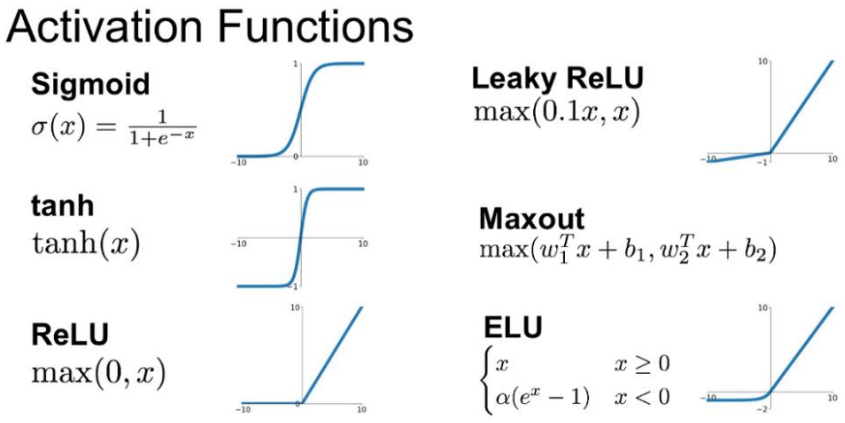


Figure 3.7: Activation functions[8]

[8] https://medium.com/@shrutijadon10104776/survey-on-activation-functions-for-deep-learning-9689331ba092

For the scope of this research, decision level fusion , sometimes also referred as Late fusion, has been used to merge the information from each of the modality. In order to perform a decision level fusion, the final softmax output layer which classifies emotion is removed from each of the unimodal models. And the information extracted from each of the modality just before that layer, is fused to form a concatenated vector which contains information from both the modalities. All the contextual features extracted in each of the Unimodal model before the final output layer are converted into single dimension beforehand using the flatten layer. This eases the process of concatenating the multi-dimensional output from each of the modality in the later stages. The figure 3.8 below depicts the basic working of flatten layer of keras.

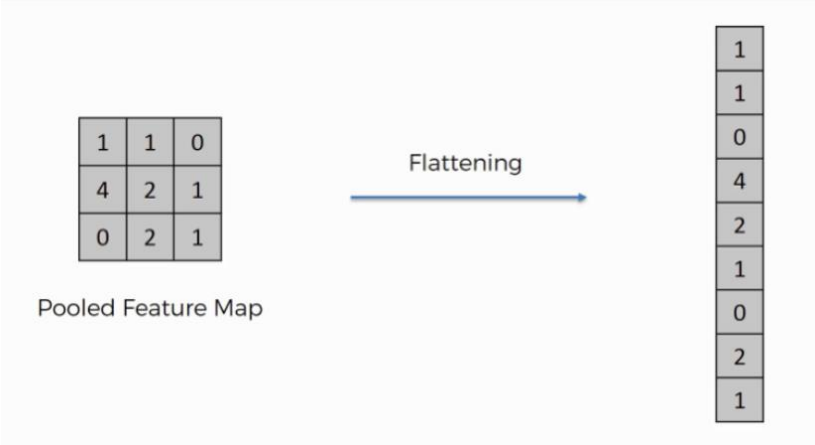


Figure 3.8: Keras Flatten Layer[9]

Once the various Bimodal frameworks are designed by the fusion of Audio and Visual models, the next step involves evaluating those models and finding out the best model for Experiment 2. The best Bimodal model is further investigated by changing the architecture of individual models with additional layers and the wide final layer to investigate any noticeable change in its performance after fusion. The reason behind conducting this experiment is to examine how the number of nodes in the final layer of each individual model can impact the decision-level fusion. As part of this experiment, two Bimodal models are implemented. For one of the Bimodal models, just before the flatten layer in both the Unimodal models, a wide layer with 128 neurons is kept in order to create a Wide layered network which would result in more output features.

[9] https://www.superdatascience.com/blogs/convolutional-neural-networks-cnn-step-3-flattening

The number of neurons have been kept 128 in order to avoid GPU allocation(out of memory) issues.

Contrary to this, another Bimodal model is built with the individual networks (unimodal models) which have additional layers and 8 neurons in their final layer just before the flattening. It would result in lesser number of output features from both the modality channels. Both the models after training will give more insights about the impact of change in the architecture of Unimodal networks.

3.7 Performance Evaluation

There are various evaluation metrics like Accuracy, Precision, Recall, F1 score etc. which could be used for judging the performance of the developed models. In some of the cases, accuracy of the model can be enough to test the performance of the model. However, taking other metrics into consideration could give further insights regarding the robustness of the model.

Classification accuracy is defined as the number of instances in the dataset which are classified correctly from the overall predictions made by the classifier. The calculation for classification accuracy is done using formula: (TP+TN)/(TP+TN+FP+FN) where TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives.

Confusion matrix compares the predicted target values generated by our model with the target values of our test data. True positives (TP) are the number of instances which were classified as positive by the classifier that are actually positive while True Negatives (TN) are the number of instances classified as negative by the classifier that are actually negative. On the other hand, False positives (FP) are the number of instances classified as positive by the classifier that are actually negative and False negatives (FN) are the number of instances classified as negative by the classifier that are actually positive.

Precision is often referred to as the Positive Predictive Value. It gives the fraction of positive instances classified correctly by the classifier. It is calculated using formula : (TP)/(TP+FP).On the other hand, *Recall* is also called the True Positive Rate or Sensitivity as it measures the proportion of actual positives that are correctly identified as such. It can be calculated using formula: (TP)/(TP+FN).

F1 score is also known as F Score or F Measure and uses precision and recall for computing the score. It depicts the balance between precision and recall. In simple terms if precision increases, recall decreases, and recall increases if precision decreases. It can be calculated using the formula: 2*(Precision*Recall)/(Precision+Recall).

For the scope of this research, the models will be evaluated on the Accuracy and the weighted average of Precision, Recall and F1 score because of the multiple classes in the target variable. The performance evaluation for all the models is covered in detail in Chapter 5.

4 CHAPTER 4 – IMPLEMENTATION AND RESULTS

The previous chapter covered the Research Objectives, Data Preprocessing and the Modeling and evaluation metrics which would be used for the scope of the proposed research. This chapter covers the range of experiments which have been performed in accordance with the research needs. It explains the technical implementation of individual Audio and Video classifiers as well as the Bimodal Model developed with the decision level fusion of those classifiers. The performance of each individual classifier is compared with the merged model in the Results section for both the experiments mentioned in the Modeling sub-section of the previous chapter.

4.1 Experiment 1

To serve the purpose of this experiment , different comparative models were developed based on the performance of the individual classifiers which resulted in the creation of 3 final Bimodal models. The technical details for the same are discussed below.

4.1.1 Model 1 : Dense Only Networks

The first model is designed by the concatenation of Dense Only networks built for each Audio and Video Inputs. Dense Only networks were chosen just to reduce the complexity and computational cost of the merged model. The audio model with the input shape of (Batch Size, 40,1) is directly fed to the Dense layer consisting of 64 nodes. Batch Size of 32 has been kept constant throughout the experiments. The activation function ‘Relu’ was assigned to the Dense layer. Flatten layer is added before the final layer to convert the data into single dimension keeping the concatenation in mind needed in the later stages for the final Bimodal model. The final dense layer computes the Softmax probabilities for each of the 8 categories corresponding to 8 emotions. Figure 4.1 below shows a summary of the model which is then compiled with adam optimizer, sparse_categorical_crossentropy loss function and Accuracy as the performance metrics.

```
Model: "functional_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_audio_ (InputLayer)    [(None, 40, 1)]           0
_________________________________________________________________
dense (Dense)                (None, 40, 64)            128
_________________________________________________________________
flatten (Flatten)            (None, 2560)              0
_________________________________________________________________
output_audio_ (Dense)        (None, 8)                 20488
=================================================================
Total params: 20,616
Trainable params: 20,616
Non-trainable params: 0
_________________________________________________________________
```

Figure 4.1: Audio Model 1

On the other hand, the Video model receives input in the form three-dimensional matrix (Height , Width, Number of channels) along with the Batch Size. The first two dimensions correspond to Height and Width of the image and the third dimension represents the 3 Channels(RGB) for the colored image. The input is passed to the keras dense layer which consists of 64 neurons. It is followed by the Dropout layer with '0.2' value. This inhibits some of the nodes to train and cause overfitting. The flatten layer added afterwards, converts the extracted information into single dimension which is passed to the final dense layer made up of 8 nodes having Softmax activation function classifies the multi-class emotions. Figure 4.2 below shows a summary of the model which is then compiled with adam optimizer, sparse_categorical_crossentropy loss function and keeping Accuracy as the performance metrics.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_frame_ (InputLayer)    [(None, 256, 512, 3)]     0
_________________________________________________________________
dense (Dense)                (None, 256, 512, 64)      256
_________________________________________________________________
dropout (Dropout)            (None, 256, 512, 64)      0
_________________________________________________________________
flatten (Flatten)            (None, 8388608)           0
_________________________________________________________________
output_frame_ (Dense)        (None, 8)                 67108872
=================================================================
Total params: 67,109,128
Trainable params: 67,109,128
Non-trainable params: 0
_________________________________________________________________
```

Figure 4.2: Video Model 1

While the individual Audio and Video models discussed above, were trained and tested for Accuracy, the main research objective was to test the Accuracy of the Multimodal model formed using the same architecture. Since the decision level fusion combines the decisions of both the classifiers into a common decision, the output layer consisting of Softmax activation function is removed from both the models and the flatten layers of each model is concatenated. The concatenated layer now consists of information from each of the modality. The final dense layer with 8 nodes and Softmax activation function is further added to classify the emotions.

The final Bimodal model is then compiled with efficient adam optimizer, sparse_categorical_crossentropy loss function suitable for sparse multi-class classification problem and Accuracy as the performance metrics. The model was trained with the batch size of 32 and 50 epochs, callbacks for EarlyStopping has been set to monitor the validation loss with min mode and patience value of 10. This would basically add a delay to the trigger in terms of number of epochs which could be considered even if there is no improvement. The model will stop training once the chosen performance measures stop to show any improvement. Figure 4.3 below shows the architecture of the final Bimodal model developed by the fusion of audio and video model at decision level.

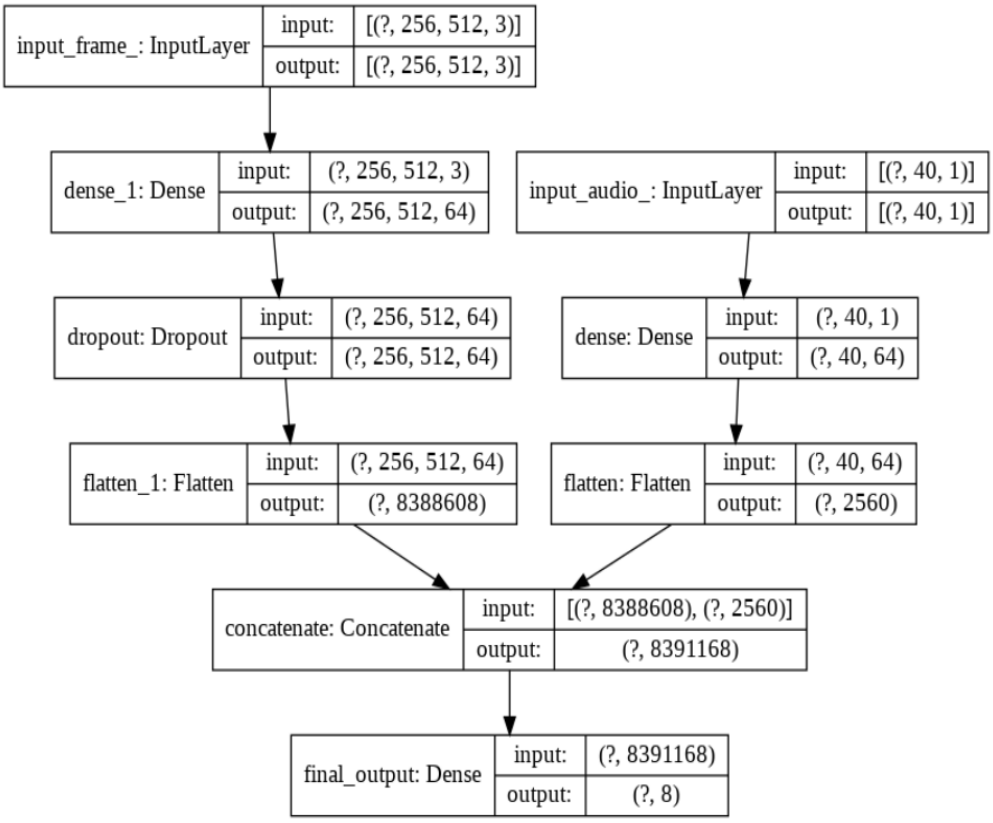


Figure 4.3: Model 1 Framework

4.1.2 Model 2: Convolutional Neural Networks

The second model is motivated by the success of Convolutional Neural Networks specially in the field of Image classification. The research has shown that 1D CNN has worked well for Audio classification as well. To serve the purpose of this experiment, 1D CNN model for Audio and 2D CNN model for Video is concatenated at the decision level removing the final output layer from each of the model.

The first layer of the 1D Audio CNN consists of 64 filters with kernel size 3. The activation function is set to be 'relu'. The input shape is specified to be (Batch Size, 40 , 1) where 40 represents the number of input features. The dropout layer with the value '0.2' has been added to minimize the loss between training and validation set. It is followed by the dense layer consisting of 16 nodes and 'relu' activation function. The next layer is Flatten layer which converts the data into single dimension followed by the final output dense layer. It is made up of 8 units to predict the probabilities of

emotions using Softmax activation function. Figure 4.4 below shows a summary of the Audio model which is then compiled with adam optimizer, sparse_categorical_crossentropy loss function and accuracy as the performance metrics.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_audio_ (InputLayer)    [(None, 40, 1)]           0
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 36, 64)            384
_________________________________________________________________
dropout_3 (Dropout)          (None, 36, 64)            0
_________________________________________________________________
dense_4 (Dense)              (None, 36, 16)            1040
_________________________________________________________________
flatten_4 (Flatten)          (None, 576)               0
_________________________________________________________________
output_audio_ (Dense)        (None, 8)                 4616
=================================================================
Total params: 6,040
Trainable params: 6,040
Non-trainable params: 0
_________________________________________________________________
```

Figure 4.4: Audio CNN Model

The first layer of the video model consists of 2D Convolutional layer with 64 filters and kernel size of 3 representing same value for height and width of the 2D convolution window. The input shape of (256,512,3) is fed to this convolutional layer where 3 represents the number of channels for RGB pictures. The convolution layer is followed by the MaxPooling layer with the pool size of (2,2) which reduces the spatial dimensions of the output layer. Dropout layer with '0.1' is added afterwards to stop the training of defined nodes and reduce overfitting. It is followed by the Dense layer which consist of 64 neurons with 'Relu' activation function. The flatten layer then converts all the extracted features into single dimension which is connected to the final dense output layer of 8 units. Softmax activation function in the final layer calculates the probabilities of all the emotions. Figure 4.5 below shows a summary of the Video model which is then compiled with adam optimizer, sparse_categorical_crossentropy loss function and keeping Accuracy as the performance metrics.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_frame_ (InputLayer)    [(None, 256, 512, 3)]     0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 256, 512, 64)      1792
_________________________________________________________________
max_pooling2d_5 (MaxPooling2 (None, 128, 256, 64)      0
_________________________________________________________________
dropout_6 (Dropout)          (None, 128, 256, 64)      0
_________________________________________________________________
dense_6 (Dense)              (None, 128, 256, 64)      4160
_________________________________________________________________
flatten_6 (Flatten)          (None, 2097152)           0
_________________________________________________________________
output_frame_ (Dense)        (None, 8)                 16777224
=================================================================
Total params: 16,783,176
Trainable params: 16,783,176
Non-trainable params: 0
_________________________________________________________________
```

Figure 4.5: Video CNN Model

The final Bimodal CNN model is formed by combining the decisions of both the classifiers at the decision level. The output layer is removed from both the models and the flatten layers of each model is concatenated. The concatenated layer now consists of information from both the modalities. The final dense layer with 8 nodes and Softmax activation function is further added to classify the emotions.

The model is then compiled with the adam optimizer, sparse_categorical_crossentropy loss function and Accuracy as the performance metrics. This model was also trained with the batch size of 32 and 50 epochs, callbacks with EarlyStopping has been created to check for the validation loss with min mode and patience value of 10. The model will stop training once the chosen performance measures stop to show any improvement. Figure 4.6 below shows the architecture of the final Bimodal CNN model developed by the fusion of audio and video model at decision level.

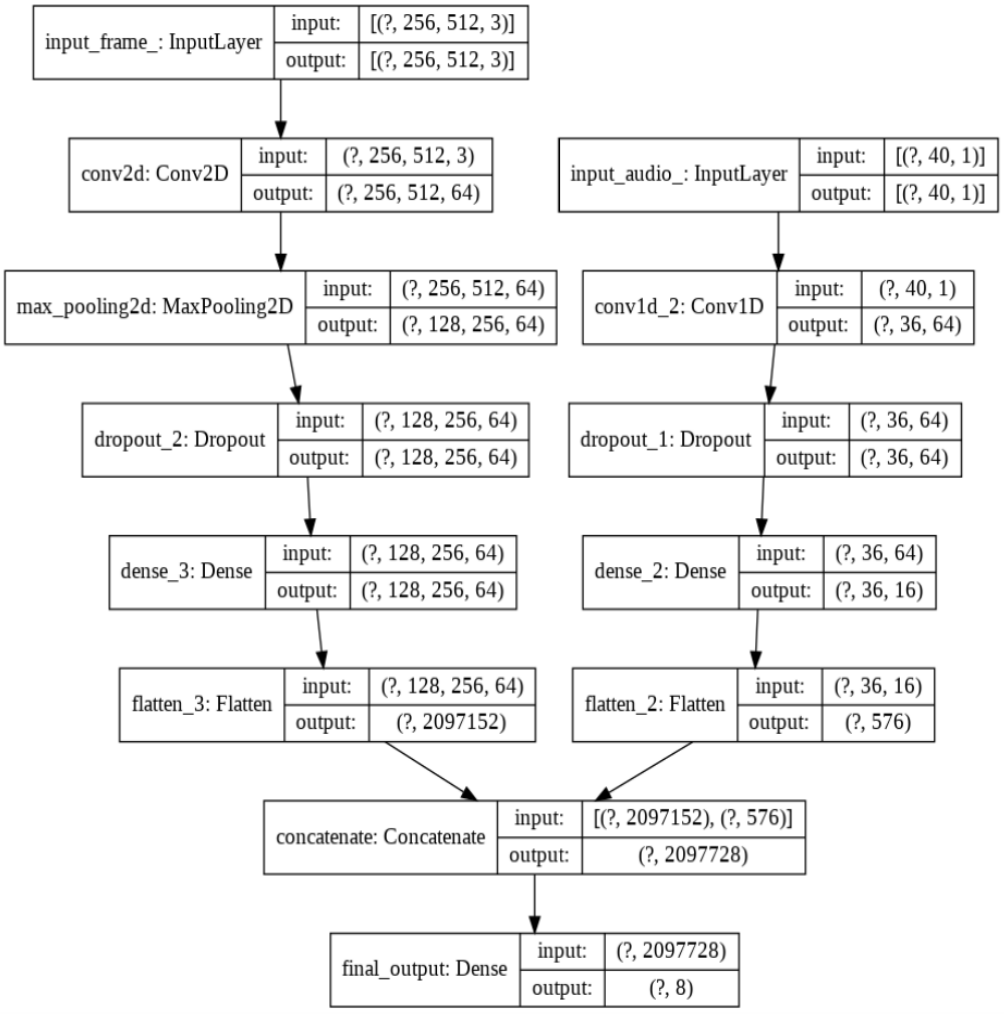


Figure 4.6: Model 2 Framework

4.1.3 Model 3: CNN – Dense Hybrid Network

This model is inspired by the results of the previously designed models. The same 1D CNN architecture for Audio model from Model 1 and Dense Only architecture for Video modality from Model 2 has been used to merge the information at the decision level and monitor the performance of the final Bimodal Classifier.

The Audio model consists of 576 features after the flatten layer while the Video model has 8388608 features at the flatten level layer. These decision features are concatenated to create the output of 8389184 features which is later connected to the final dense layer of 8 nodes with Softmax activation function to classify multi-category

emotions. Figure 4.7 below represents the architecture of the Bimodal CNN-Dense hybrid model developed by the decision level fusion.

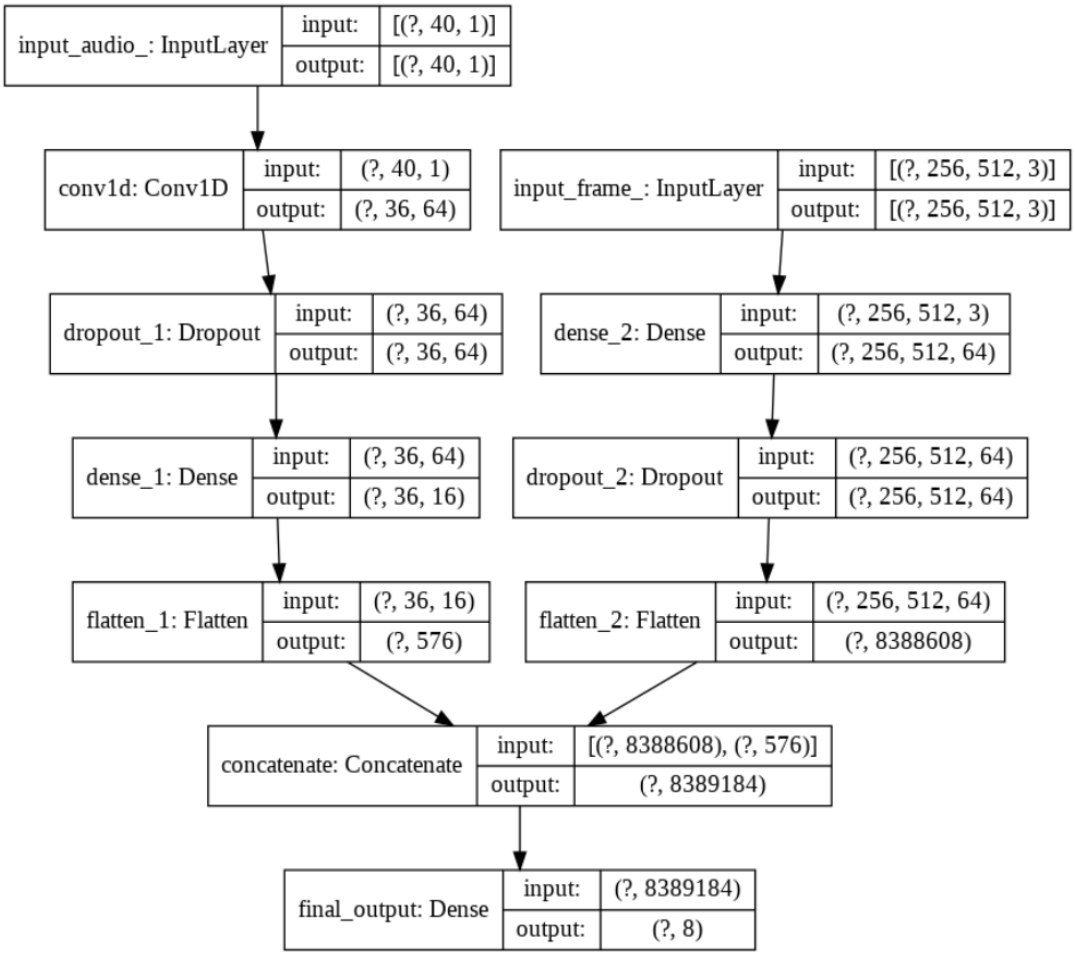


Figure 4.7: Model 3 Framework

The developed model is then compiled choosing adam optimizer, sparse_categorical_crossentropy loss function and Accuracy as the performance metrics. The batch size of 32 and 50 epochs was kept for training the model with the callbacks option. The validation loss with min mode and patience value of 10 will be monitored with the EarlyStopping option. This will ensure that the model doesn't over-train or under-train and will stop once the chosen performance measures do not show any significant improvement.

4.2 Experiment 2

The experiment 2 is based on the sub-question of the proposed research which is basically Task 2 discussed earlier. While creating a Multimodal classifier is computationally expensive, it could be proved crucial to find out if the performance of such models is deeply affected by the number of nodes in the final layer of each individual models. The higher number of nodes before the flattening layer outputs huge number of features, and lesser nodes outputs less features, how these features interact at the time of fusion remains the basis of this experiment.

4.2.1 Model 4: Wide CNN-Dense

Since the objective of the Wide CNN-Dense model is to test if increasing number of neurons in the final layer before flattening is affecting the performance of individual classifiers or the Bimodal classifier, the 1D Audio CNN is created using the same architecture defined in the model 3 with only 1 difference. The Dense layer before the flattening layer consist of 128 neurons. The same number of neurons have also been increased from 64 to 128 in the Dense layer of the Dense Only Video classifier. Due to Resource Exhaustion Issue(GPU- running out of memory), the number of neurons were sufficed to 128. This resulted into 4608 Audio features and 1,67,77,216 Image features after flatten layer which is about double of the CNN-Dense based Bimodal model developed earlier. The individual classifiers are compiled using *adam* optimizer, *sparse_categorical_crossentropy* loss function and *Accuracy* performance metrics. They are later trained with the batch size 32 and 50 epochs with callbacks for *EarlyStopping*.

In order to create the final Bimodal model, the output features obtained after the flatten layer from both the classifiers is concatenated resulting into *1,67,81,824* features which is followed by the final dense layer with 8 neurons and Softmax activation function classifying 8 categories of emotions.

The developed model is then compiled using the same parameters as used earlier. The adam optimizer, sparse_categorical_crossentropy loss function and Accuracy performance metrics is used for compiling the model. The model is then trained with specifying batch size as 32 and epochs to 50 using callbacks option. The validation loss with min mode and patience value of 10 is given in the callback to trigger the early stopping if the chosen performance measures are not showing any significant improvement. Below is the architecture of the Wide Bimodal Model formed by the late fusion of Audio and Video classifiers.

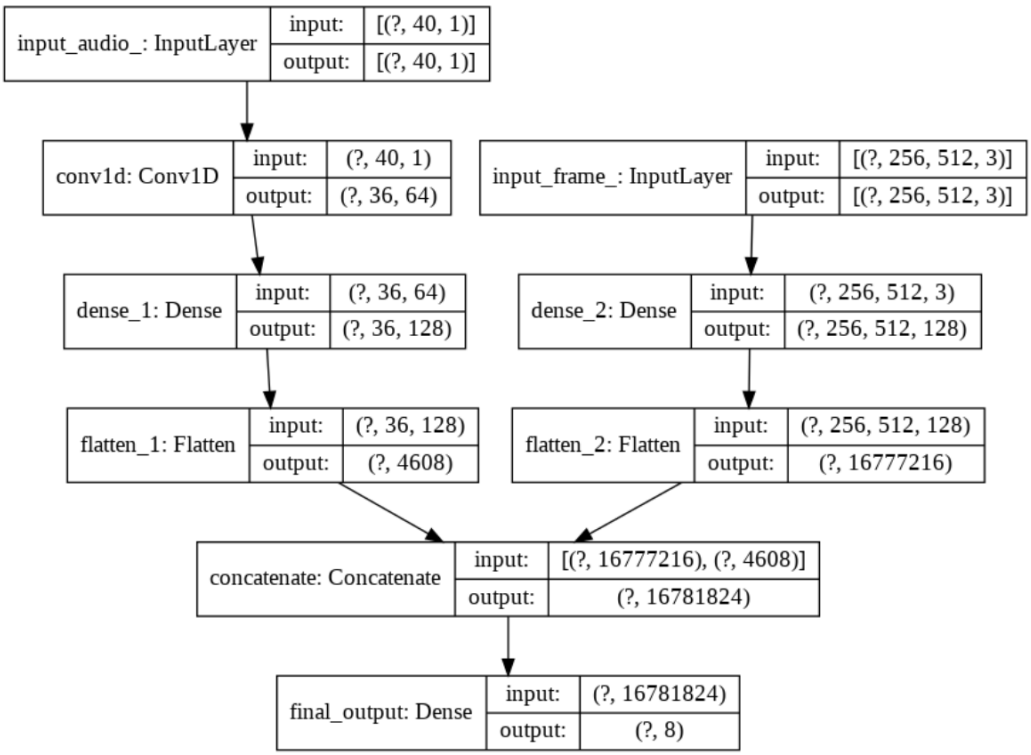


Figure 4.8: Model 4 Framework

4.2.2 Model 5: Narrow CNN-Dense

The idea behind the creation of this Narrow CNN-Dense model is to investigate how the late fusion will be influenced if the number of layers are more and has lesser number of nodes in the final layer before concatenation.

In order to do so, the first layer of the 1D Audio CNN consists of 64 filters with kernel size 3 with ‘relu’ activation function. The dropout layer with the value ‘0.2’ is added to exclude nodes while training. It is followed by three dense layers, each consisting of 32,16 and 8 nodes respectively with ‘relu’ activation function. The next layer is Flatten layer which converts the data into single dimension giving 288 features as output followed by the final output dense layer.

Likewise, the Video model is formed by using three dense layers each consisting of 128,64 and 32 nodes respectively with activation function ‘relu’. The flatten layer added afterwards, converts the extracted information into single dimension giving 41,94,304 features which is passed to the final output layer made up of 8 neurons. Both the models were then compiled separately with adam optimizer, sparse_categorical_crossentropy loss function and keeping Accuracy as the performance metrics.

This resultant 288 Audio features and 41,94,304 Image features obtained after flatten layer are comparatively less than the CNN-Dense based Bimodal model developed earlier, which is perfectly suited for the proposed experiment. The final Bimodal model after concatenation of both the features at decision level outputs to *41,94,592* features which is connected to the final dense layer made up of 8 neurons and Softmax activation function responsible for categorizing 8 emotions. The model is compiled and trained using the same options as discussed in the previous models. Below figure 4.9 shows the architecture of the Narrow Bimodal Model formed by the late fusion of Audio and Video classifiers.

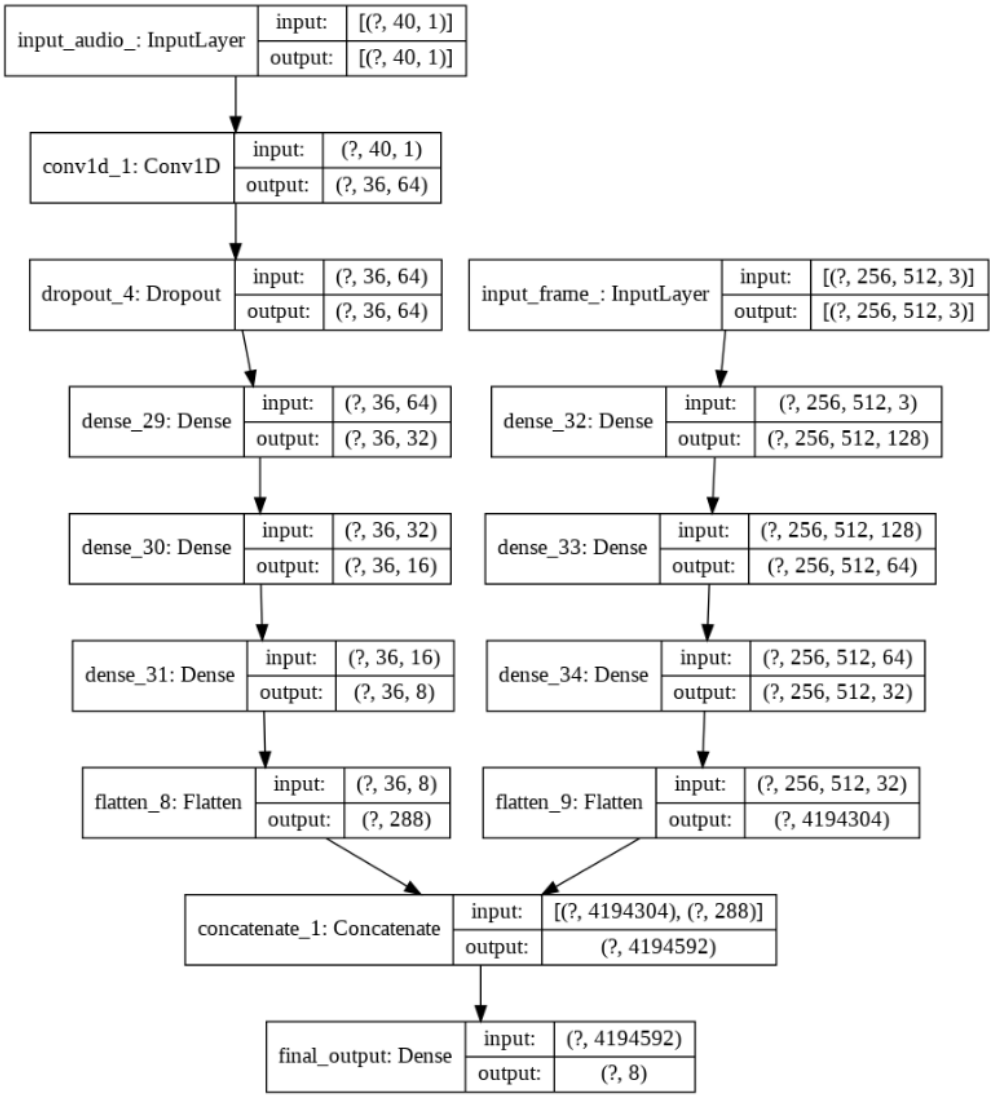


Figure 4. 9: Model 5 Framework

4.3 Results

The previous sections explained the architecture of each model and parameters which were used for its compilation and training. All the Bimodal and Unimodal models were later tested for accuracy on the Test data. The Table 4.1 shows the performance of each classifier on the test data. The results obtained from each model can be further compared and make decisions about the best performing model for the human emotion classification.

Experiments	Audio	Video	Bimodal
Experiment 1	**Test Accuracies**		
Dense Only Model	52.34%	79.23%	80.65%
CNNs Model	67.21%	59.67%	70.67%
CNN – Dense Hybrid model	67.21%	79.23%	83.30%
Experiment 2			
Wide CNN-Dense	65.38%	73.93%	80.86%
Narrow CNN-Dense	65.17%	69.25%	68.23%

Table 4. 1: Experiment Results

It is observed that the models developed as part of Experiment 1, testified the fact that the Bimodal classifiers in general perform better than the Unimodal models. The combination of CNNs did not do well with 70.67% accuracy, majorly because of the performance of Video classifier giving accuracy of 59.67%. Dense Only networks on concatenating the individual classifiers gave a good accuracy of 80.65%, mostly attributed to Video Dense Only network. Out of all the developed models, the CNN-Dense Hybrid Bimodal model outperformed all the models and gave test accuracy of 83.30%, whereas individual Audio and Video classifiers of the same network gave accuracy of 67.21% and 79.23% respectively. It was expected as Audio CNN and Video Dense classifiers were performing better than the rest individually.

In experiment 2, the same hybrid model is used with variations in the number of nodes and layers in the architecture. The result showed that Wide layered individual networks performed better than the Narrow individual networks. The Narrow Video model suffered with the problem of overfitting and could not give a decent accuracy while the Video model with extra nodes performed better and gave accuracy of 73.93%. It was evident that the performance of the Bimodal classifier was getting impacted if one of the unimodal models could not perform well. Further detailed analysis of the performance of the models is done in Chapter 5: ‘Evaluation and Analysis’ Section.

5 CHAPTER 5 – EVALUATION AND ANALYSIS

This chapter covers the detailed evaluation of all the final developed Bimodal models. The Unimodal models concatenated to form a Bimodal model will be compared with each other in terms of the Accuracy as part of Experiment 1. Also, for experiment 2, the final Bimodal classifiers developed in each of the experiments will be compared based on the Accuracy and the Precision value obtained for each class while making predictions. The section will be concluded with the Strengths and Limitations of the obtained results.

5.1 Evaluation of the Result

This section covers the results obtained after performing each of the experiments for the research sub-questions outlined in Chapter 1.

5.1.1 Experiment 1

As part of Experiment 1, three Bimodal models were designed to test whether the decision-level or late fusion of Unimodal models show any improvement in terms of accuracy.

The Model 1 is formed by the merging of two Dense Only models for Audio and Video. From the experiment, it is observed that the accuracy of Audio classifier alone on Test set was around 52.34% while Video classifier gave a better accuracy of 79.23%. Later, both these classifiers were fused at the decision level to form a Bimodal Emotion Recognition System.

The model stopped training after 20 epochs after noticing no improvement in the loss between training and validation. The model with the best accuracy was saved. The late fusion of both these individual models showed an overall increase of around 2% on unimodal video model corresponding to accuracy of 80.65%. In order to find out the best performing model in classifying emotions for comparative study, the model was

later tested on Test data. Figure 5.1 represents the classification report of the model 1. It shows that the precision value for Fearful class is equivalent to 100% while the other class had precision value between 78% - 94%. The lowest precision value of 64% was recorded for the Neutral emotion.

	precision	recall	f1-score	support
Neutral	0.64	1.00	0.78	34
Calm	0.90	0.92	0.91	66
Happy	0.79	0.96	0.87	57
Sad	0.89	0.86	0.87	76
Angry	0.78	0.78	0.78	65
Fearful	1.00	0.54	0.70	67
Surprise	0.67	0.94	0.78	66
Disgust	0.94	0.53	0.68	60
accuracy			0.81	491
macro avg	0.83	0.82	0.80	491
weighted avg	0.84	0.81	0.80	491

Figure 5.1: Model 1 Classification Report

The ***Model 2*** developed by the concatenation of Convolutional Neural Networks for each of the modality, did not perform well majorly because of the Video CNN model which was prone to overfitting because of the lesser number of samples. It gave only the test accuracy of 60%. However, the audio CNN model showed improved accuracy of 67.21% than the dense Audio model. Upon performing the late fusion with the Video CNN, the Bimodal model was trained with EarlyStopping parameter which stopped after 16 epochs after seeing no improvement in the accuracy and loss. Figure 5.2 shows the classification report of the model 2 on test set in predicting the emotions.

	precision	recall	f1-score	support
Neutral	0.84	0.91	0.87	34
Calm	0.57	0.98	0.72	66
Happy	0.71	0.77	0.74	57
Sad	0.61	0.87	0.72	76
Angry	0.85	0.54	0.66	65
Fearful	0.83	0.66	0.73	67
Surprise	0.79	0.52	0.62	66
Disgust	0.85	0.47	0.60	60
accuracy			0.71	491
macro avg	0.76	0.71	0.71	491
weighted avg	0.75	0.71	0.70	491

Figure 5.2: Model 2 Classification Report

It shows that the precision value for Angry and Disgust class was the highest equivalent to 85%, while Neutral and Fearful class had the precision value of 84% and 83% respectively. Precision value for the rest of the emotions were in the range of 57% to 79%. The lowest precision value of 57% for the CNN based model was recorded for Calm class.

The ***Model 3*** is the hybrid model formed by the merging of Audio CNN and Dense Only Video networks. It was developed observing the fact that for Audio modality CNN gave better accuracy while for Video modality; Dense networks performed better than CNNs. It gave accuracy of 84.90% on Validation set and 83.30% on Test set which is again higher than the individual audio and video models. This model stopped training after 24 epochs. The plots 5.3 and 5.4 below represent loss and accuracy during the training of the model.

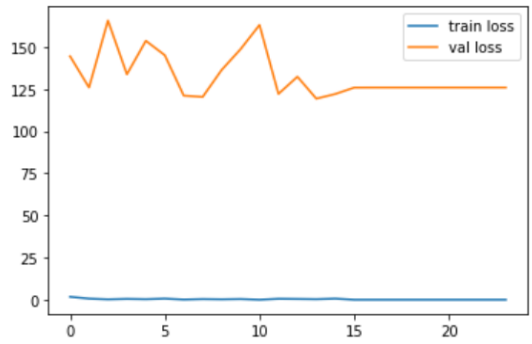


Figure 5.3: Model 3 Loss Plot

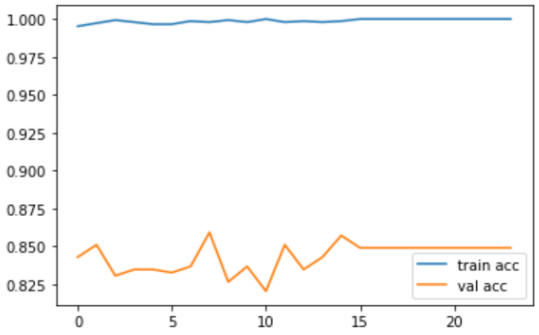


Figure 5.4: Model 3 Accuracy Plot

The classification report in figure 5.5 shows that Model 3 formed by the fusion of Audio CNN and Video Dense network performed better in classifying the emotions.

Classification Report

	precision	recall	f1-score	support
Neutral	0.85	0.97	0.90	34
Calm	0.92	0.92	0.92	66
Happy	0.85	0.93	0.89	57
Sad	0.88	0.83	0.85	76
Angry	0.75	0.86	0.80	65
Fearful	0.81	0.78	0.79	67
Surprise	0.82	0.71	0.76	66
Disgust	0.79	0.73	0.76	60
accuracy			0.83	491
macro avg	0.83	0.84	0.84	491
weighted avg	0.83	0.83	0.83	491

Figure 5.5: Model 3 Classification Report

The precision value for the Calm class is the highest for this model i.e. 92% . The lowest precision value is for the Disgust class equivalent to 79%. Rest of the emotions showed the precision value between 80-88% which is by far the most stable and highest value. The classification accuracy of the same model on test data was also the highest than the other two models.

A confusion matrix for model 3 is shown in Figure 5.6. Each row value in the matrix represents the actual class value whereas the column value represents the predicted

value of the emotions. The diagonal entries represent the instances which were classified correctly.

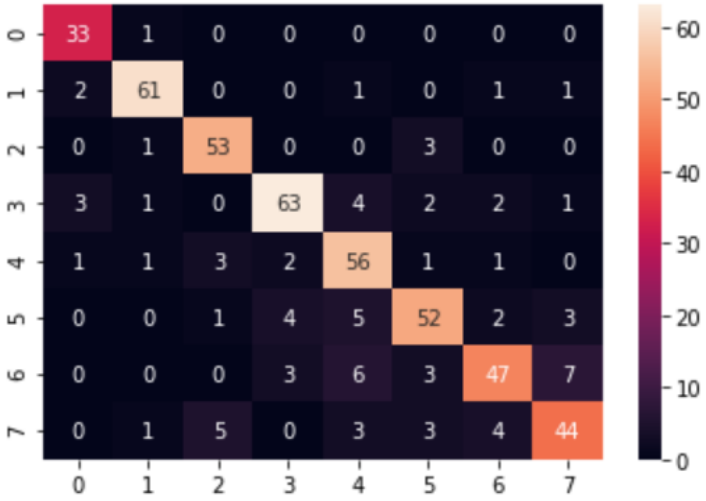


Figure 5.6: Confusion Matrix for Model 3

Based on the results from different architecture of Bimodal models, it is clear that late fusion of both the modalities has shown improvement over the Unimodal models as shown earlier in table 4.1 . And to form the basis of Experiment 2, classification report for all the models was also generated. Since the problem is of Multiclass emotion classification, the weighted average of Precision, Recall and F1 Score for each of the model is compared shown in table 5.1 which further attested the fact that Model 3 has performed better than the other two models in classifying the eight emotions not only in terms of Accuracy but also the Precision and F1 Score.

	Weighted Average		
	Precision	**Recall**	**F1 Score**
Model 1	84%	82%	80%
Model 2	75%	71%	70%
Model 3	83%	83%	83%

Table 5. 1: Performance Summary

Thus, it was chosen as the base model in order to tweak some changes in the architecture for the experiment 2 explained in the earlier sections.

5.1.2 Experiment 2

As part of Experiment 2, the effect of two different architectures - 'wide' and 'narrow' was investigated on the best performing Model 3 developed earlier.

The ***Model 4*** developed with the concatenation of wide layered individual audio and video models showed consistency in terms of regularization of loss between training and validation. The individual audio and video models with such architecture showed accuracy of 65.38% and 73.93% respectively on test set. The Bimodal model stopped training after 37 epochs while monitoring the validation loss which was set with the patience value of 10. It gave the accuracy of 99.46% on training set, 81.63% on validation set and 80.86% on test set. There was a significant rise of 7-7.5% in the accuracy after the late fusion of both the classifiers. Figure 5.7 and 5.8 show the state of the model during the training process. The loss between both is almost converging which signifies that the model is more robust than the other models trained earlier.

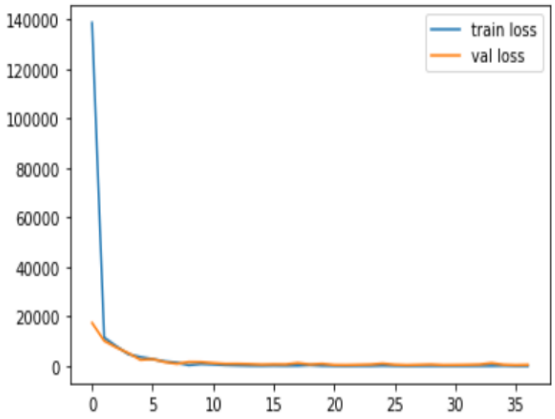


Figure 5.7: Model 4 Loss Plot

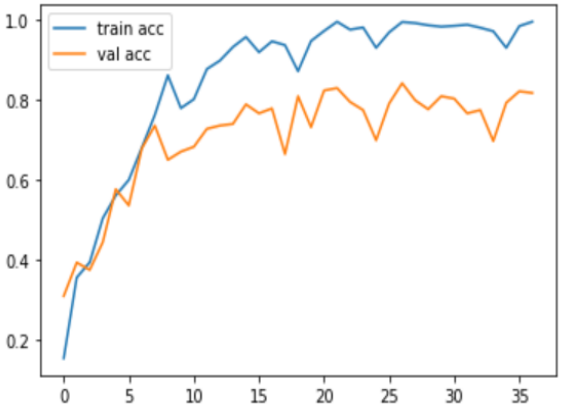


Figure 5.8: Model 4 Accuracy Plot

The classification report for the model generated after testing it on the test set, shows that the precision value for four of the emotions calm ,happy, sad and disgust is in the 90s, while the lowest is for the surprise emotion which is 67%. The overall classification accuracy of this model is 81% which is 2% lower than the model 3.

Classification Report

	precision	recall	f1-score	support
neutral	0.87	0.97	0.92	34
calm	0.98	0.89	0.94	66
happy	0.90	0.79	0.84	57
sad	0.95	0.72	0.82	76
angry	0.71	0.85	0.77	65
fearful	0.69	0.87	0.77	67
surprise	0.67	0.91	0.77	66
disgust	0.91	0.53	0.67	60
accuracy			0.81	491
macro avg	0.84	0.82	0.81	491
weighted avg	0.83	0.81	0.81	491

Figure 5.9: Model 4 Classification Report

Figure 5.10 represents the confusion matrix for the model 4, showing the instances in diagonal which were classified correctly.

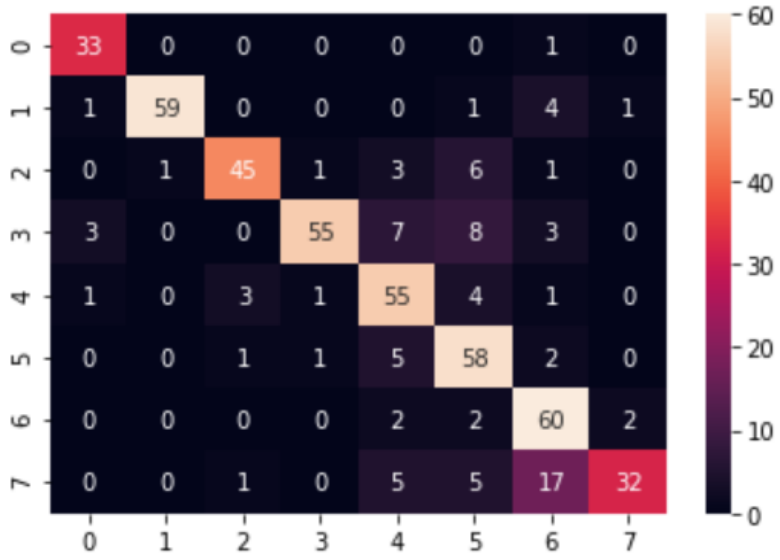


Figure 5.10: Confusion Matrix for Model 4

The ***Model 5*** is created in order to check if the deeper network with more layers performs better than the wider networks with a high number of nodes. There was no significant impact on the performance of the Audio model on increasing the number of layers which gave the test accuracy of 65.17% . However, it was observed that the

performance of the Video model dropped significantly and suffered the problem of overfitting. It gave the individual accuracy of 69.25% on test data. Both the individual classifiers were later merged to observe any change in the classification after the fusion. The model stopped training after 12 epochs due to early stopping seeing no further improvement in the accuracy. Figure 5.11 and 5.12 show the plot between the training and validation loss as well as training and validation accuracy.

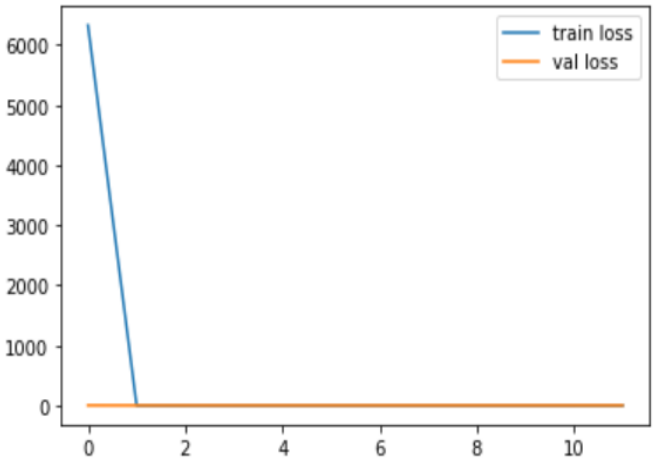


Figure 5.11: Model 5 Loss Plot

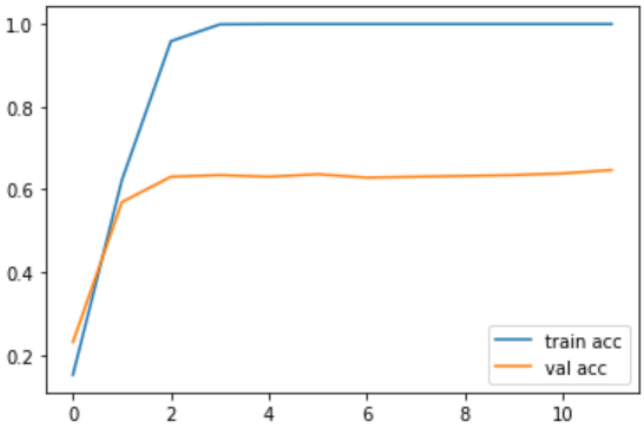


Figure 5.12: Model 5 Accuracy Plot

The model suffered with the problem of overfitting while refining captured features through the further layers. The figure 5.13 below shows the classification report generated after testing the model on the test data. It shows that the precision value for all the emotions is between 58%-72% except Neutral class which is 90%.

Classification Report

	precision	recall	f1-score	support
Neutral	0.90	0.79	0.84	34
Calm	0.72	0.82	0.77	66
Happy	0.69	0.63	0.66	57
Sad	0.70	0.71	0.71	76
Angry	0.72	0.58	0.64	65
Fearful	0.62	0.54	0.58	67
Surprise	0.66	0.73	0.69	66
Disgust	0.58	0.70	0.63	60
accuracy			0.68	491
macro avg	0.70	0.69	0.69	491
weighted avg	0.69	0.68	0.68	491

Figure 5.13: Model 5 Classification Report

Figure 5.14 below shows the confusion matrix for model 5 representing the correctly classified instances in coloured diagonally.

Figure 5.14: Confusion Matrix for Model 5

The table 5.2 shows the weighted average of Precision, Recall and F1 score which confirms that Model 4 developed with the concatenation of Wide layer unimodal models performed better than the Model 5 created by increasing the number of layers and lowering the nodes.

	Weighted Average		
	Precision	**Recall**	**F1 Score**
Model 4	84%	81%	81%
Model 5	69%	68%	68%

Table 5.2: Performance Summary of Model 4 and 5

The experiments showed that Bimodal models developed by combining the information from both the modality at the decision-level has performed better than the all of the Unimodal models. Since, the experiment 2 was to test the impact of early stage and late stage decision-level fusion, wide layer and narrow layered networks were formed respectively. The results showed that early stage decision level fusion helped in achieving the better accuracy than the late stage decision level fusion of Unimodal networks. The strengths and limitations of the result is discussed in the following sub-section.

5.2 Strength and Limitations of Result

As part of the research objectives, the various architecture of Unimodal and Bimodal classifiers were developed. While the number of samples chosen for this experiment is not as big as it would be required in general, to build a stable and effective Multimodal framework yet it provided the insights to test the proposed hypothesis. The designed Bimodal framework tested with different architectures further proved the fact that information extracted from two different modalities, if fused at decision level, could enhance the Accuracy in classifying human emotions. It also gave the insights that if the data is comparatively less and the model is overlearning the features, reducing the number of layers, and increasing the number of nodes in each of the Unimodal model before fusion is a good option. Deep networks with more layers work better when the

number of samples are more, which help the network in learning the high-level representation of the features filtering layer by layer. But this is not the case with the less samples where increased layer caused overfitting which resulted in the poor performance of Bimodal network.

Building a Multimodal emotion classifier requires the pre-processed data from multiple modalities to feed it as the input. The training of individual models and then, the final fused model require huge computational power. The optimization of models based on the training performance is not only time consuming but also, computationally expensive. Thus, keeping all these things in mind, a lesser number of samples were used to develop this Bimodal framework which resulted in most of the models suffering from the problem of overfitting. Also, efficient pre-processing of more data and optimizing the models would require additional computational power and time, but because of lack of both, not much focus was put on the same. The discussed approaches and results could be only effective for the smaller datasets

6 CHAPTER 6 – DISCUSSION AND CONCLUSION

This chapter covers the brief summary of the outcomes generated from the various experiments as per the proposed research. It also details out the future work which could be done on top of this research.

6.1. Research Overview

As per the existing research, combining inputs from multiple modalities is a challenging task because of the steps involved in pre-processing of multi-dimensional data plus the amount of time and computing power it requires. Thus, the proposed research to develop a Bimodal emotion classifier was conducted on smaller dataset RAVDES in two phases. The first phase involved the creation of different simple Unimodal networks using Dense and CNN for each of the modality and then performing fusion at the decision level which could help in investigating the fact that combining modalities can enhance the performance in classifying human emotions. The second phase was to make changes in the architecture of the best performing model in terms of the narrow and wide layered network. Both the models in the second phase were evaluated and compared to decide on the proposed hypothesis.

6.2. Problem Definition

The research problem was originally described by the research question:
"Can early-stage decision-level fusion improve the accuracy of a Bimodal model in emotion-classification problems?"
It was later expanded by the below two sub-questions derived from the same:
Sub-question 1: Do Bimodal models give better accuracy than the Unimodal models alone?
Sub-question 2: Should the individual models output a large number of nodes (a "wide" final layer) or a small number of nodes (additional layers leading to a "narrow" node) ?

The main purpose of the research is to validate the hypothesis which states that the performance of Bimodal deep learning model designed to classify human emotions with the decision-level fusion of Audio and Video model is not affected by the number of nodes in the final layer of each individual models. And also, draw conclusions which could answer the research sub-questions discussed earlier.

The comparison between the different models showed that Bimodal models performed better than Unimodal. The results also showed that the Bimodal network developed by the decision level fusion of wide layered(more number of nodes in final layer) CNN-Dense network outperformed the narrow CNN-Dense Bimodal network giving basis to reject the proposed Null hypothesis.

6.3. Experimentation, Evaluation and Results

The primary objective of this research was to develop an efficient Bimodal framework which could classify human emotions. There have been a lot of successful projects in Audio and image classification alone however, extracting emotions by combining modalities is far more complex and is relatively new as discussed in the Chapter 2. Various levels of data pre-processing and feature level or decision level fusion of Unimodal networks in general, increases the overhead time. In order to do so with limited amount of time and computing power, not much of the focus was put on the pre-processing and further optimization of the developed models. Instead, this research primarily focused on combining the decisions of both the modalities at the early and late stage and observe the way fusion interacts if there are any changes in the number of nodes in the final layer.

It was observed from the experiments that combining modalities can enhance the performance in classifying emotions. The more complex individual networks were not developed to reduce the complexity at the time of performing fusion. The CNN network for audio classification and Dense only network for Video performed better individually and thus, were used to form a final Bimodal model. It was also observed that the final wide layer with additional nodes/neurons in the individual Audio and Video models performed way better than the individual models which output a smaller number of nodes in the final layer. The early stage decision-level fusion of the output

features from each of the modality performed better than late stage decision-level fusion.

The accuracy for the wide CNN-Dense network was 80.86% while the narrow CNN-Dense network gave accuracy of 68.23%. The models were compared not only on the basis of Accuracy but also, Precision, Recall and F1 score which confirmed the fact that the wide layered network models are robust in case of smaller datasets. This gave the basis to reject the null hypothesis which states that the number of nodes in the final layer of each of the Unimodal networks do not impact the performance of Bimodal networks at the time of fusion.

6.4. Contributions and Impact

The dataset RAVDES used for developing a Bimodal emotion classifier is new. Most of the work on this dataset have been related to Audio classification but combining its video data together for further Bimodal classification is new and has been explored as part of this research. The experiments done on this data also gave interesting insights about the impact of narrow and wide networks based on the number of nodes in the final layer at the time of decision level fusion.

While building a Multimodal emotion classifier is computationally expensive, the conducted work shows the possibility of developing such classifiers with the limited computational capacity. The work presented here can be well suited for a lesser amount of data for each modality. One key take away from this research is the use of wide layered individual networks. Wide layered network even with the single layer can improve the performance of Bimodal classifiers at the time of decision level fusion.

6.5. Future Work and Recommendations

With the efficient systems, the conducted research can be put to a very good use. Inclusion of other data preprocessing techniques specific to each modality like using mel-spectrograms for Audio or noise injection, augmentation etc. can further help in exploring the possibilities of developing an efficient bimodal system. Another

interesting concept would be to observe the performance of the Bimodal framework when fusion is done at the feature level.

Because of lack of time, attention layers could not be tried with the emotion recognition framework. Introducing attention layers to extract the relevant output features from each modality before fusion remains the future work for this research. One of the key recommendations for conducting such types of research would be to acquire a system that could provide a lot of computational power. It would not only save time in experimenting with high dimensional data but also could help in fine tuning the developed framework. However, with not so powerful computing systems building a wide-layered neural network instead of narrow-layered networks for each of the modality can help in achieving considerable accuracy in classifying emotions.

BIBLIOGRAPHY

Abdulsalam, W., Alhamdani, R. & Abdullah, M.(2019). Facial Emotion Recognition from Videos Using Deep Convolutional Neural Networks. *International Journal of Machine Learning and Computing*, *9*(1), 4-19.

Agarwal, A., Yadav,A. and Vishwakarma, D., K.(2019).Multimodal Sentiment Analysis via RNN variants. IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD), pp. 19-23. doi: 10.1109/BCD.2019.8885108.

Asghari, M., H. and Jalali, B.(2015). Edge detection in digital images using dispersive phase stretch. International Journal of Biomedical Imaging, pp. 1–6.

Atmaja, B., T. & Akagi, M. (2019). Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model. *IEEE International Conference on Signals and Systems*, p. 5.

Batbaatar, E., Li, M. & Ryu, K.(2019). Semantic-Emotion Neural Network for Emotion Recognition From Text. *IEEE Access, 7,* 111866-111878. doi: 10.1109/ACCESS.2019.2934529

Cambria, E., Hazarika, D., Poria, S., Hussain, A., & Subramanyam, R. (2017). Benchmarking multi- modal sentiment analysis. *In International conference on computational linguistics and intelligent text processing*, pp. 166–179.

Chen, F., Luo, Z., Xu, Y. & Ke, D.(2019). Complementary Fusion of Multi-Features and Multi- Modalities in Sentiment Analysis. *IEEE International Conference of Computing*, *2*(1), 9.

Chen, M. , Jiang, L., Ma, C., & Sun, H. (2019). Bimodal Emotion Recognition Based on Convolutional Neural Network*. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing (ICMLC '19),* 178–181. Doi: https://doi.org/10.1145/3318299.3318347

Cornejo, J. & Pedrini, H. (2019). Bimodal Emotion Recognition Based on Audio and Facial Parts Using Deep Convolutional Neural Networks. *IEEE International Conference On Machine Learning And Applications (ICMLA),* 111-117, doi: 10.1109/ICMLA.2019.00026.

Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product review. *In Proceedings of the 12th international conference on World Wide Web* (pp 519-528).ACM.

Do, H., Huynh, H., Nguyen, K., Nguyen, N. & Nguyen, A. (2019). Hate Speech Detection on Vietnamese Social Media Text using the Bidirectional-LSTM Model.

Gallo, I., Calefati, A., Nawaz, S. & Janjua, M.(2019). Image and Encoded Text Fusion for Multi- Modal Classification. *IEEE International Conference of Computing, 5*(1).

Gan, L., Benlamri, R. & Khoury, R.(2017). Improved Sentiment Classification by Multi-Modal Fusion. *IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService)*,11-16. doi: 10.1109/BigDataService.2017.11.

Hagen, A., Pellom, B., Van Vuuren, S., & Cole, R. (2004). Advances in children's speech recognition within an interactive literacy tutor. In *Proceedings of HLT-NAACL 2004: Short Papers* (pp. 25-28).

Hammad, M., Liu, Y. and Wang, K.(2019). Multimodal Biometric Authentication Systems Using Convolution Neural Network Based on Different Level Fusion of ECG and Fingerprint. *IEEE Access, 7,* 26527-26542. doi: 10.1109/ACCESS.2018.2886573

Hu, A., & Flaxman, S. (2018). Multimodal Sentiment Analysis To Explore the Structure of Emotions. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. doi:10.1145/3219819.3219853

Huang, K. Y., Wu, C. H., Hong, Q. B., Su, M. H., & Chen, Y. H. (2019). Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5866-5870). IEEE.

Kaur R. & Kautish S. (2019). Multimodal Sentiment Analysis: A Survey and Comparison. *International Journal of Service Science, Management, Engineering, and Technology, 10*(2),38-58.

Lalitha, S. & Thyagharajan, K.(2019). Micro-Facial Expression Recognition in Video Based on Optimal Convolutional Neural Network (MFEOCNN) Algorithm. *International Journal of Engineering and Advanced Technology (IJEAT) , 9*(1), 10.

Lee, C.W., Song, K.Y., Jeong, J., & Choi, W.Y. (2018). Convolutional Attention Networks for Multimodal Emotion Recognition from Speech and Text Data. doi: 10.18653/v1/w18-3304

Lu, H., Zhang, H., & Nayak, A. (2020). A Deep Neural Network for Audio Classification with a Classifier Attention Mechanism. *arXiv preprint arXiv:2006.09815*.

Nefian, A. V., Liang, L., Pi, X., Liu, X., & Murphy, K. (2002). Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP Journal on Advances in Signal Processing*, *2002*(11), 783042.

Nemati, S., Rohani, R., Basiri, M., Abdar, M., Yen, N. & Makarenkov, V.(2019). A Hybrid Latent Space Data Fusion Method for Multimodal Emotion Recognition. *IEEE Access, 7,* 172948- 172964.

Orjesek, R., Jarina, R., Chmulik, M. & Kuba, M., 2019. DNN Based Music Emotion Recognition from Raw Audio Signal. IEEE International Conference on Signals and Systems, 3(1), p. 6.

Poria, S., Cambria, E., Howard, N., Huang, G., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing, 174*, 50-59. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0925231215011297

Poria, S., Cambria, E., Hazarika, D., Mazumder, N., Zadeh, A. and Morency L. (2017).Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis. *IEEE International Conference on Data Mining (ICDM),*1033-1038. doi: 10.1109/ICDM.2017.134.

Saravanan, A., Perichetla, G. & Gayathri, D., K. (2019). Facial Emotion Recognition using Convolutional Neural Networks. *International Journal of Machine Learning and Computing, 3*(6),6.

Sharma, A. & Mansotra, V.(2019). Multimodal Decision-level Group Sentiment Prediction of Students in Classrooms. *International Journal of Innovative Technology and Exploring Engineering , 8*(12), 4902-4909.

Snoek, C. G., Worring, M., & Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 399-402).

Soleymani, M., García, D., Jou, B., Schuller, B.W., Chang, S., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image Vis. Comput.,* 65, 3-14.

Tocoglu, M., Ozturkmenoglu, O. & Alpkocak, A.(2019).Emotion Analysis from Turkish Tweets Using Deep Neural Networks. *IEEE Access, 7,* 183061-183069.

Tripathi, S., Tripathi, S. & Beigi, H.(2019).Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning. *IEEE International Conference of Computing,* 10(1).

Ullah, M., A., Islam, M., M., Azman, N., B., and Zaki, Z., M. (2017).An overview of Multimodal Sentiment Analysis research: Opportunities and Difficulties. *IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR),*1-6. doi: 10.1109/ICIVPR.2017.7890858.

Verma, S., Wang, C., Zhu, L. & Liu, W.(2019). DeepCU: Integrating both Common and Unique Latent Information for Multimodal Sentiment Analysis. *International Joint Conference on Artificial Intelligence, 1*(10), 3627-3634.

Wang, H., Meghawat, A., Morency, L. and Xing, E.,P. (2017). Select-additive learning: Improving generalization in multimodal sentiment analysis. *IEEE International Conference on Multimedia and Expo (ICME)*,949-954. doi: 10.1109/ICME.2017.8019301

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding Abuse: A Typology of Abusive Language Detection Subtasks. the proceedings of the 1st Workshop on Abusive Language Online., pp. 066-149.

Xi, C., Wang, Lu, G., & Yan, J. (2020). Multimodal sentiment analysis based on multi-head attention mechanism. *Proceedings of the 19th ACM International Conference on Multimodal Interaction,*163-171. Retrieved from https://doi.org/10.1145/3380688.3380693

Zhang, D., Wu, L., Li, S., Zhu, Q. and Zhou, G.(2019).Multi-Modal Language Analysis with Hierarchical Interaction-Level and Selection-Level Attentions. *IEEE International Conference on Multimedia and Expo (ICME)*,724-729. doi: 10.1109/ICME.2019.00130.

Zheng, H. & Yang, Y., 2019. An Improved Speech Emotion Recognition Algorithm Based on Deep Belief Network. *IEEE International Conference on Power, Intelligent Computing and Systems, 10*(2), p. 5.

Zheng, J. & Zheng, L., 2019. A Hybrid Bidirectional Recurrent Convolutional Neural Network Attention-Based Model for Text Classification. *IEEE Access, 7,* pp.106673-106685.

APPENDIX

Model 1:

Classification Report

	precision	recall	f1-score	support
Neutral	0.64	1.00	0.78	34
Calm	0.90	0.92	0.91	66
Happy	0.79	0.96	0.87	57
Sad	0.89	0.86	0.87	76
Angry	0.78	0.78	0.78	65
Fearful	1.00	0.54	0.70	67
Surprise	0.67	0.94	0.78	66
Disgust	0.94	0.53	0.68	60
accuracy			0.81	491
macro avg	0.83	0.82	0.80	491
weighted avg	0.84	0.81	0.80	491

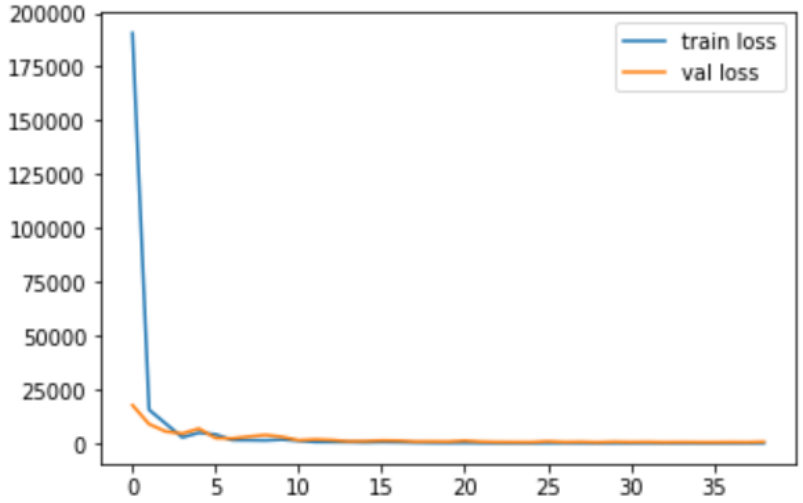
Model 2:

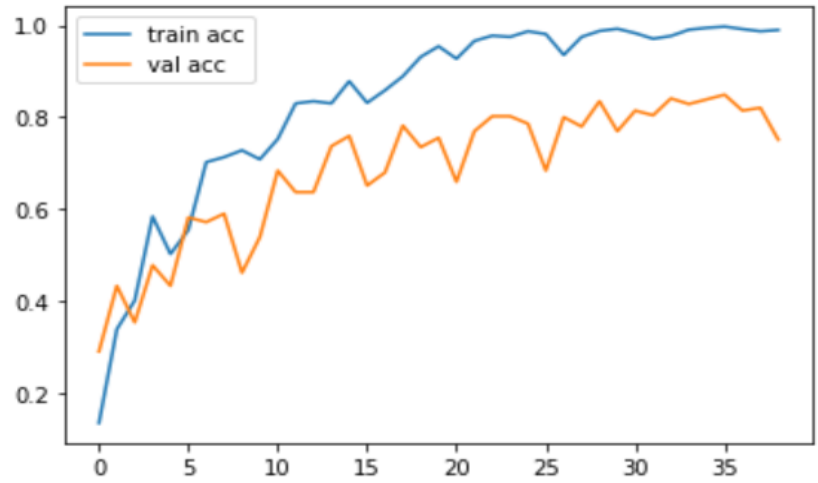
Classification Report

	precision	recall	f1-score	support
Neutral	0.84	0.91	0.87	34
Calm	0.57	0.98	0.72	66
Happy	0.71	0.77	0.74	57
Sad	0.61	0.87	0.72	76
Angry	0.85	0.54	0.66	65
Fearful	0.83	0.66	0.73	67
Surprise	0.79	0.52	0.62	66
Disgust	0.85	0.47	0.60	60
accuracy			0.71	491
macro avg	0.76	0.71	0.71	491
weighted avg	0.75	0.71	0.70	491

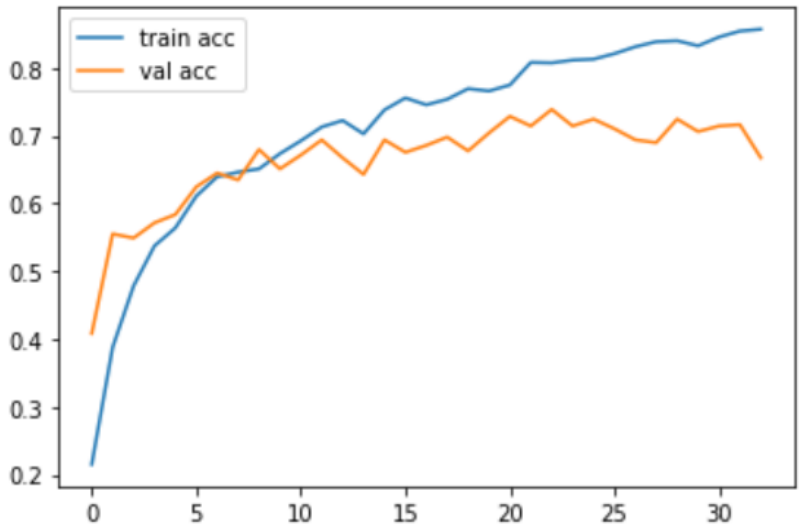
Model 4:

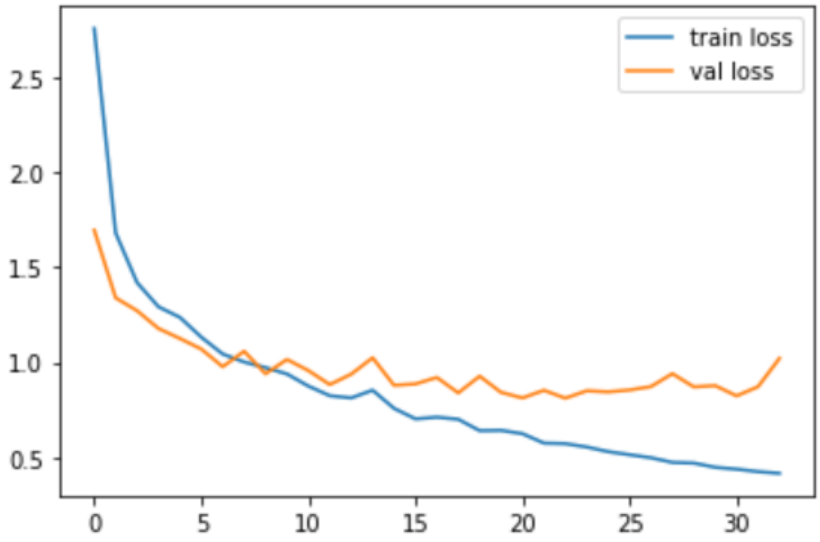
Video Model





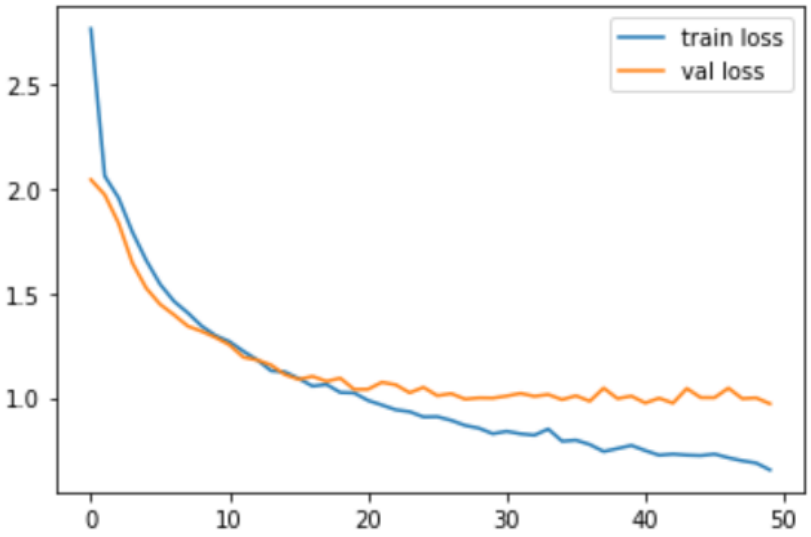
Audio Model

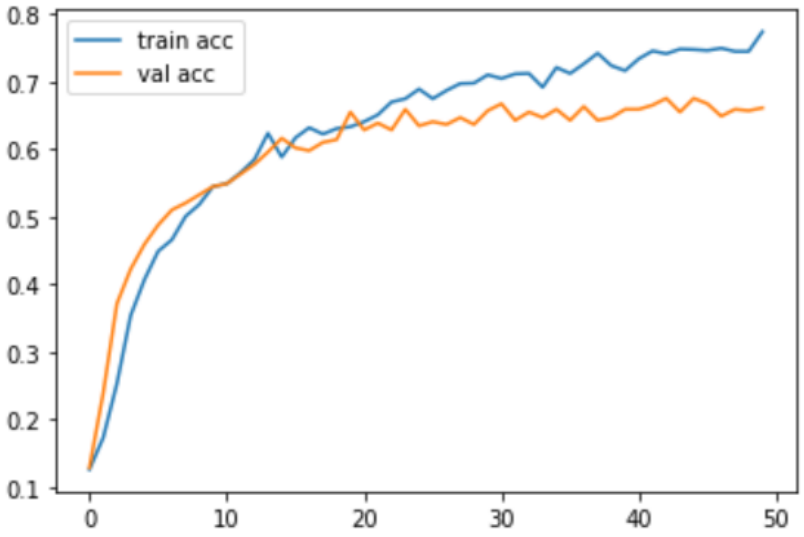

train acc
val acc
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
5
10
15
20
25
30


train loss
val loss
2.5
2.0
1.5
1.0
0.5
0
5
10
15
20
25
30

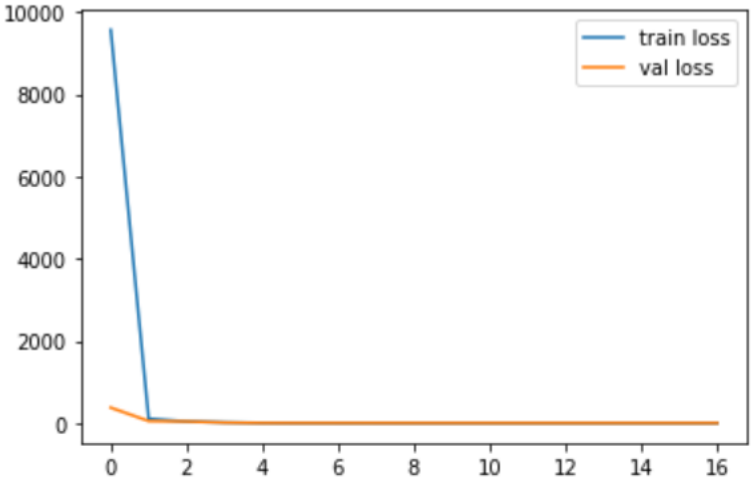
Model 5:

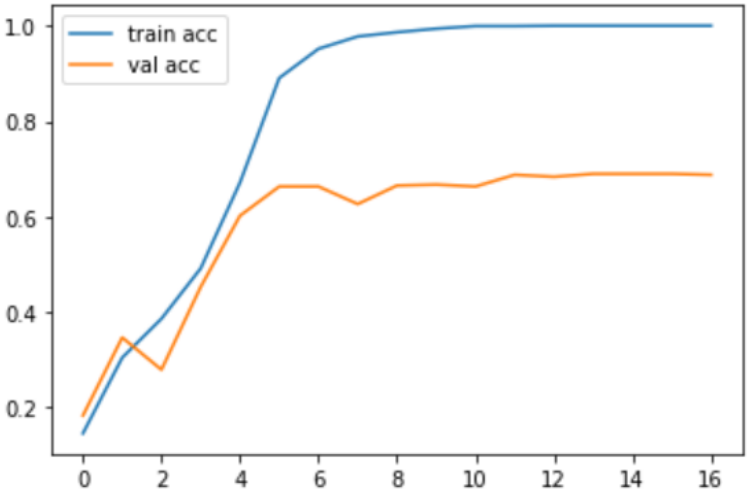
Audio Model





Video Model


train loss
val loss
10000
8000
6000
4000
2000
0
0
2
4
6
8
10
12
14
16


train acc
val acc
1.0
0.8
0.6
0.4
0.2
0
2
4
6
8
10
12
14
16