Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Speech Emotion Recognition (SER) is an emerging field at the intersection of communication and technology, focusing on deciphering emotional cues embedded in spoken language. SER has evolved from traditional signal processing techniques to encompass advanced methodologies like k-nearest Neighbours (KNN), Support Vector Machines (SVM), and Random Forest, which leverage engineered features and statistical analysis to uncover emotional content. More recently, deep learning techniques like Recurrent Neural Networks (RNNs) have captured sequential patterns inherent in speech, while Transformers utilise self-attention mechanisms to decode nuanced connections within speech data. Convolutional Neural Networks (CNNs), originally designed for images, have also found application in SER by autonomously extracting hierarchical features from speech signals. This paper presents a comprehensive exploration of the integration of CNNs into the SER landscape, offering insights into its evolution, methodologies, challenges, and future directions.

This chapter provides background on the evolution of SER. Traditional machine learning algorithms initially dominated the field before the introduction of deep neural networks enabled new advances. The chapter explains the motivation to leverage state-of-the-art deep learning architectures including recurrent, convolutional, and Transformer-based models to push the performance boundaries of SER. The research objectives and dissertation outline are also presented. Overall, the chapter lays the groundwork for investigating novel applications of deep learning to sentiment analysis and emotion recognition in the text domain.

1.1. Background

Speech is a fundamental aspect of human cognitive abilities, highlighting a crucial skill that sets our species apart in terms of language and communication. But before speech even forms, emotions serve as fundamental indicators of our internal state of mind. Long before infants utter their first words, their sounds are animated by expressions that manifest their emotional states – a precursor to the development of verbal language. This highlights the inherent and all-encompassing presence of emotions as instruments of communication that manifest ahead of structured linguistic frameworks.

Within the complex framework of human communication, emotions stand as a vital thread that weaves depth and context into the fabric of our interactions. Speech, as the primary vehicle for conveying thoughts and ideas, serves as a conduit through which emotions find expression. Beyond the explicit content of the spoken words, speech carries the intonations, cadences, and pauses that lend emotional texture to our communication. These subtleties provide a nuanced understanding of not only the intended message but also the underlying sentiments, intentions, and emotional states of the speaker. The capability to decipher and recognize these emotional cues embedded within spoken language holds substantial implications across a spectrum of domains, which will be discussed in Section 5.2.

The journey of Speech Emotion Recognition (SER) has been one of continual evolution, traversing a spectrum of methodologies from conventional signal processing techniques to the powerful realm of deep learning. In the initial stages, attempts at emotion recognition from speech data relied heavily on handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and energy variations. Although these early endeavours represented significant strides in understanding emotional cues, they often fell short in capturing the intricate and dynamic patterns woven into speech data by human emotions.

1.1.1. Traditional algorithms

The landscape of SER thrives on a diverse ecosystem of techniques, each with distinct strengths in deciphering the emotional nuances embedded within spoken language. Traditional algorithms, such as k-nearest Neighbours (KNN), Support Vector Machines (SVM), (Huang et al., 2014) and Random Forest (Lim et al., 2016), have contributed significantly. These methods leverage carefully engineered features and statistical analyses to unearth patterns indicative of emotional content in speech. While their contributions are noteworthy, the intricate nature of emotions often surpasses the capacity of predefined features to capture their complexity fully.

1.1.2. RNNs and Transformer Models

The emergence of deep learning models has further broadened the spectrum of techniques employed in Speech Emotion Recognition (SER). Among these, Recurrent Neural Networks (RNNs) have played a pivotal role by introducing the crucial concepts of memory and temporal dynamics. By adeptly capturing sequential patterns inherent in spoken language, RNNs have led us to revolutionise the way we approach SER. Emotions often unveil themselves gradually over time, making the inherent capability of RNNs to model temporal dependencies exceptionally valuable in this context. However, the journey of innovation did not merely halt with RNNs (Kerkeni et al., 2022). A more recent breakthrough has been witnessed with the adaptation of Transformers, in the realm of SER, which makes use of self-attention mechanisms. While initially designed to enhance natural language processing tasks, Transformers have seamlessly transitioned into the domain of SER. They can capture extensive temporal relationships and nuanced connections within speech data.

1.1.3. Convolutional Neural Networks (CNNs)

CNNs, initially created for image analysis, marked a turning point in various fields. Their capability to extract hierarchical features from raw data revolutionised SER. While designed for images, CNNs extended to discern temporal hierarchies in sequential data like speech signals, uncovering emotional signatures. In addition to image analysis, CNNs excel at recognising subtle spectral patterns in audio data.

1.1.4. ImageNets

Amidst this dynamic landscape, the incorporation of pre-trained convolutional models originally designed for image analysis, such as Residual Networks (ResNets) and Inception Networks, added an intriguing layer to the SER narrative. ResNets, renowned for their exceptional prowess in mitigating

the vanishing gradient predicament, have seamlessly translated their success from the realm of computer vision to the intricate domain of emotion analysis. The vanishing gradient problem, a notorious hindrance in deep neural networks, arises when gradients during training diminish exponentially, impeding effective weight updates. ResNets' ingenious architecture, featuring residual connections, enables the seamless flow of gradients and facilitates the training of deeper models. Inception Networks, characterised by parallel convolutions, provided the capability to simultaneously capture multi-scale features. This architectural innovation empowers the network to simultaneously process information at various spatial scales, akin to the multi-layered nature of human emotions. Human emotional expression is a complex amalgamation of subtle cues and pronounced signals, often spanning different temporal and contextual dimensions (Szegedy et al., 2015).

1.2. Aims and Objectives

This paper embarks on a comprehensive exploration of the integration of CNNs into the landscape of SER. By synthesising the advancements in this dynamic field, we seek to empower researchers, practitioners, and enthusiasts alike with insights that foster deeper comprehension and innovation.

Our main goal is to

- Leverage deep convolutional neural networks to accurately classify emotions.
- Explore state-of-the-art ImageNet architectures such as ResNet and InceptionNet to recognize emotions from speech with higher performance.
- Aggregate multiple public emotion speech datasets to train robust models that generalise across different speakers and contexts.
- Develop a reliable end-to-end emotion recognition system from speech using deep learning techniques to extract emotionally salient features.
- Provide a basis for a reliable Emotion Recognition System

1.3. Outline of Dissertation

In this dissertation, we embark on a comprehensive journey through the landscape of Speech Emotion Recognition (SER), driven by the potent capabilities of Convolutional Neural Networks (CNNs).

- While introducing our research, we delve into the world of Speech Emotion Recognition (SER). We discuss the evolution of this field, its significance, and the motivations behind our research. Furthermore, we outline our research objectives and the goals that drive our exploration.
- The second chapter is a comprehensive review of the existing body of research in SER. We analyse various studies, their contributions, and how they have shaped the current landscape of SER.
- In the third chapter, we present our research methodology. We detail the tools, datasets, and techniques we have selected along with a roadmap to approach the problem of SER.

- The next portion is dedicated to the presentation and analysis of the results obtained. We discuss the outcomes of our research and examine the factors that have influenced these results.
- The final chapter serves as the conclusion to our dissertation. Here, we summarise our research journey, discuss its implications, and acknowledge its limitations. We also outline potential directions for future research and development in the dynamic field of SER.

Chapter 2: Literature Review

Speech Emotion Recognition (SER) is a specialised area within the broader field of speech processing, dedicated to recognizing and interpreting emotional states from spoken language (Ayadi et al., 2011). It has been an active research domain due to its applications in human-computer interaction, affective computing, and psychological analysis. Over the years, extensive research has been conducted in the field of SER. For the past decade, traditional machine learning algorithms like Support Vector Machines (SVM), K-Nearest Neighbors (kNN), and Artificial Neural Networks (ANN), have been widely utilised for this task. These methods provided valuable insights, but they often faced limitations in handling the complexities of speech data and capturing intricate emotional nuances. However, in recent times, the introduction of advanced deep learning techniques has revolutionised SER, leading to state-of-the-art results.

In SER, simple and widely used architectures include Support Vector Machines (SVM) (Song et al., 2005) and Artificial Neural Networks (ANN) (Grossi & Buscema, 2008). SVM is a supervised learning algorithm that finds optimal hyperplanes for classification, while ANN consists of a layered architecture. As the data propagates through the layer, it is processed and filtered. Selvaraj, Bhuvana, and Padmaja (2016) conducted research using SVM and ANNs. Their strategy consisted of first framing the audio, extracting features like Mel Frequency Cepstral Coefficients (MFCC) and pitch. MFCC is a widely used feature extraction technique in speech processing and recognition. It represents the short-term power spectrum of audio signals, converting the frequency domain information into a compact, perceptually relevant representation (Hossan et al., 2011), which is particularly useful for tasks like Speech Emotion Recognition (Slaney 2000). Using SVM for classification, the audio could be classified into classes like gender based on characteristics like pitch. The results of SVM are used as input for an ANN. Two types of ANN were tested, i.e., with Radial Basis Function (RBF) activation function and a simple Back Propagation Network (BPN). The results were approximately 80% and 90% accuracy for RBF and BPN respectively. The figures are definitely impressive, but it needs to be noted that the reason for these figures is primarily two. First, the study classifies the speech samples only in 4 classes and second, it uses a consistent 200-sample dataset for testing, extracting results and building models. The model may have learned to give good accuracy on this data, but totally unseen data would be surely difficult for the model to even classify in the meagre set of 4 emotions. In other words, just relying on accuracy to judge a model's strength is not an appropriate way.

(Aouani & Ayed, 2020) further explored innovative approaches to improve speech emotion classification accuracy using SVM. One notable addition to the traditional feature set was the usage of Harmonic to Noise Ratio (Fernandes et al., 2018) in conjunction with Teager Energy Operator, MFCC, and Zero Crossing Rate (Jalil et al., n.d.). This expanded feature set aimed to capture more comprehensive acoustic information from the speech data, potentially enhancing emotion recognition performance. To further optimise the input data, the researchers employed an auto-encoder. The auto-encoder's purpose was to minimise the dimensions of the input vector while retaining relevant information. By reducing the data complexity, the model could focus on the most discriminative features for emotion classification, and at the same time could improve the performance. The researchers have evaluated the emotion recognition system using Support Vector

Machines (SVM) as the classifier. They tested the system on the RML database (Dalvi et al., 2021), which contained a diverse range of emotional speech samples. Experimenting with the auto-encoder, the researchers found that a Stacked Artificial Neural Network (ANN) with three hidden layers as an encoder yielded the best accuracy. This stacked architecture was capable of capturing intricate patterns in the data, leading to an emotion recognition accuracy of 74%.

Another research by (Deshmukh et al., 2019), has classified speech emotions into only 3 classes i.e., Happy, Angry or Sad. To ensure accurate feature extraction from the analogue audio signal, they applied several essential signal processing techniques, including Sampling, Pre-emphasis, De-silencing, Framing, and the application of a Hamming Window. These techniques collectively prepared the raw audio data for further analysis. The researchers had an extensive array of features including Energy, MFCC (Hossan et al., 2011), Fast Fourier Transform (FFT), Power Spectral Density (PSD), and the MEL filter bank coefficients (Kopparapu & Laxminarayana, 2010). However, after thorough experimentation, they narrowed down the features to the most relevant ones: Short Term Energy (STE), MFCC, and Pitch. To make the emotion classification process more efficient, the researchers had multiple options to choose from when summarising the feature values for each frame. They could use Mean or Mode, and after extensive evaluation, they found that computing the Mean yielded the best results. Intriguingly, the impact of including STE in the feature set varied for each emotion. For every emotional category, STE contributed to an accuracy boost of 15-20%. The researchers' comprehensive approach, combining essential signal processing techniques with feature selection and summarization, allowed them to achieve notable results in emotion classification. Their efforts contributed to an overall accuracy of 61% using a combination of MFCC and Pitch features for emotion classification. The addition of STE led to an increase in accuracy, raising it to 79%. But that also involves classifying into 3 classes.

(Kerkeni et al., 2022) followed a systematic four-step approach in their speech-emotion recognition study. Sample collection was done and then proceeded with feature extraction. The features they were able to extract included Linear Predictive Coefficients (LPC), MFCC and Modulation Spectral Features (MSFs) generated through Spectro-temporal processing. For feature selection, they employed Recursive Feature Elimination (RFE) to streamline the dataset and enhance model efficiency. In the classification stage, they evaluated the performance of three algorithms: Multivariate Linear Regression (MLR), Support Vector Machines (SVM), and Recurrent Neural Networks (RNNs), with and without Speaker Normalisation (SN). The dataset used for training and testing consisted of speech samples from two languages: Berlin and Spanish. Despite RNNs demonstrating the highest accuracy in one of the aspects of emotion classification, their long training time posed a challenge for real-world applications. In contrast, SVM and MLR, while slightly less effective in that particular aspect, offered consistent results and faster training times, making them more practical choices for emotion recognition tasks.

Using a Chinese language dataset, the researchers (Huang et al., 2014) explored the application of Deep Belief Networks (DBNs) in SER. DBNs are a type of deep learning architecture, which is a stacked form of Restricted Boltzmann Machines (RBMs) (Montufar, 2018). In DBNs, each hidden layer is connected to the visible layer and vice versa, forming a hierarchical structure. However, no hidden layer is directly connected to another hidden layer, and the same holds for visible layers. The researchers chose DBNs as their primary model for emotion classification due to their ability to

extract emotional cues layer-by-layer, instead of just classifying directly. To evaluate the effectiveness of DBNs, the researchers performed a comparison with the Support Vector Machine (SVM) classification technique. The DBNs achieved an accuracy of approximately 86 in classifying emotions in the Chinese language dataset. In contrast, the SVM technique achieved an accuracy of 79%, which while respectable, was outperformed by the DBNs.

With the context of clustering algorithms used in machine learning, K-Nearest Neighbors (k-NN) is a popular and straightforward algorithm used in various fields, including Speech Emotion Recognition (SER). It relies on finding the most similar data points in the training set to classify new inputs based on their proximity. k-NN can be applied to classify audio through its emotional characteristics in speech data. (Singh et al., 2023) conducted research focused on the Berlin dataset, employing clustering. Multiple classifiers, including Medium, Cubic, Cosine, Fine, and Weighted, were tested for evaluation. The results indicated that the Fine and Weighted classifiers outperformed the others, achieving accuracy rates of 88% and 86%, respectively. (Venkata Subbarao et al., 2022) utilised two databases, Berlin and SAVEE, for the study. Their extracted features included Base Frequency, Energy, and Zero Crossing Rate (FEZ) along with Fourier Parameters (FP). Support Vector Machine (SVM) and k-nearest Neighbours (KNN) classifiers were both tested. The results demonstrated an accuracy of 87% for the German (Berlin) dataset and an even higher accuracy of 90% for the English dataset using KNN.

With the advancement in the field of machine learning various complex models like RNN and its variations were introduced. RNNs are useful where the order of data matters when drawing conclusions, like Speech Recognition. So, researchers employed RNNs in the SER, as it is also the sub-branch of speech recognition. (Tzinis & Potamianos, 2017) use of the IEMOCAP (Busso et al., 2008) database for emotion classification. They employed a model with two LSTM layers (512 and 256 units) followed by a SoftMax layer. The study investigated the impact of varying chunk lengths for calculating Low-Level Descriptors (LLDs) and changing segment lengths on classification accuracy. The best results were obtained with an LLD chunk length of 0.09 seconds, achieving a weighted accuracy of 59% and an unweighted accuracy of 54%. For segment lengths, the highest performance was achieved with 3-second segments, resulting in a weighted accuracy of 64% and an unweighted accuracy of 60%.

Another study on the application of Recurrent Neural Networks using the IEMOCAP dataset is presented by (Mirsamadi et al., 2017). The researchers explored different configurations for emotion classification. They experimented with several architectures, including two dense layers followed by a SoftMax layer, a dense layer followed by a recurrent layer and then a SoftMax layer (without pooling, with mean pooling, and with weighted pooling using an attention model). Among the various configurations tested, the best performance was achieved with an architecture that utilised a Recurrent Neural Network (RNN) with weighted pooling and an attention model. This model yielded a weighted accuracy of 63.5% and an unweighted accuracy of 58.8%. Here it is to be noted that this low accuracy is due to the fact that the models used are too simple and they just rely on LSTM layers, whose primary purpose is to process sequential data. If some complex model was used with some mixed configuration of LSTMs, ANNs and SVMs, it would have provided better results. Such a technique is used by (Singh et al., 2023) and has resulted in better results. It will be discussed later in the same chapter.

The most effective way of SER is generating spectrums from the audio and using image classification on the spectrums to conclude. The extremely popular architecture for image classification is Convolutional Neural Networks (CNNs). It is a deep learning architecture designed specifically to process grid-like data such as images and spectrograms. CNNs are particularly effective in feature extraction from raw data by utilising convolutional layers to detect patterns and spatial relationships, followed by pooling layers to reduce dimensionality. In the context of Speech Emotion Recognition, CNNs have been widely used for their ability to automatically learn relevant acoustic features from audio signals, contributing to promising results in emotion classification tasks. (Singh et al., 2023) used the architecture that contains a series of Conv2d layers, and 2 LSTM layers with attention mechanisms. The model was tested and trained on the combination of RAVDESS, SAVEE (Haq et al., 2008) and TESS dataset. An extensive feature selection was done, by using combinations of features manually. The accuracy achieved was about 90%.

The researchers (Lim et al., 2016) propose three models using the dataset of RAVDESS and TESS. The first model combines Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks, followed by flattening and classification, achieving a precision rate of 87%. The second model comprises three layers of LSTM units with 128 units each, followed by a SoftMax Layer of classification, resulting in a precision rate of 79%. The third and most successful model, termed "Time Distributed CNNs," involves 2D Convolutional layers, a Max Pooling layer, an LSTM layer, and a SoftMax layer, yielding an impressive precision rate of 88%. (Zheng et al., 2018) conducted research utilising the Chinese Database. The researchers employed a feature extraction process followed by a Convolutional Neural Network (CNN) architecture. The extracted features were then passed through a Flatten layer before being classified using a Random Forest (RF) Classifier. The results of this approach demonstrated an average accuracy of 75%. However, the happy and sad accuracy was limited to approximately 60%, whereas other emotions have quite high accuracy.

With the advent and increased popularity of transfer learning, pre-trained CNNs and also some state-of-the-art models like RESNETS (pawangfg, 2023), Transformer (Vaswani et al., 2023), VGGNETs (Simonyan & Zisserman, 2015) have been used extensively in applications like SER, as such models have amassed vast knowledge from a diverse dataset. This transfer of knowledge enables the model to grasp essential features and nuances related to emotions, especially when dealing with speech data.

(Togootogtokh & Klasen, 2021) introduced the DeepEMO framework. This approach comprised two vital pipelines to achieve remarkable results. The first pipeline involved employing advanced signal processing techniques to extract potent speech features. By subjecting the raw speech signal to Fast Fourier Transform (FFT), they transformed it into the frequency domain and generated a spectrum. Further mathematical computations on the spectrum yielded the Mel spectrogram, a representation known for enhancing human perception of audio data and facilitating emotion recognition. In the second pipeline, leveraging the power of pre-trained models, they opted for the ResNet18 (Anon., 2018) architecture. The model attained an outstanding training and validation accuracy of 1.0, indicating its ability to precisely predict emotions for both datasets. Additionally, the reported low loss value of 0.006 showcased the model's capacity to make highly accurate emotion predictions with minimal errors.

Yet another team-led research (Lech et al., 2020) using the Alex Net (Krizhevsky, 2014) pre-trained model, which essentially is a Convolutional Neural Network (CNN). It was fine-tuned for the task of audio signal processing and classification. The study experimented with four different frequency scales for the spectrum (linear, log, MEL, and ERB) and tested multiple sampling frequencies and bandwidths. Through careful fine-tuning and hyperparameter exploration, the model achieved a weighted accuracy of approximately 80%.

Another research (S, N. R. et al., 2018) proposed the use of the Inception NET v3 model (Szegedy et al., 2015) for this purpose, which basically is a convolutional model. They trained the model on the dataset IEMOCAP (Busso et al., 2008). The results were a loss of 0.71 on training data and 0.95 on validation data, whereas only a low accuracy of 35.6% was attainable. The researchers argue that the accuracy was low because of the loss of data during transfer learning.

(An & Ruan, 2021) utilised the RAVDESS database, which comprised a substantial collection of 7356 recordings. To increase the dataset's size and enhance its variability, Additive White Gaussian Noise was applied to the original recordings. For feature extraction, the researchers employed Mel Frequency Cepstral Coefficients (MFCC). To build an effective model, the researchers designed a novel architecture involving parallel Convolutional Neural Networks (CNNs) and a Transformer layer. The CNNs excelled at capturing spatial features from the speech data, while the Transformer layer, known for its strength in sequential data processing, focused on capturing temporal patterns.

After extracting distinctive features from both CNNs and the Transformer layer, the researchers merged the results in a dense layer using a SoftMax activation function. This final layer facilitated the classification of emotions based on the combined features extracted from the parallel models. The outcomes of their experiments revealed promising results. The proposed architecture achieved an accuracy of 80% in speech emotion recognition, demonstrating its effectiveness in accurately identifying emotions from the RAVDESS dataset (Livingstone & Russo, 2018). In comparison, a traditional Deep Convolutional Neural Network (DCNN) achieved an accuracy of 72%.

The above research demonstrates the use of various machine learning models, including traditional algorithms like KNN, SVM, and Random Forest, as well as advanced deep learning models like CNN, RNN, and Transformers. It has been observed that advanced deep learning techniques outperform traditional methods in Speech Emotion Recognition (SER). This motivates us to further explore deep learning techniques for SER. The study in reference (An & Ruan, 2021) encourages us to adopt CNN for SER, and our intuition is to utilise multiple advanced pre-trained CNNs instead of simple 2-layered models. Pretrained models like Resnets (Anon., 2018), and Inception Net (Szegedy et al., 2015) on the ImageNet dataset have already shown their significance in image classification, leading us to believe that fine-tuning these pre-trained models could yield better results in SER.

Chapter 3: Methodology

In this chapter, we dive into the essential components of our research journey. We'll begin by discussing the tools and resources we utilised in our research environment, giving you insights into the programming language and libraries that powered our model development. We'll also provide a clear and detailed account of how we crafted and trained our models to achieve the remarkable results we'll present. Additionally, we'll shed light on our dataset selection process, offering transparency about data acquisition and preparation.

This study has received ethical approval, and the ethical approval letter is included in the Appendix Material.

3.1. Research Framework and Tools

This section presents an in-depth overview of the research framework and the suite of tools and technologies harnessed for the execution of the study focusing on Speech Emotion Recognition (SER) using Convolutional Neural Networks (CNN). The chosen tools and technologies were instrumental in ensuring the successful implementation of the proposed methodology, encompassing audio processing, data management, visualisation, and the development of the CNN-based model.

3.1.1. Research Environment

The research environment was meticulously established on the Google Colab platform. Google Colab provided access to a dynamic hardware configuration comprising an Intel(R) Xeon(R) CPU @ 2.20GHz, a Tesla T4 GPU and 12GBs of RAM. (Google, n.d.)The integration of these processing resources endowed the research with the computational capacity required for intricate tasks involved in CNN-based SER model development, training, and evaluation.

3.1.2. Programming Language

Python 3.10.12 was adopted as the core programming language for this research endeavour. Renowned for its versatility and a vast array of libraries (Python, 2023), Python offered an ideal foundation for implementing diverse components of the SER methodology.

3.1.3. Libraries and Packages

The study employed a selection of libraries and packages, each tailored to specific phases of the research process:

Librosa (v0.10.1): Librosa, a preeminent Python package for audio analysis, played a pivotal role in the audio processing stage. Leveraging Librosa, the research extracted pertinent features from audio files, such as Mel-Frequency Cepstral Coefficients (MFCCs), spectrograms, and chroma features (McFee et al., 2015).

Scikit-learn (v1.2.2): The versatile capabilities of Scikit-learn were harnessed for intermediary tasks such as dataset partitioning into training and testing subsets. Additionally, Scikit-learn facilitated categorical and one-hot encoding of emotion labels, vital steps for preparing the data for CNN model training (Pedregosa et al., 2011).

NumPy (v1.23.5) and Pandas (v1.5.3): NumPy (Harris et al., 2020) and Pandas (McKinney, 2010) formed the cornerstone of data manipulation, shaping, and management. These libraries enabled the efficient organisation and preprocessing of the dataset, transforming it into a suitable format for consumption by the CNN model.

Matplotlib (v3.7.1) and Seaborn (v0.12.2): For insightful data visualisation, Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021) were harnessed. These visualisation libraries facilitated the creation of intuitive graphs, plots, and visual summaries of the dataset's distribution, feature extraction outcomes, and model performance.

TensorFlow (v2.12.0) and Keras (packaged with the same version): TensorFlow (Abadi et al., 2015)and its high-level interface, Keras, formed the bedrock of the CNN model development. These frameworks facilitated the design and construction of the neural network architecture, training configurations, and performance evaluation.

3.2. Overview of the Dataset

This section presents a comprehensive overview of the dataset utilised for the training and testing phases of the CNN model designed for SER The dataset amalgamates four well-known and widely used collections, namely RAVDESS, SAVEE, TESS, and CREMA-D, each contributing diverse emotional expressions that enrich the scope of analysis. These datasets are readily available on Kaggle at https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en. (BABKO, 2020)

3.2.1. Composition of the Dataset

The training and testing data included samples from the following four prominent datasets. All these datasets are specifically made for the purpose of SER and therefore do not need any cleaning or consistency management. These include:

Table 1: Dataset distribution

Dataset	No. of Audio Samples
Crema	7442
Tess	2800
Ravdess	1440
Savee	480
Total	**12,162**

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song): This dataset comprises a total of 1440 samples, evenly split between male and female recordings. The samples are recorded by 24 professional actors with North American English accents. The dataset encompasses a rich emotional spectrum, featuring recordings manifesting anger, disgust, fear, happiness, neutrality, sadness, and surprise. Notably, the neutral emotion category contains a slightly larger share of 288 samples, offering a comprehensive view of baseline vocal expressions.

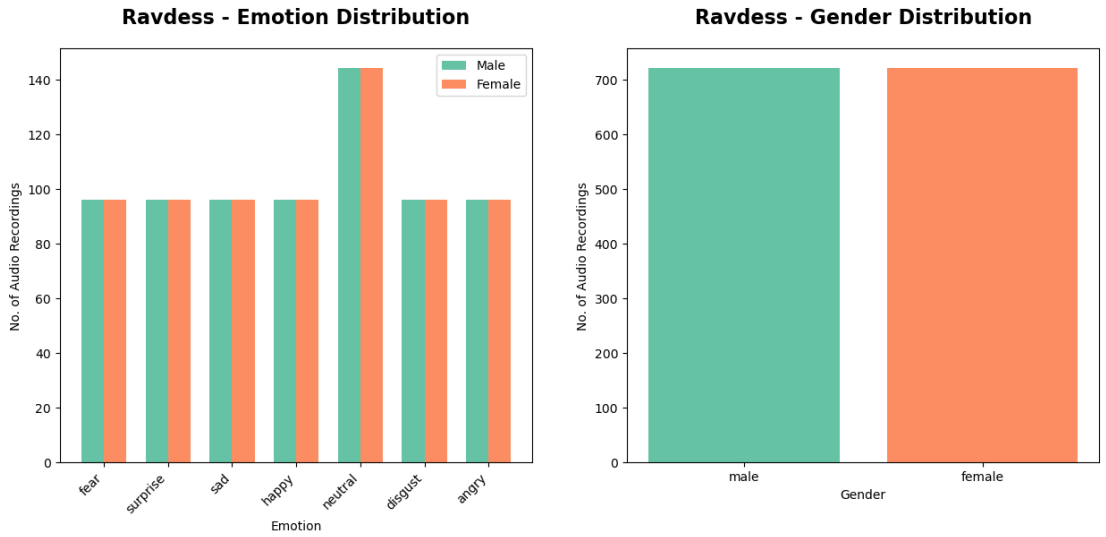


Figure 1: Ravdess Distribution

SAVEE (Surrey Audio-Visual Expressed Emotion): With a total of 480 samples, the SAVEE dataset exclusively features male speakers. Among the emotional categories, each emotion is represented by 60 samples. This dataset serves as a valuable addition, particularly in the context of male vocal expressions, adding a unique dimension to the overall emotional diversity.

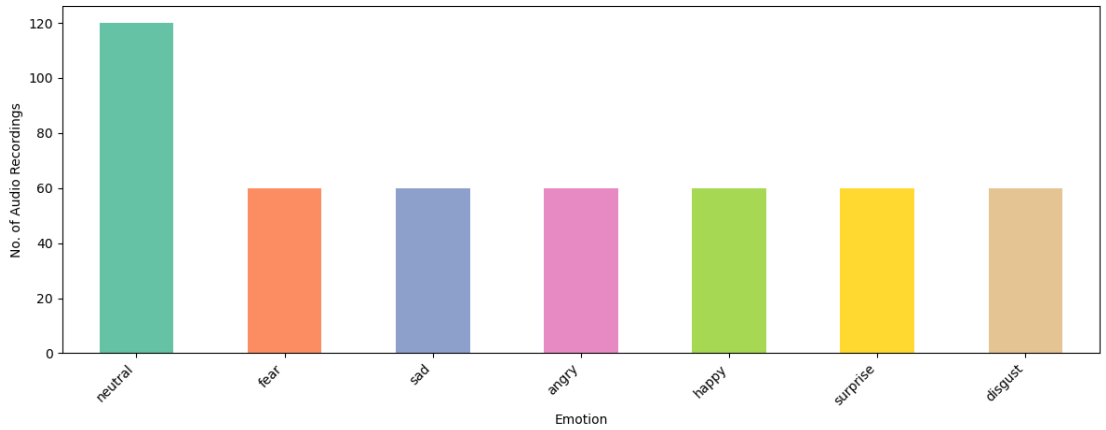


Figure 2: Savee Distribution

TESS (Toronto Emotional Speech Set): Comprising 2800 samples solely from female speakers, the TESS dataset brings a distinct feminine vocal perspective to the dataset. It encompasses an equal distribution of 400 samples per emotional category, ensuring a well-balanced representation of anger, disgust, fear, happiness, neutrality, sadness, and surprise.

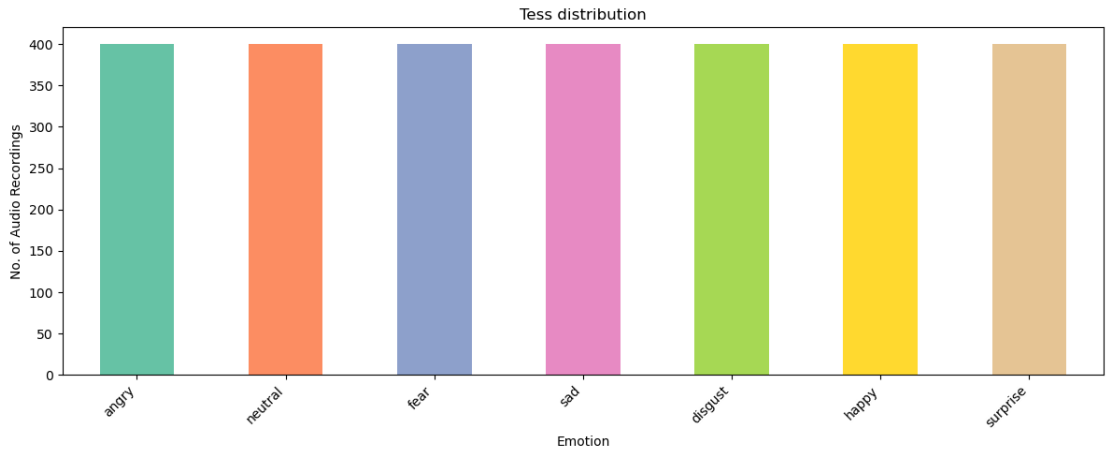


Figure 3: Tess Distribution

CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset): The CREMA-D dataset boasts a substantial volume of 7442 samples, encompassing a diverse demographic of 3512 female and 3930 male speakers. This dataset provides a comprehensive collection of emotional expressions, featuring a substantial share of samples across the anger, disgust, fear, happiness, neutrality, and sadness categories.

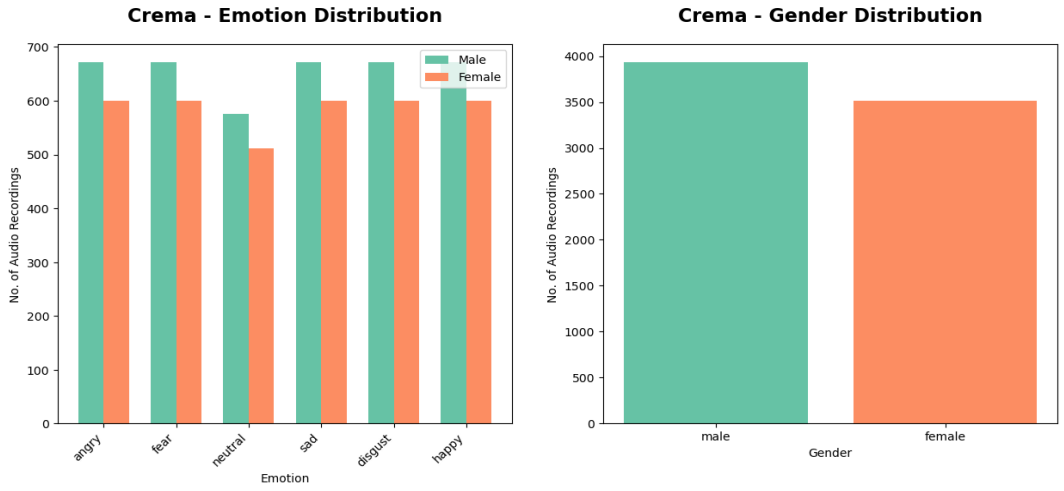


Figure 4: Crema Distribution

3.2.2. Total Emotion Distribution Visualisation

To provide an encompassing visualisation of the overall dataset distribution, the bar chart in Figure 5 showcases the cumulative distribution of emotional categories across all four datasets.

The collective dataset composition and the distribution of emotional expressions within each dataset provide the groundwork for the subsequent phases of the SER methodology. The diversity and volume of data gathered from these datasets establish a solid foundation for training and evaluating the CNN model's aptitude in recognizing emotions from speech.

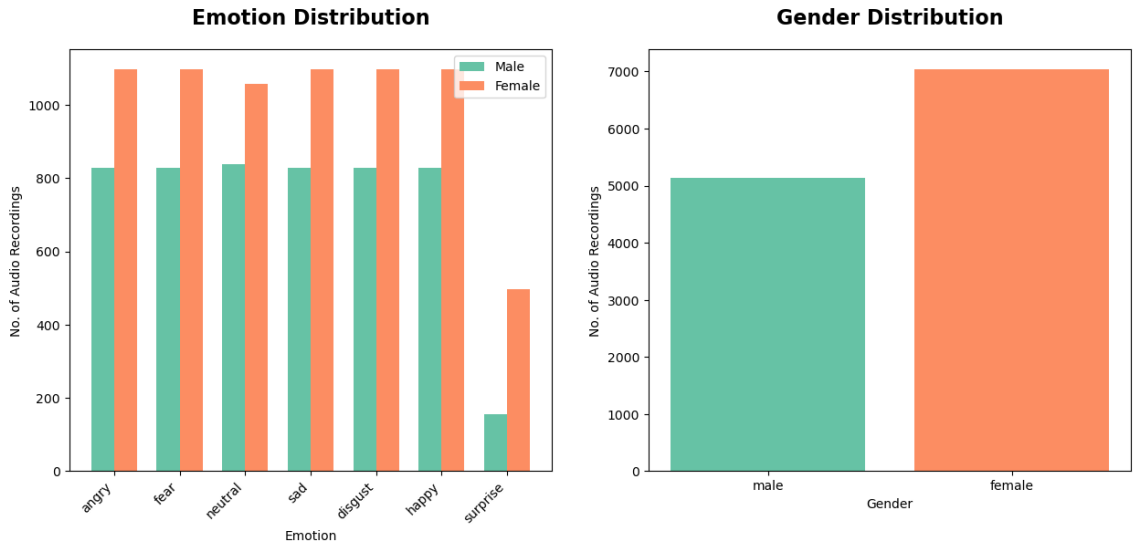


Figure 5: Total Data Distribution

3.3. Data Augmentation

3.3.1. The Need for Data Augmentation

Data augmentation is an essential technique in the realm of machine learning and deep learning, particularly when working with limited datasets. Its significance becomes particularly pronounced when dealing with intricate tasks such as SER. Augmentation enhances the diversity and richness of the dataset by introducing variations into the existing samples. This process aids the model in generalising better to different scenarios and variations in the real world, ultimately improving its performance and robustness. In the context of audio data like speech, augmentation techniques play a crucial role in mimicking real-world conditions and challenges.

3.3.2. Augmentation Techniques

To inject diversity into the audio data and improve the model's resilience, a variety of augmentation techniques were applied to the dataset. We tried implementing augmentations one by one but as we increased the number of augmentations the results became better. So, we applied all the techniques to the data. We describe the augmentation techniques employed below. Figure 6 contains a mel-spectrogram and waveform of audio for demonstrating and explaining the difference the augmentation makes. We will compare this with the graphs of augmented audio.

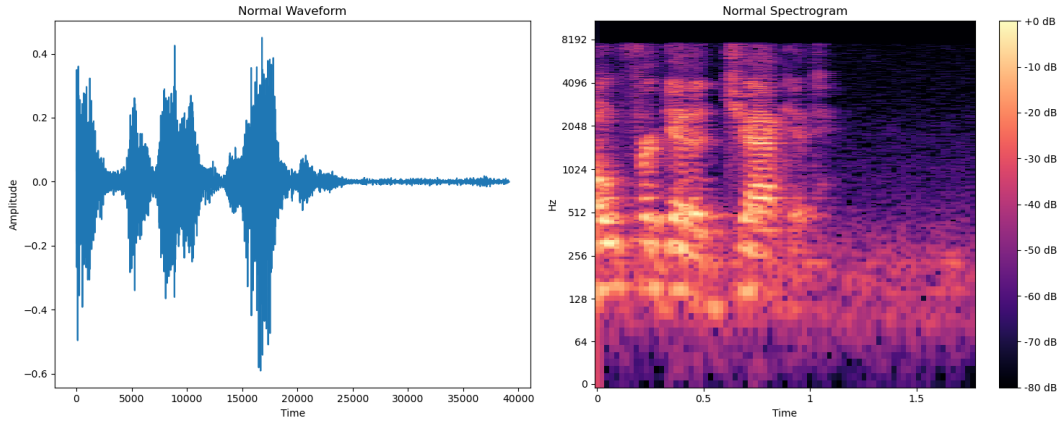


Figure 6: Waveform and Mel Spectrogram of a normal audio

Noise Augmentation: The noise augmentation technique (Eklund, 2019) introduces random ambient noise to the audio samples, simulating the presence of background noise that often occurs during speech communication. By adding a controlled amount of noise to the audio data, the model learns to recognize emotional cues even in the presence of varying noise levels. This augmentation serves to enhance the model's robustness to acoustic challenges. We can generate noise by calculating a random amplitude of noise based on the maximum value in the input data and a scaling factor. Then we generate random noise samples from a normal distribution and add this noise to the input audio data.

```
function noise(data):
    noise_amplitude = 0.04 * random_uniform() * max(data)
    noise_array = generate_normal_noise(size(data))
    noisy_data = data + (noise_amplitude * noise_array)
    return noisy_data
```

In the waveform and spectrogram from Figure 7, we can see that the noise has been added to the normal audio in Figure 6. The orange dots here and there in the spectrogram and the roughness in the curve of the waveform represent noise. The normal audio in Figure 6 was simple and contained a large black-and-blue area, demonstrating a clear voice signal. But in Figure 7, there is a lot of noise added.

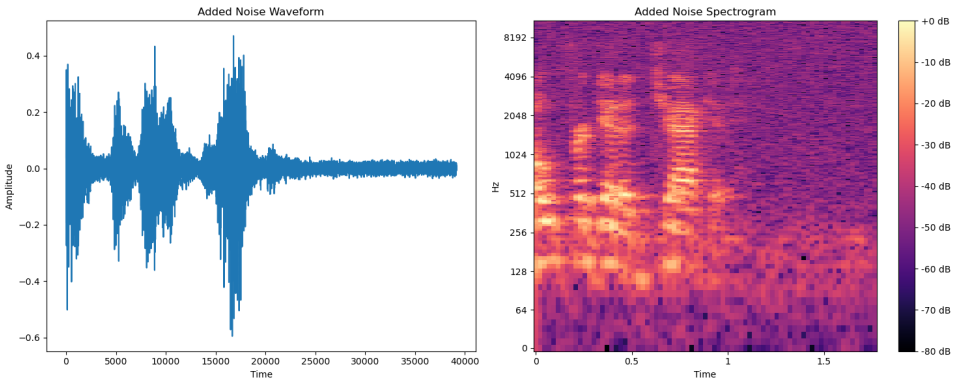


Figure 7: Waveform and Mel Spectrogram of a noise-added audio

Time Stretch Augmentation: Time stretching involves altering the temporal duration of the audio while preserving its pitch. This augmentation simulates variations in speaking rates and helps the model adapt to different pacing and timing patterns in speech. It aids in enhancing the model's sensitivity to the rhythm and tempo of spoken emotions. For this purpose, librosa provides a function `librosa.effects.time_stretch(...)`.

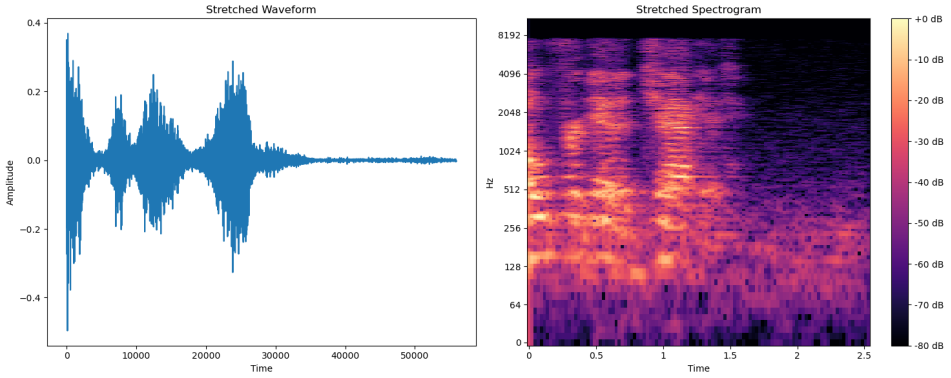


Figure 8: Waveform and Mel Spectrogram of a time-stretched audio

In Figure 8, the audio has been stretched. The normal audio in Figure 6 had only a time frame of 2s, whereas the stretched audio had a time frame of about 2.5 seconds.

Shifting Augmentation: The shifting augmentation technique displaces the audio waveform by a certain time duration, akin to slight temporal misalignments that can occur during communication. By introducing controlled shifts, the model learns to accommodate variations in speech delivery and better decipher emotional cues. We can implement it using the NumPy function `np.roll(...)`.

Figure 9 represents an audio signal with some of the data from the start of Figure 6 being shifted to the end in the waveform and spectrogram.

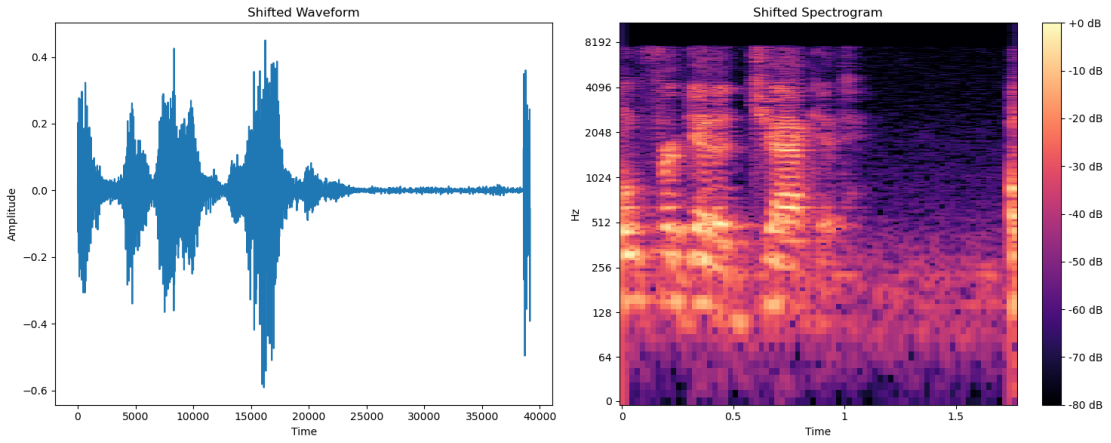


Figure 9: Waveform and Mel Spectrogram of a time-shifted audio

Pitch Augmentation: Pitch augmentation involves altering the pitch of the audio without affecting its temporal characteristics. This technique reflects instances where individuals might exhibit varied vocal pitches while expressing emotions. By training on augmented samples with adjusted pitch levels, the

model becomes more adept at capturing emotional content across diverse vocal pitches. Librosa assists us to implement this using `librosa.effects.pitch_shift(…)`.

The higher pitch is represented by more ups and downs or in other words more density in the spectrogram and waveforms where there were already some audio waves. This can be demonstrated in Figure 10.

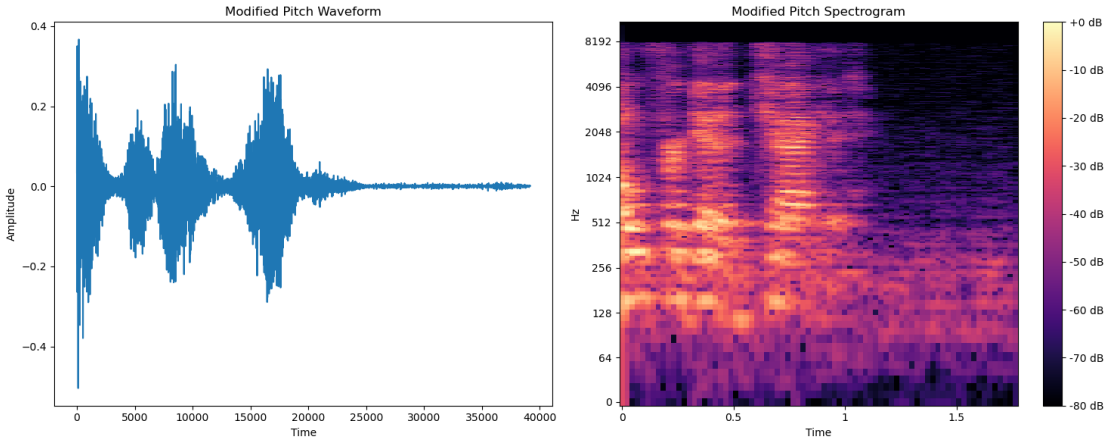


Figure 10: Waveform and Mel Spectrogram of a modified pitch audio

Speed Augmentation: Speed augmentation manipulates the temporal duration of the audio samples, affecting both pitch and speed. This technique is akin to variations in speech rate that occur naturally. By training on samples with different speaking speeds, the model acquires the ability to decipher emotions effectively, regardless of varying speaking rates. For slowing we interpolate the values between the arrays, whereas for fastening we can drop some values while keeping the sampling rate the same. We can implement this using the following algorithms:

```
function change_speed(data, factor):
        # data: Input audio data (e.g., a NumPy array)
        # factor: Speed change factor (e.g., 0.5 for half speed)
        output_length = factor x length(data)
        new_audio = create_empty_array(output_length)
        if (factor < 1):
                for each output_idx in new_audio and input_idx in data:
                        new_audio[output_idx] = interpolate_sample(data,input_idx)
        else:
                for each output_idx in new_audio and input_idx in data:
                        new_audio[output_index] = audio_data[input_index]
        return new_audio
```

We can see in the waveforms and spectrograms in Figure 10 and Figure 11 that the speed of the audio has changed. The same audio now plays faster in one case and slower in other cases compared to the normal audio in Figure 6.

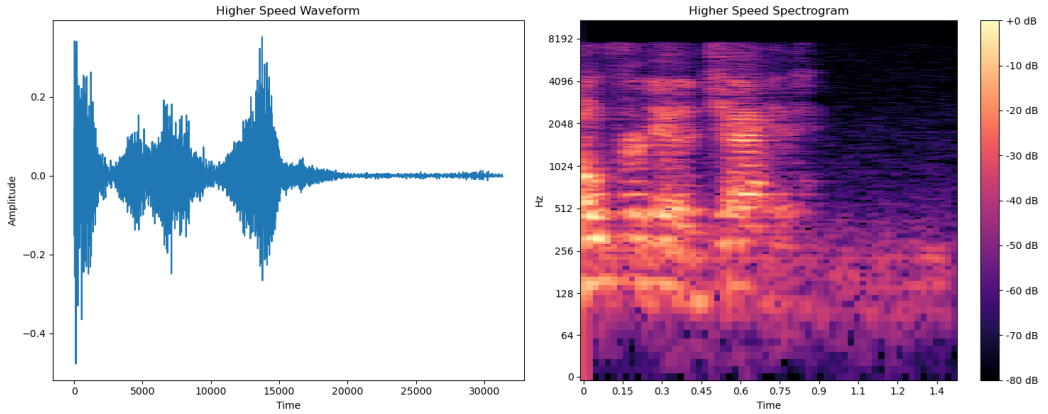


Figure 11: Waveform and Mel Spectrogram of a fastened audio

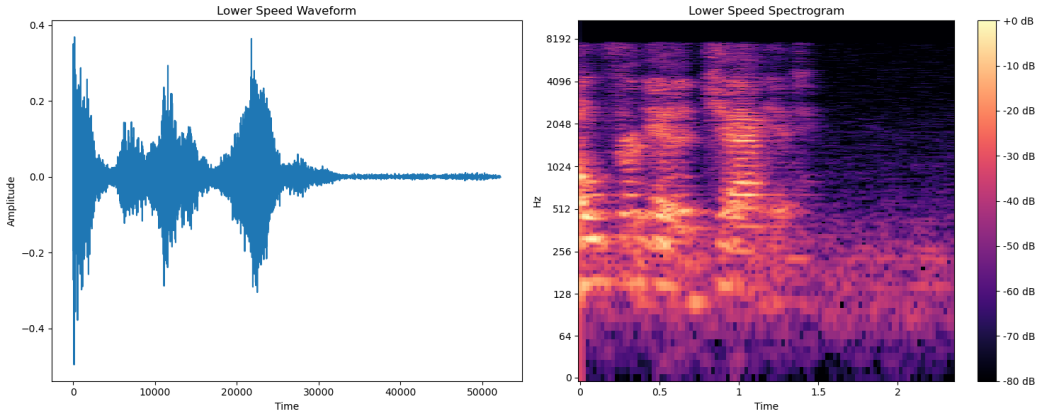


Figure 12: Waveform and Mel Spectrogram of a slowed audio

The integration of these augmentation techniques enriches the dataset by introducing variations in acoustic conditions, pacing, pitch, and speed. This augmentation process equips the CNN-based SER model to robustly recognize emotional expressions in diverse real-world scenarios.

3.4. Feature Extraction

3.4.1. Introduction to Feature Extraction

Feature extraction is a pivotal step in Speech Emotion Recognition (SER), where the raw audio signals are transformed into a compact representation that captures essential information for emotion analysis. These extracted features serve as the input to the Convolutional Neural Network (CNN), enabling the model to decipher emotional cues effectively. This section delves into the extraction of five key features: Mel-Spectrogram, Mel-Frequency Cepstral Coefficients (MFCC), Teager Energy, Zero Crossing Rate, and Short-Time Energy (STE).

3.4.2. Feature Extraction Techniques

Mel-Spectrogram: The Mel-Spectrogram is a visual representation of the spectral content of an audio signal over time. It effectively encapsulated the distribution of frequency components within the audio, highlighting patterns and shifts that were indicative of different emotions. To extract the

Mel-Spectrogram, the raw audio signal was first divided into short overlapping frames. For each frame, the Fast Fourier Transform (FFT) was computed, and the power spectrum was transformed using a Mel filter bank. The resulting Mel-Spectrogram captured frequency variations over time, forming a two-dimensional representation that can be a foundational feature for emotion recognition.

Figure 13 shows a Mel-Spectrogram. The brighter portion of the spectrogram indicates the presence of frequency at higher amplitude. For example, in Figure 13, between 0.5 and 0.7 seconds intervals there are higher amplitudes of 512 to 4096 Hz, whereas there are lesser amplitudes of 512 to 4096 Hz between time intervals of 0.7 to 1.0 seconds.

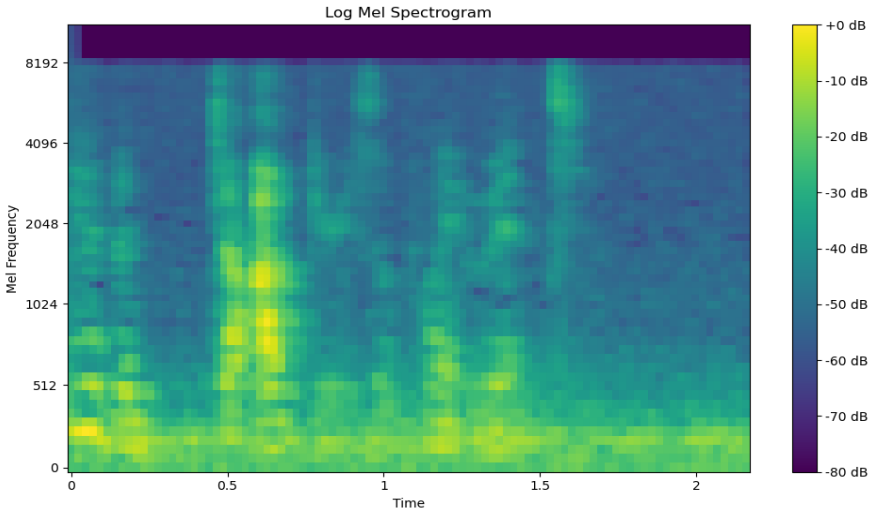


Figure 13: Example of a Mel-Spectrogram

We can generate a mel-spectrogram using the librosa library. It provides a function `librosa.feature.melspectrogram(...)` which generates mel spectrograms for us.

Mel-Frequency Cepstral Coefficients (MFCC): MFCCs represent the spectral characteristics of the audio signal, mimicking the human auditory system's sensitivity to different frequency bands. To extract MFCCs, the Mel-Spectrogram was further processed to derive the logarithm of the power spectrum, followed by the application of the Discrete Cosine Transform (DCT). This resulted in a compact set of coefficients that encapsulated critical spectral information, making it highly suitable for feeding into the CNN model.

Just like mel-spectrogram, librosa library can be used to generate the MFCC, too. It provides a function `librosa.feature.mfcc(...)` which calculates MFCCs.

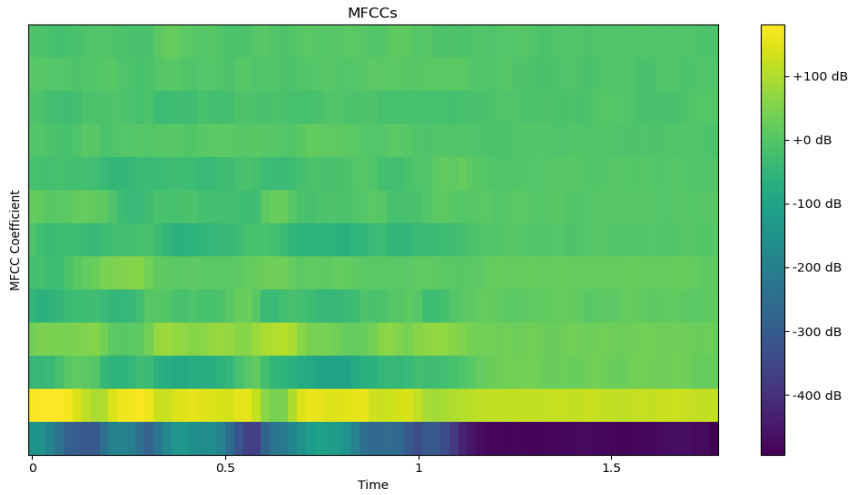


Figure 14: Representation of a MFCC constants array

Teager Energy: The Teager Energy operator is utilised to capture nonlinear dynamics in the audio signal. It focuses on detecting energy fluctuations within the waveform, often linked to abrupt changes that occur during speech modulation. This feature is particularly valuable for recognizing emotional transitions and abrupt shifts in vocal expression.

Zero Crossing Rate: The Zero Crossing Rate (ZCR) quantifies the frequency of zero crossings within the audio waveform. It is a measure of the audio's rapid changes in sign, which often correlate with abrupt changes and the presence of high-frequency components. This feature aids in discerning emotions with dynamic and varying intensity. We can calculate the ZCR per sample for each audio recording we have. The resulting array could be represented like the figure below. Librosa helps us generate ZCR. It provides a function `librosa.feature.zero_crossing_rate(..)` to produce an array of zero-crossing rates.

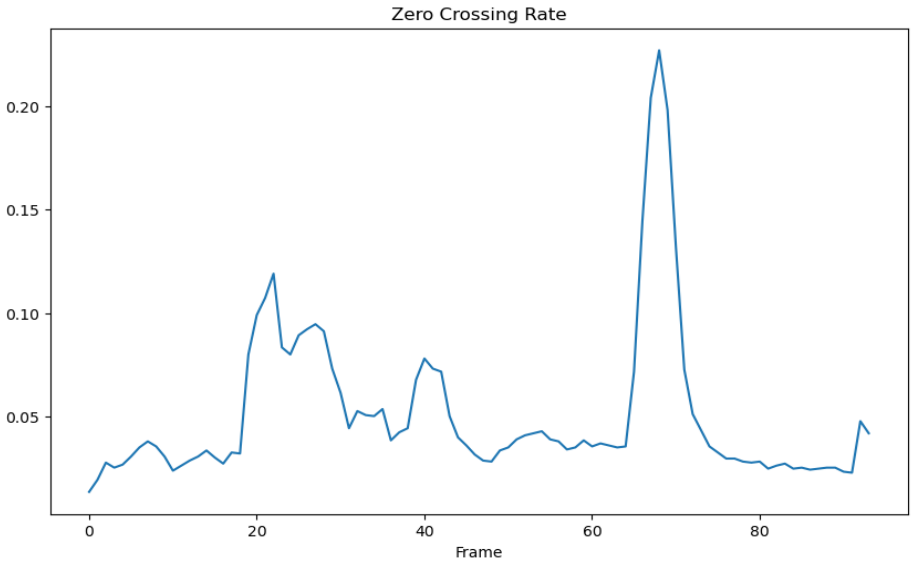


Figure 15: Example of a Zero-Crossing rate

Short-Time Energy (STE): Short-Time Energy characterises the signal's energy content within small time frames. It is computed by summing the squared values of the audio samples within each frame. STE is instrumental in capturing variations in vocal intensity, which can be indicative of emotional expression. It can be calculated as:

$$STE[n] = \frac{1}{N}\left(\sum_{k=0}^{N-1} X[n - k]^2\right)$$

Where n represents the frame number in sample audio whose STE is being calculated, N represents the size of the frame under consideration, and X contains the data of the frame under consideration.

By extracting and utilising these features, the SER model gains the ability to interpret the intricate emotional signatures encoded within the audio data. The ensuing section will detail the integration of these features into the CNN model and the subsequent stages of the methodology.

3.5. Feature Selection

In the pursuit of optimising Speech Emotion Recognition (SER) performance, a comprehensive exploration of various features was conducted. These features encompassed a spectrum of audio representations, each designed to capture different aspects of emotional expressions in speech. Among these features, Mel-Frequency Cepstral Coefficients (MFCC) and Logarithmic Mel Spectrogram emerged as the most potent candidates, consistently delivering the most optimal results across a range of model architectures.

The iterative experimentation included the evaluation of several other features, each presenting its unique characteristics for encoding audio data. However, the refined selection process highlighted that MFCC and Logarithmic Mel Spectrogram not only effectively encapsulated emotional nuances but also facilitated more accurate classification through the employed model architectures. This selective utilisation of features in the model training phase was instrumental in enhancing the precision of emotion recognition from speech data.

To provide a clearer understanding of our feature selection process, we present a summary of the results in Table 2. The method used for calculating these accuracies will be discussed in Section 3.8.2.

Table 2: Accuracy achieved using different features

Feature	Maximum Accuracy Achieved
MFCC	95%
Logarithmic Mel Spectrogram	97%
Teager Energy	88%
Zero Crossing Rate	82%
Short-Time Energy	78%

3.6. Dataset Split

We split the dataset into 2 portions randomly. 80% of the data was reserved for training models and the remaining 20% was reserved for testing the model. This 20% data was the one upon which we can test how the model would behave on the data it has never seen.

3.7. Model Architecture and Training

In this section, we will discuss in detail the four model architectures that were utilised for Speech Emotion Recognition (SER): MFCC InceptionNet, ResNet MFCC, Mel InceptionNet, and Mel ResNet. Additionally, we will delve into the loss function and optimizer employed for training these models. All the testing was done on a training set of size 68107 and test data of size 17027.

3.7.1. Model Architectures

MFCC InceptionNet: The MFCC InceptionNet model leverages the Inception architecture (Shaikh, 2018) adapted for 1D convolutional operations as in Figure 16. The input layer receives a 1D signal representing MFCC data with a shape of (58, 1). It then undergoes initial convolution, batch normalisation, and ReLU activation. Subsequent layers consist of inception blocks, each combining multiple convolutional operations of varying kernel sizes and max pooling. These blocks facilitate feature extraction at multiple scales. The model concludes with global average pooling and dense layers, eventually producing an output layer with 14 units corresponding to the emotion classes.

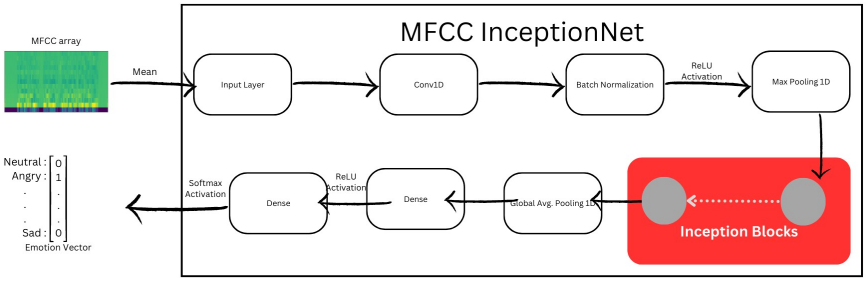


Figure 16: Architecture of MFCC InceptionNet Model

MFCC ResNet: The MFCC ResNet model shown in Figure 17 is a modified version of the ResNet architecture tailored for 1D convolutional operations. Like the previous model, the input layer receives MFCC data with a shape of (58, 1). It begins with an initial convolutional layer, followed by a series of residual blocks. These blocks consist of two convolutional layers each, interspersed with batch normalisation and ReLU activation. The shortcut connections within the residual blocks facilitate the flow of gradients during training, aiding in training deeper networks. The architecture concludes with global average pooling and dense layers, leading to an output layer with 14 units for classification.

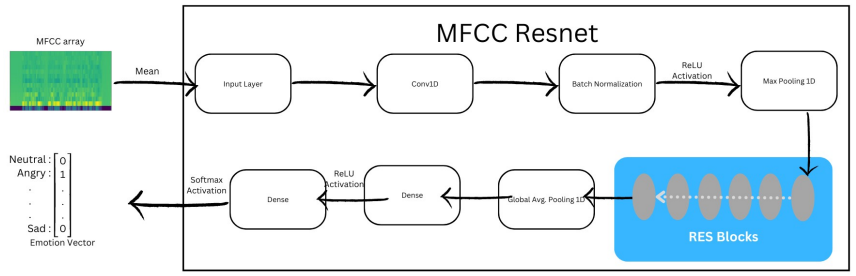


Figure 17: Architecture of MFCC ResNet Model

Mel InceptionNet: The Mel InceptionNet model adopts the Inception architecture, adapted single channel images as in Figure 18. The input layer receives Logarithmic Mel Spectrogram data with a shape of (60, 185, 1). Initial convolution, batch normalisation, and ReLU activation are followed by max-pooling operations. The inception blocks, analogous to the previous model, are designed for 2D convolutions and are aimed at capturing spatial hierarchies within the spectrogram data. The model is then concluded with global average pooling, dense layers, and an output layer of 14 units.

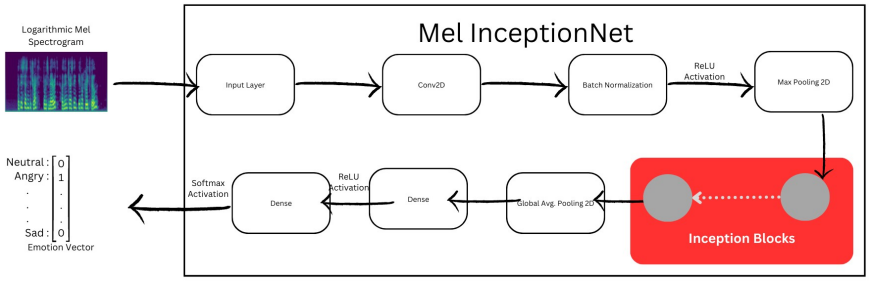


Figure 18: Architecture of Mel InceptionNet Model

Mel ResNet: The Mel ResNet model is also an adapted version of the ResNet architecture. The input layer receives Logarithmic Mel Spectrogram data with a shape of (60, 185, 1). It commences with an initial convolutional layer and subsequently includes residual blocks, like the previous ResNet models. The architecture has been tailored for 2D convolutions, with the residual blocks facilitating the extraction of complex spatial features from the spectrogram data. The model concludes with global average pooling, dense layers, and an output layer for classification. Its structure can be seen in Figure 19.

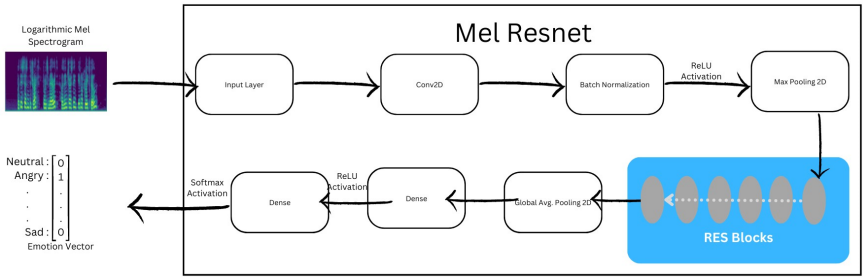


Figure 19: Architecture of Mel ResNet Model

3.7.2. Loss Function and Optimizer

For all four model architectures, the loss function utilised during training is categorical cross-entropy. This loss function is appropriate for multi-class classification tasks and measures the dissimilarity between predicted and actual class probabilities. The optimizer employed is RMSprop, a popular optimization algorithm that adapts the learning rate during training to enhance convergence and mitigate the vanishing/exploding gradient problem. RMSprop (Hinton, 2012) adjusts the learning rate for each parameter individually, making it suitable for training models with diverse architectures and data characteristics.

3.8. Evaluation Metrics

During the evaluation of our Speech Emotion Recognition (SER) models, two primary metrics were employed: Categorical Cross-Entropy Loss and Accuracy. These metrics are fundamental in assessing the performance and effectiveness of machine learning models for classification tasks.

3.8.1. Categorical Cross-Entropy Loss:

Categorical Cross-Entropy Loss (Koech, 2020), also known as softmax loss, is a widely used loss function for multi-class classification tasks. It quantifies the dissimilarity between the predicted class probabilities and the actual one-hot encoded labels. The formula for calculating categorical cross-entropy loss involves taking the negative logarithm of the predicted probability assigned to the true class label. In essence, it punishes the model more severely for confidently assigning a low probability to the correct class. By minimising this loss, the model learns to improve its predictions, moving towards accurately assigning higher probabilities to the correct classes.

$$L_{cross-entropy}\left(, \hat{y}\right) = \quad - \sum_{i=0}^{N} y_i \log log \hat{y}_i$$

$$y = Actual\ desirable\ output$$

$$\hat{y} = Predicted\ output$$

3.8.2. Accuracy:

Accuracy is a straightforward evaluation metric that quantifies the ratio of correctly predicted samples to the total number of samples in the dataset. It provides a clear understanding of the model's overall correctness in its predictions. While accuracy is an important measure, it can sometimes be misleading, especially in imbalanced datasets where a model may achieve high accuracy by simply predicting the majority class. However, in our study, as the dataset was relatively balanced, accuracy provides a reliable indication of how well the models perform in recognizing various emotion classes.

$$Accuracy\left(y, \hat{y}\right) = \frac{\sum_{i=0}^{N} \delta\left(y_i, \hat{y}_i\right)}{N}$$

$$y = Actual\ desirable\ output$$

$$\hat{y} = Predicted\ output$$

$$\delta\left(y_i, \hat{y}_i\right) = \{1,\ y_i = \hat{y}_i\ 0,\ y_i \neq \hat{y}_i$$

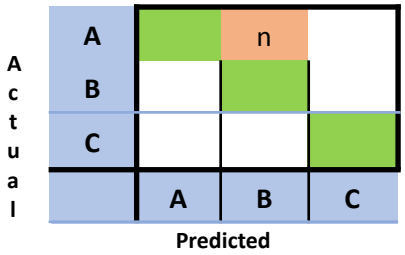
The selection of categorical cross-entropy loss and accuracy as evaluation metrics was guided by their appropriateness for multi-class classification tasks and their intuitive interpretability. The categorical cross-entropy loss helped us optimise our models by penalising incorrect predictions more heavily and encouraging the models to assign higher probabilities to the correct classes. Accuracy, on the other hand, offered a clear representation of the overall predictive performance of our models across various emotions. These metrics, in combination, provide a comprehensive insight into the capabilities and limitations of our CNN-based Speech Emotion Recognition models.

3.8.3. Confusion Matrix

Confusion matrix is an important metric for evaluating the classification model. Essentially it is a square matrix. It compares in detail how many of the samples given were classified correctly and shows detail about how many of the misclassified samples were classified under which category wrongly.

On the principal diagonal, it shows how many of the samples were predicted correctly class-wise. And all the other values represent the number of samples classified wrong. Let's say, we have a number 'n' in the matrix. On the x-axis, its label is A, and on the Y-axis its label is B. Then n represents the total number of samples whose actual label was A but were misclassified under label B. If A and B are both the same, that indicates the model has correctly predicted the label.

Table 3: Demonstration of Confusion Matrix



5.

3.8.4. Precision

Precision measures the accuracy of positive predictions made by a model. In other words, if TP (True Positives) is the number of instances correctly classified as positive, FP (False Positives) is the number of instances wrongly classified as positive when they are negative.

$$Precision = \frac{TP}{TP + FP}$$

3.8.5. Recall

Recall assesses a model's ability to capture all positive instances in the dataset. It is a magnitude of correctly identified samples from positive instances. In mathematical format, we can say, that if TP (True Positives) is the number of instances correctly classified as positive, FN (False Negatives) is the number of instances wrongly classified as negative when they are positive.

$$Recall = \frac{TP}{TP + FN}$$

3.8.6. F1 Score

The F1 Score is a harmonic mean of precision and recall, offering a balanced evaluation metric that considers both false positives and false negatives. It is particularly useful when dealing with imbalanced datasets, where one class significantly outnumbers the other.

It can be expressed as if TP (True Positives) is the number of instances correctly classified as positive, FN (False Negatives) is the number of instances wrongly classified as negative when they are positive, FP (False Positives) is the number of instances wrongly classified as positive when they are negative, then

$$F1 = \frac{2}{\frac{1}{Accuracy} + \frac{1}{Recall}} = \frac{2\,(TP)}{2TP + FN + FP}$$

3.8.7. Evaluation Metrics and Early Stopping in Our Model Training

In our model training process, we employ validation accuracy as a crucial evaluation metric. This metric measures how well the model performs on a separate validation dataset, providing a reliable gauge of its generalisation capabilities. Then after the complete training of the model, we evaluate the model's precision, recall, F1-score, and confusion matrix to compare it with other models.

To ensure optimal model performance, we implement an early stopping mechanism. This means that training will continue until there has been no improvement in validation accuracy for an extended period, preventing overfitting and ensuring that the model converges to the best possible solution.

This chapter has presented a comprehensive methodology for advancing Speech Emotion Recognition (SER) using Convolutional Neural Networks (CNNs). The research framework and tools, including the programming language, libraries, and packages, were meticulously selected to create an efficient and effective environment. The dataset's composition, augmentation techniques, and feature extraction processes were detailed, ensuring the dataset's diversity and the features' efficacy. The focus on Mel-Frequency Cepstral Coefficients (MFCC) and Logarithmic Mel Spectrogram as optimal features showcases a strategic feature selection process. Moreover, the design of diverse CNN architectures, such as MFCC InceptionNet, ResNet MFCC, Mel InceptionNet, and Mel ResNet, was discussed along with the employed loss function and optimizer. The following chapter delves into the results and subsequent discussions, shedding light on the performance of these models in recognizing emotions from speech data and their implications across various domains.

Chapter 4: Results and Discussion

4.1. Attained Metrics and Model Performance

Following the training and evaluation of our four Speech Emotion Recognition (SER) models – MFCC ResNet, MFCC InceptionNet, Mel InceptionNet, and Mel ResNet – we meticulously analysed an array of metrics to gauge their performance across various emotional classes. These metrics provide a comprehensive understanding of the models' capabilities, highlighting their strengths and areas for improvement.

4.1.1. MFCC ResNet Model:

The MFCC ResNet model achieved an overall accuracy of 95%, substantiating its competency in recognizing emotions from audio data. Precision, recall, and F1-score measures were assessed for each emotion class. Notably, the model demonstrated high precision and recall for most classes, illustrating its effectiveness in both minimising false positives and capturing true positives. The F1-score, which balances precision and recall, indicated the model's harmonious performance in capturing nuanced emotions.

Table 4: MFCC ResNet Model metrics

Gender	Emotion	Precision	Recall	F1-score	Support
Female	Sad	0.97	0.97	0.97	1510
Female	Neutral	0.96	0.95	0.96	1536
Male	Fear	0.96	0.97	0.96	1552
Male	Happy	0.96	0.96	0.96	1538
Female	Disgust	0.96	0.96	0.96	1486
Female	Happy	0.96	0.97	0.97	1568
Male	Angry	0.99	1	0.99	726
Male	Neutral	0.96	0.94	0.95	1180
Male	Disgust	0.92	0.91	0.91	1149
Female	Angry	0.93	0.91	0.92	1101
Female	Fear	0.92	0.93	0.92	1123
Male	Sad	0.93	0.94	0.94	1209
Female	Surprise	0.91	0.93	0.92	1149
Male	Surprise	0.95	0.95	0.95	200
Accuracy				0.95	17027
Macro Avg.		0.95	0.95	0.95	17027
Weighted Avg.		0.95	0.95	0.95	17027

The discerning performance of the MFCC ResNet model was prominently showcased in its precision and recall scores. For numerous emotion classes, the model demonstrated exceptional precision and recall, underscoring its ability to discern genuine positives while minimising false positives. This was particularly evident in its remarkable precision-recall balance for emotions such as anger, happiness, and sadness. For instance, the model achieved a precision and F1-score of 99% in detecting male anger. Moreover, the confusion matrix in Figure 20 corroborated its efficacy, with just a mere two misclassified instances for this emotion. This exceptional performance can be attributed to the distinct nature of voice signals associated with anger, characterised by distinctive patterns in loudness, pitch, and other acoustic features.

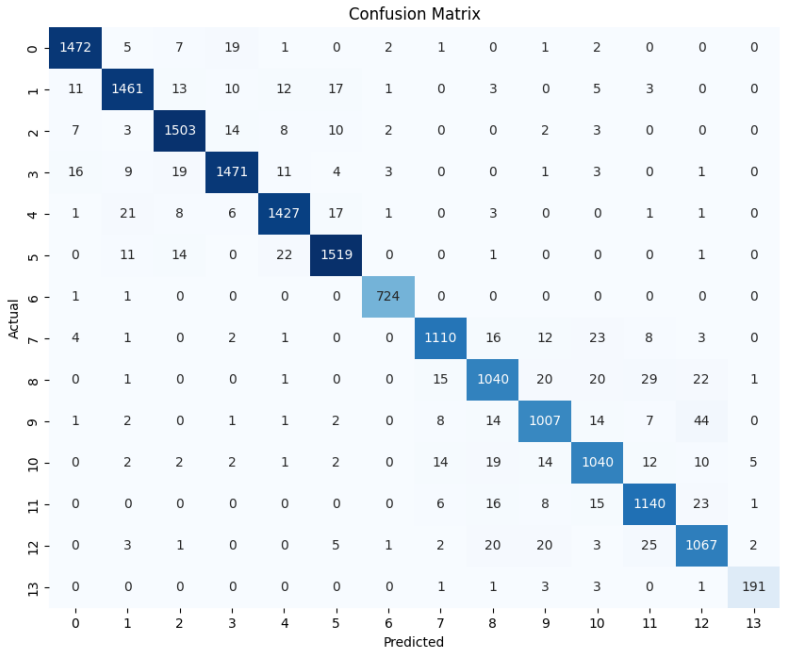


Figure 20: Confusion Matrix of MFCC Resnet

Conversely, the model's relatively lower performance was in detecting female surprise, achieving a precision of 91% and an F1-score of 92% (Visible in Table 4). Although these metrics are respectable, they pale in comparison to the model's performance on other emotions. The intricate nature of surprise, often a blend of diverse emotions like happiness, neutrality, or others not included in the classification, presents a challenge for the model. This complexity, inherent in human emotions, contributes to the lower precision and recall rates for this particular emotion.

4.1.2. MFCC InceptionNet Model:

The MFCC InceptionNet model displayed a commendable accuracy rate of 93%, once again highlighting its capability in emotion recognition from audio data. This outcome, although slightly lower in comparison to the MFCC ResNet model, reflects the model's reliability in capturing the essence of emotions. The evaluation metrics, including precision, recall, and F1-score, provide a comprehensive perspective on the model's performance across diverse emotion classes in Table 5.

Table 5: MFCC InceptionNet Model Metrics

Gender	Emotion	Precision	Recall	F1-score	Support
Female	Sad	0.97	0.95	0.96	1510
Female	Neutral	0.94	0.93	0.93	1536
Male	Fear	0.95	0.95	0.95	1552
Male	Happy	0.94	0.94	0.94	1538
Female	Disgust	0.94	0.94	0.94	1486
Female	Happy	0.94	0.95	0.94	1568
Male	Angry	0.99	1.00	0.99	726
Male	Neutral	0.93	0.93	0.93	1180
Male	Disgust	0.88	0.87	0.88	1149
Female	Angry	0.89	0.86	0.87	1101
Female	Fear	0.88	0.90	0.89	1123
Male	Sad	0.92	0.91	0.91	1209
Female	Surprise	0.87	0.91	0.89	1149
Male	Surprise	0.95	0.94	0.95	200
Accuracy				0.93	17027
Macro Avg.		0.93	0.93	0.93	17027
Weighted Avg.		0.93	0.93	0.93	17027

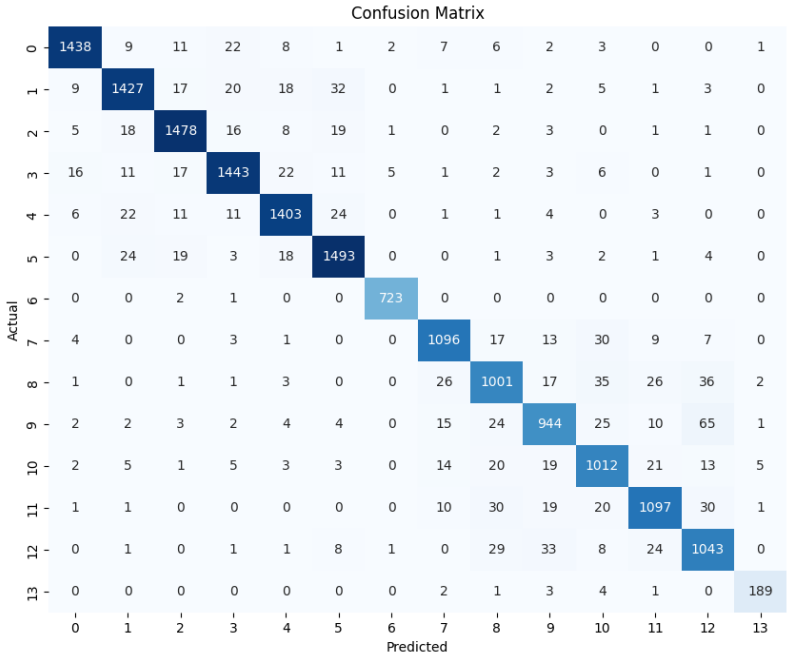


Figure 21: Confusion Matrix of MFCC InceptionNet

Notably, the model excelled in correctly identifying instances of the "angry" emotion, further underscoring the distinctive nature of this emotion's acoustic characteristics. This aligns with the previous observation and emphasises that either the quality of data related to anger is exceptional or that the emotion possesses inherent acoustic traits that differentiate it distinctly from other emotions. Within this category, the model demonstrated an astonishing F1 score and precision rate of 99%, marking a nearly perfect performance. However, the model did encounter challenges in a few cases, misclassifying three audio samples, suggesting that there might be room for improvement in fine-tuning the model's ability to handle subtle variations.

Conversely, the MFCC InceptionNet model faced difficulties in accurately identifying the "Female Fear" emotion, as is evident in the confusion matrix in Figure 21. The highest number of samples misclassified fall under this category. This contrasts with the results obtained from the MFCC ResNet model, which displayed superior performance in this particular emotion. This divergence indicates that the dataset contains sufficient discriminative information for fear, but the InceptionNet model might have encountered challenges in deciphering intricate acoustic patterns associated with fear. An examination of the confusion matrix in Figure 23 reveals that several misclassifications were made as "sad" and "disgust," hinting at the underlying complexities of distinguishing these emotions based on acoustic features.

4.1.3. Mel InceptionNet Model:

The Mel InceptionNet model stands out as a remarkable performer, boasting an exceptional accuracy rate of 96%. This remarkable achievement underscores the model's proficiency in effectively discerning intricate emotional variations within audio samples. The precision, recall, and F1-score metrics further corroborate the model's success across the spectrum of emotions, highlighting its prowess in capturing subtle nuances that define different emotional states.

Table 6: Mel InceptionNet Model Metrics

Gender	Emotion	Precision	Recall	F1-score	Support
Female	Sad	0.98	0.98	0.98	1510
Female	Neutral	0.97	0.98	0.97	1536
Male	Fear	0.97	0.97	0.97	1552
Male	Happy	0.98	0.97	0.98	1538
Female	Disgust	0.98	0.97	0.98	1486
Female	Happy	0.97	0.97	0.97	1568
Male	Angry	0.99	1.00	1.00	726
Male	Neutral	0.96	0.96	0.96	1180
Male	Disgust	0.93	0.95	0.94	1149
Female	Angry	0.95	0.95	0.95	1101
Female	Fear	0.94	0.94	0.94	1123
Male	Sad	0.95	0.95	0.95	1209
Female	Surprise	0.92	0.94	0.93	1149
Male	Surprise	0.98	0.95	0.96	200
Accuracy				0.96	17027

Macro Avg.	0.96	0.96	0.96	17027
Weighted Avg.	0.96	0.96	0.96	17027

Intriguingly, a closer examination of the model's metrics reveals a significant improvement compared to the MFCC-based approaches. This indicates that the use of the entire Mel spectrogram arrays as inputs, rather than averaging MFCC coefficients, led to an enhanced understanding of the acoustic characteristics of emotions. This heightened comprehension facilitated the model's impressive accuracy of 96%, a testament to its ability to unravel complex patterns within the audio data. Notably, the metrics for individual emotions consistently exceed the 90% threshold, further emphasising the model's robustness and adaptability. A particularly noteworthy outcome is the recognition of the "Male Anger" emotion, where the model demonstrated remarkable precision and recall, misclassifying only two audio samples. This reaffirms the distinctiveness of anger's acoustic characteristics and the model's capability to capture them accurately.

On the other hand, the "Female Surprise" emotion displayed the weakest performance, though even in this scenario, the F1-score and precision metrics stood at an impressive 0.93% and 92%, respectively (can be seen in Table 6). While this might be considered the least favourable outcome, it's important to note that achieving such high precision and F1-score values even for the least performing class is indicative of the model's overall strong performance. Interestingly, misclassifications for the "Female Surprise" emotion often leaned towards emotions like "Disgust," "Angry," and "Sad." This could be attributed to potential overlaps or complexities in the acoustic features of these emotions, making it challenging for the model to precisely differentiate "Surprise" from them. This also highlights the intricacies of emotion recognition and the nuances that exist in audio data.

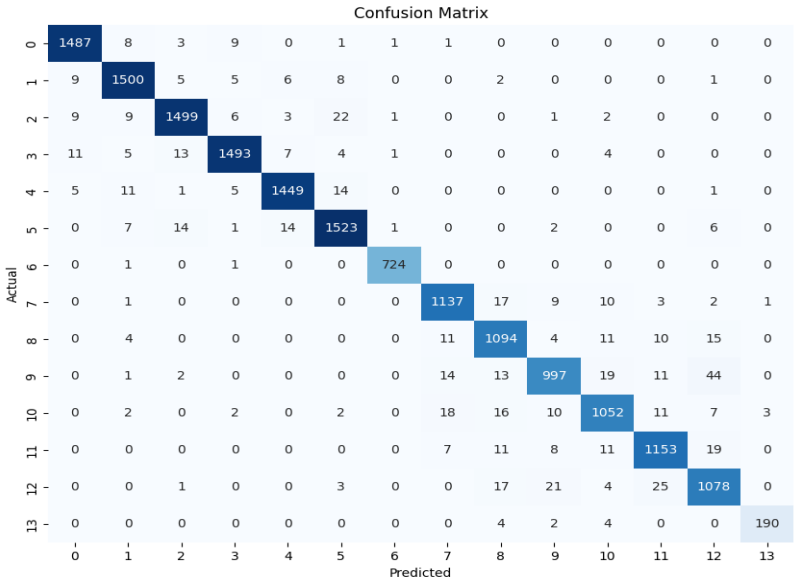


Figure 22: Confusion Matrix of Mel InceptionNet

4.1.4. Mel ResNet Model:

The Mel ResNet model stands out as the pinnacle of performance, boasting an impressive accuracy rate of 97%. The precision, recall, and F1-score metrics further underscore its excellence by consistently reflecting exceptional performance across all emotion classes. A defining characteristic of this model is its remarkable equilibrium between precision and recall, which signifies its proficiency in both making accurate positive predictions and capturing the actual positives within the dataset.

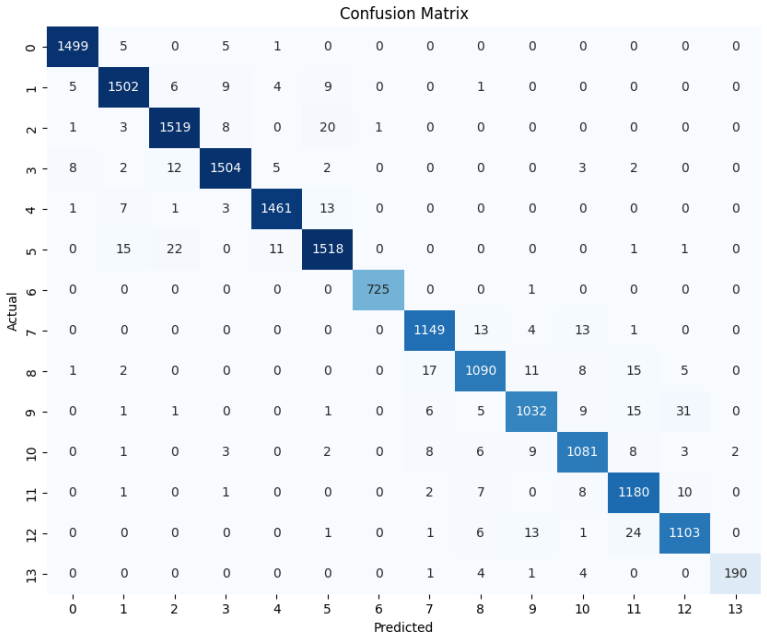


Figure 23: Confusion Matrix of Mel Resnet

Table 7: Mel ResNet Model metrics

Gender	Emotion	Precision	Recall	F1-score	Support
Female	Sad	0.99	0.99	0.99	1510
Female	Neutral	0.98	0.98	0.98	1536
Male	Fear	0.97	0.98	0.98	1552
Male	Happy	0.99	0.98	0.98	1538
Female	Disgust	0.99	0.98	0.98	1486
Female	Happy	0.97	0.97	0.97	1568
Male	Angry	1.00	1.00	1.00	726
Male	Neutral	0.97	0.97	0.97	1180
Male	Disgust	0.96	0.95	0.96	1149
Female	Angry	0.96	0.94	0.95	1101
Female	Fear	0.96	0.96	0.96	1123
Male	Sad	0.95	0.98	0.96	1209
Female	Surprise	0.96	0.96	0.96	1149

Male	Surprise	0.99	0.95	0.97	200
Accuracy				0.97	17027
Macro Avg.		0.97	0.97	0.97	17027
Weighted Avg.		0.97	0.97	0.97	17027

Noteworthy is the model's exceptional achievement in achieving consistent accuracy across all emotions. The minimum we got was 96% in case of female anger, fear, surprise, and male disgust as expressed in Table 7. Moreover, Figure 23 shows the extreme density of the figures on the diagonal, which shows the correctly identified samples.

This profound consistency serves as a testament to the power of ResNet architecture, as well as a testament to the significance of our decision to move away from the loss of information through mean calculation.

It is remarkable to observe that almost all emotions achieved metrics hovering around the 95% mark. This outcome signifies that the model has essentially reached a point of equilibrium, beyond which incremental improvements might prove challenging. The uniformity in the metrics for various emotions underscores the model's well-rounded ability to recognize a wide spectrum of emotions with consistent accuracy.

4.2. Effect of Augmentation

The role of data augmentation in achieving the remarkable results presented earlier cannot be overstated. Before the integration of augmentation techniques, the achieved accuracy hovered around the modest 60-70% range. However, the inclusion of augmentation marked a turning point in the performance of the models, facilitating a significant leap in accuracy and proficiency.

Augmentation serves as a pivotal strategy for enhancing the dataset's richness by generating a diverse array of samples. In doing so, it broadens the model's exposure to a wider spectrum of variations inherent in emotional speech patterns. Consequently, the model's learning process becomes more robust, as it begins to discern and internalise the underlying patterns that define distinct emotional expressions.

One of the remarkable outcomes of augmentation is the model's improved capability to recognize and generalise patterns present in the training data. By encountering a multitude of augmented instances, the model refines its understanding of the intricate characteristics that differentiate various emotional states. This refined comprehension leads to enhanced performance when presented with unseen data, illustrating the augmentation-driven boost in the model's generalisation power.

The significance of augmentation in the realm of audio applications, particularly in emotion recognition, is noteworthy. While augmentation is comparatively straightforward to implement in audio, its impact on model outcomes is nothing short of transformative. The technique capitalises on the inherent nature of audio data, where slight variations in pitch, tempo, or background noise can significantly influence the emotional context. By mimicking these real-world variations through augmentation, the model becomes sensitised to the subtleties present in different emotions, thus elevating its discernment capabilities.

Additionally, augmentation acts as a form of regularisation by injecting controlled variability into the training data. This regularisation dampens the risk of overfitting, where the model might become overly specialised to the training data and struggle to generalise to unseen instances. The introduction of augmented samples encourages the model to distil the core features that define each emotion while being less susceptible to noise or irrelevant fluctuations.

In essence, the profound impact of data augmentation in this study is a testament to its ability to bridge the gap between training data and real-world variability. (S, N. R. et al., 2018) tried to make a model without using any augmentation based on the architecture of InceptionNet. But the results were not good. The research yielded an accuracy of 35.6%.

And if we compare our model to the work of (An & Ruan, 2021) – which was one of our inspirations – they were able to achieve an accuracy of 80%. The only augmentation they added was noise. We took another step forward. We added more variations of augmentation. As a result of this step, we were able to achieve 17% better accuracy. Through multiple augmentations, the models transcend their initial limitations and achieve results that underscore the power of leveraging diverse training samples for robust and accurate emotion recognition.

4.3. Implications and Future Directions:

The comprehensive evaluation of these metrics underscores the potential of Mel spectrogram-based models, particularly the Mel ResNet, in capturing complex emotional cues from audio data. These models showcase the intricate interplay between deep learning architectures and feature representations, illuminating the importance of data preprocessing and architecture selection in SER tasks.

As we delve further into the discussion of these results, we will scrutinise the implications of these findings in the context of existing literature, addressing the challenges and opportunities in Speech Emotion Recognition. The subsequent section will shed light on the significance of these models' performance in real-world applications, guiding the development of effective emotion-aware systems.

In conclusion, the meticulous analysis of attained metrics elucidates the varying capacities of the four models, while also laying the foundation for deeper insights into the intricate world of emotional expression in audio data.

Chapter 5: Conclusion

5.1. Remarks and observation:

With this research, we embarked on a journey to research and develop a reliable Speech Emotion Recognition model using Convolutional Neural Networks. We combined four sets of audio data and followed a careful process to work through the complexities of enhancing audio and pulling out useful features. We looked at things like Mel Spectrograms, MFCC constants, Zero Crossing Rate, and Short-Time Energy to build a strong foundation for our study. After testing and experimenting, we decided that Mel Spectrograms and MFCC constants were the most important features to focus on.

We explored a range of model types, starting from simple Convolutional Neural Networks (CNNs) to more complex ResNet structures. Through thorough analysis, we identified four models that consistently performed the best. Among these, the InceptionNet and ResNet models were the strongest, especially when using both MFCC constants and Mel Spectrogram data. We carefully examined these chosen models and learned a lot from their results. It's worth noting that all four models achieved an accuracy of over 90% on new, unseen data. Our most exciting achievement was creating a fine-tuned version of the ResNet model tailored to our needs. We trained it on a large dataset of about 68,000 samples. Our model converged to the point where it achieved an astonishing macro and weighted average of 97% and F1-Score of 0.97 on unseen test data of size 17027.

This project gives a clear roadmap of how we went from improving audio data to choosing the best model designs. Our findings show that advanced neural networks can be powerful, especially when we pick the right features and adapt the appropriate technique. As technology gets better and we have more data to work with, the field of recognizing emotions in speech will keep getting better too.

5.2. Applications of Speech Emotion Recognition

The implications of our research reverberate beyond the confines of academic inquiry, finding resonance in real-world scenarios. The culmination of our efforts in the realm of Speech Emotion Recognition (SER) holds the promise of transformative applications across various domains.

In the field of mental health, our model's prowess in decoding emotional cues embedded within vocal expressions can be harnessed to facilitate early detection of emotional distress. By analysing speech patterns, the system could alert healthcare professionals to potential signs of anxiety, depression, or other emotional states. A few examples are cited as (Leung et al., 2022) and (Espinola et al., 2020) This proactive approach could pave the way for timely interventions and support, thereby mitigating the severity of mental health challenges.

The entertainment industry also stands to reap the benefits of our research. In media production, our SER model could be employed to tailor content to the emotional engagement of the audience. (Song et al., 2014) Whether in movies, advertisements, or video games, adaptive content creation based on the real-time emotional response of the viewer could enhance engagement and immersion. Also, it could be used to evaluate and analyse the performance of an actor’s voice and his emotions in his dialogue delivery.

Moreover, in customer service and market research, our model's capacity to decipher customer sentiment from spoken interactions could prove invaluable. Businesses could analyse customer

feedback calls, enabling them to gauge customer satisfaction, uncover pain points, and fine-tune their products and services accordingly. This data-driven approach to customer interactions could lead to enhanced customer experiences and brand loyalty.

5.3. Drawbacks and Limitations

In our pursuit of developing effective Speech Emotion Recognition (SER) models, several noteworthy limitations and drawbacks have come to the forefront, underscoring the need for cautious interpretation and continuous improvement.

5.3.1. Variability in Real-Life Data:

One of the substantial challenges we encountered was the variation between our synthesised training datasets and real-life data. While our models exhibited commendable performance on the synthesised datasets, their performance on real-life data can be considerably different. Real-life recordings encompass a wide array of conditions, including diverse recording environments, microphone characteristics, and inherent variability in human speech patterns. This can lead to unpredicted variations in the models' predictive accuracy when applied to real-world scenarios.

5.3.2. Complexity of Mixed Emotions:

Emotions are intricate and can often manifest as combinations of different emotional states. Our models were primarily trained to recognize distinct emotional categories. Consequently, their ability to accurately predict mixed emotions, where two or more emotional states coexist, remains limited. This limitation underscores the need for more sophisticated approaches that can decipher and classify complex emotional blends.

5.3.3. Synthetic Training Data vs. Real-Life Variation:

Despite our efforts to diversify the training data by incorporating four datasets, it's important to note that all of these datasets were synthesised for research purposes. Real-life data introduces an added layer of complexity due to its inherent variability. Natural recordings encompass a broader spectrum of vocal characteristics, dialects, accents, and cultural influences that might not have been fully captured by our synthesised datasets. As a result, the models' performance on these real-life variations might differ from our synthesised data evaluations.

5.3.4. Bias in Dataset Selection:

The selection of training data is pivotal in determining the performance and fairness of machine learning models. Our datasets were sourced from various online repositories, potentially introducing biases present in those sources. These biases can influence the models' predictions and limit their effectiveness across diverse demographic groups. Efforts were made to mitigate bias, but ensuring complete fairness and equity remains a persistent challenge.

In conclusion, while our study represents a significant advancement in Speech Emotion Recognition, these limitations and drawbacks emphasise the complexities inherent in training models to understand human emotions accurately. Addressing these challenges requires interdisciplinary collaboration, extensive real-life data collection, and the implementation of bias-mitigation strategies. Through such endeavours, we can advance the field and create more reliable, inclusive,

and effective models that contribute positively to human-computer interaction and emotional understanding.

5.4. Further Research and Development

As we conclude our study on Speech Emotion Recognition (SER), it becomes evident that there are numerous avenues for future research and refinement in this dynamic field. Our work has illuminated various potential areas of focus that could pave the way for more accurate, robust, and versatile SER systems. Here, we outline some promising avenues for future exploration:

5.4.1. Real-Life Data Collection

The incorporation of genuine, diverse, and extensive real-life data can substantially enhance the applicability and generalisation capabilities of SER models. Collecting a comprehensive dataset that represents various cultures, accents, and emotional expressions can address the disconnect between synthesised data and real-world scenarios.

5.4.2. Multimodal Approaches

Emotions are not solely conveyed through speech; facial expressions, gestures, and physiological signals also play a crucial role. Exploring multimodal approaches that combine audio and visual cues can lead to more comprehensive and accurate emotion recognition systems.

5.4.3. Handling Mixed Emotions

Developing models that can effectively detect and classify mixed emotions will be crucial, as human emotions often exist in complex blends. Investigating techniques that decipher the intricate interplay of emotions within a single utterance can lead to a deeper understanding of emotional nuances.

5.4.4. Continuous Emotion Recognition

Most existing models operate on static audio clips. Developing systems that perform continuous emotion recognition on dynamic speech streams can capture evolving emotional states and offer real-time insights for applications like mental health monitoring or human-computer interaction.

5.4.5. Clinical and Therapeutic Applications

Emotion recognition technology holds potential in clinical and therapeutic contexts, such as assessing emotional states in therapy sessions or aiding individuals with autism spectrum disorder in recognizing emotions. Exploring these applications can lead to impactful contributions.

In essence, our study offers a stepping stone towards more advanced, nuanced, and adaptable SER systems. By addressing the aforementioned areas and exploring innovative avenues, researchers can contribute to the evolution of emotion recognition technology, enabling its integration into a plethora of real-world applications that benefit human well-being and communication. The journey of understanding emotions in speech continues, with the promise of even greater achievements on the horizon.

References

Abadi, M. et al., 2015. *TensorFlow: Large-scale machine learning on heterogeneous systems.* s.l.:Tensorflow.org.

Anon., 2018. *resnet18.* [Online]

Available at: https://www.mathworks.com/help/deeplearning/ref/resnet18.html

An, X. & Ruan, Z., 2021. Speech Emotion Recognition algorithm based on deep learning algorithm fusion of temporal and spatial features. *Journal of Physics: Conference Series.*

Aouani, H. & Ayed, Y. B., 2020. Speech Emotion Recognition with deep learning. *Procedia Computer Science,* Volume 176, pp. 251-260.

Ayadi, M. E., Kamel, M. S. & Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition,* Volume 44, pp. 572-587.

BABKO, D., 2020. *Speech Emotion Recognition (en).* [Online]

Available at: https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en

[Accessed 08 2023].

Busso, C. et al., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang Resources & Evaluation,* Volume 42, pp. 335-359.

Dalvi, C. et al., 2021. A Survey of AI-based Facial Emotion Recognition: Features, ML & DL Techniques, Age-wise Datasets and Future Directions. *IEEE Access.*

Deshmukh, G., Gaonkar, A., Golwalkar, G. & Kulkarni, S., 2019. Speech-based Emotion Recognition using Machine Learning. *International Journal of Research and Analytical Reviews,* pp. 812-817.

Eklund, V.-V., 2019. *Data augmentation techniques for robust audio analysis.* s.l.:s.n.

Espinola, C. W., Gomes, J. C., Pereira, J. M. S. & Santos, W. P. d., 2020. *Detection of Major Depressive Disorder Using Vocal Acoustic Analysis and Machine Learning.* s.l.:Cold Spring Harbor Laboratory Press.

Fernandes, J. et al., 2018. Harmonic to Noise Ratio Measurement - Selection of Window and Length. *Procedia Computer Science,* Volume 138, pp. 280-285.

Google, n.d. *Google Colaboratory.* [Online]

Available at: https://research.google.com/colaboratory/faq.html

[Accessed 8 2023].

Grossi, E. & Buscema, M., 2008. Introduction to artificial neural networks. *European Journal of gastroenterology & hepatology,* pp. 1046-54.

Haq, S., Jackson, P. & Edge, J., 2008. Audio-visual feature selection and reduction for emotion classification. In: *Proc.\ Int.\ Conf. on Auditory-Visual Speech Processing (AVSP'08),.* Tangalooma, Australia: s.n.

Harris, C. R. et al., 2020. Array programming with NumPy. *Nature,* Volume 585, pp. 357-362.

Hinton, G., 2012. *Coursera - Neural Networks for Machine Learning - Geoffrey Hinton.* s.l.:s.n.

Hossan, M., Memon, S. & Gregory, M., 2011. *A novel approach for MFCC feature extraction.* s.l., s.n., pp. 1-5.

Huang, C., Gong, W., Fu, W. & Feng, D., 2014. *A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM.* s.l., s.n.

Hunter, J. D., 2007. Matplotlib: A 2D graphics environment. *Computing in Science \& Engineering,* Volume 9, pp. 90-95.

Jalil, M., Butt, F. A. & Malik, A., n.d. *Short-Time Energy, Magnitude, Zero Crossing Rate and Autocorrelation Measurement for Discriminating Voiced and Unvoiced segments of Speech Signals,* s.l.: s.n.

Kerkeni, L. et al., 2022. *Automatic Speech Emotion Recognition Using Machine Learning.* s.l., s.n.

Koech, K. E., 2020. *Medium.* [Online]

Available at: https://towardsdatascience.com/cross-entropy-loss-function-f38c4ec8643e

[Accessed 26 8 2023].

Kopparapu, S. K. & Laxminarayana, M., 2010. *CHOICE OF MEL FILTER BANK IN COMPUTING MFCC OF A RESAMPLED SPEECH.* Kuala Lumpur, Malaysia, s.n.

Krizhevsky, A., 2014. *One weird trick for parallelizing convolutional neural networks.* s.l.:arXiv.

Lech, M., Stolar, M., Best, C. & Bolia, R., 2020. Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding. *Frontiers in Computer Science,* Volume 2.

Leung, F. Y. N. et al., 2022. Emotion recognition across visual and auditory modalities in autism spectrum disorder: A systematic review and meta-analysis. *Developmental Review,* Volume 63.

Lim, W., Jang, D. & Lee, T., 2016. *Speech emotion recognition using convolutional and Recurrent Neural Networks.* s.l., s.n.

Livingstone, S. R. & Russo, F. A., 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE,* Volume 13, p. 135.

McFee, et al., 2015. librosa: Audio and music signal analysis in Python. *Proceedings of the 14th Python in science conference,* pp. 18-25.

McKinney, W., 2010. Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference,* Volume 445, pp. 56-61.

Mirsamadi, S., Barsoum, E. & Zhang, C., 2017. *Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention.* s.l., s.n., pp. 2227-2231.

Montufar, G., 2018. *Restricted Boltzmann Machines: Introduction and Review.* s.l.:arXiv.

pawangfg, 2023. *Residual Networks (ResNet) – Deep Learning.* [Online]

Available at: https://www.geeksforgeeks.org/residual-networks-resnet-deep-learning/

Pedregosa, F. et al., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* Volume 12, pp. 2825-2830.

Python, 2023. *Applications for Python.* [Online]

Available at: https://www.python.org/about/apps/

[Accessed 08 2023].

Selvaraj, M., Dr.R.Bhuvana & S.Padmaja, 2016. HUMAN SPEECH EMOTION. *International Journal of Engineering and Technology,* Volume 8, pp. 311-323.

Shaikh, J., 2018. *Deep Learning in the Trenches: Understanding Inception Network from Scratch.* [Online]

Available at: https://www.analyticsvidhya.com/blog/2018/10/understanding-inception-network-from-scratch/

Simonyan, K. & Zisserman, A., 2015. *Very Deep Convolutional Networks for Large-Scale Image Recognition.* s.l.:arXiv.

Singh, J., Saheer, L. B. & Faus, O., 2023. Speech Emotion Recognition Using Attention Model. *International Journal of Environmental Research and Public Health,* Volume 20.

Singh, J., Saheer, L. B. & Faust, O., 2023. Speech Emotion Recognition Using Attention Model. *International Journal of Environmental Research and Public Health,* Volume 20.

Slaney, M., 2000. An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank.

S, N. R., M, P. & Betty. P, 2018. Speech Emotion Recognition using Deep Learning. *International Journal of Recent Technology and Engineering (IJRTE),* 7(4S).

Song, K.-T., Han, M.-J. & Wang, S.-C., 2014. Speech signal-based emotion recognition and its application to entertainment robots. *Journal of the Chinese Institute of Engineers,* Volume 37, pp. 14-25.

Song, M.-h.et al., 2005. Support Vector Machine Based Arrhythmia Classification Using Reduced Features. *International Journal of Control Automation and Systems,* Volume 3, pp. 571-579.

Szegedy, C., Liu, W. & Jia, Y., 2015. *Going Deeper with Convolutions.* s.l., s.n.

Szegedy, C., Vanhoucke, V., Ioffe, S. & Shlens, J., 2015. *Rethinking the Inception Architecture for Computer Vision; Zbigniew Wojna.* s.l.:arXix.

Togootogtokh, E. & Klasen, C., 2021. *DeepEMO: Deep Learning for Speech Emotion Recognition.* s.l.:arXiv.

Tzinis, E. & Potamianos, A., 2017. Segment-Based Speech Emotion Recognition Using Recurrent Neural Networks. In: *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII).* s.l.:s.n., pp. 190-195.

Vaswani, A. et al., 2023. *Attention Is All You Need.* s.l.:arXiv.

Venkata Subbarao, M., Terlapu, S. K., Geethika, N. & Harika, K. D., 2022. Speech Emotion Recognition Using K-Nearest Neighbor Classifiers. In: *Recent Advances in Artificial Intelligence and Data Engineering.* Singapore: Springer Singapore, pp. 123-131.

Waskom, M. L., 2021. seaborn: statistical data visualization. *Journal of Open Source Software,* Volume 6, p. 3021.

Zheng, L., Li, Q., Ban, H. & Liu, S., 2018. *Speech emotion recognition based on convolution neural network combined with random forest.* s.l., s.n., pp. 4143-4147.

APPENDIX A: ETHICAL APPROVAL

The ethical approval letter is included in the Appendix Material.

APPENDIX B: Training Graphs of Models

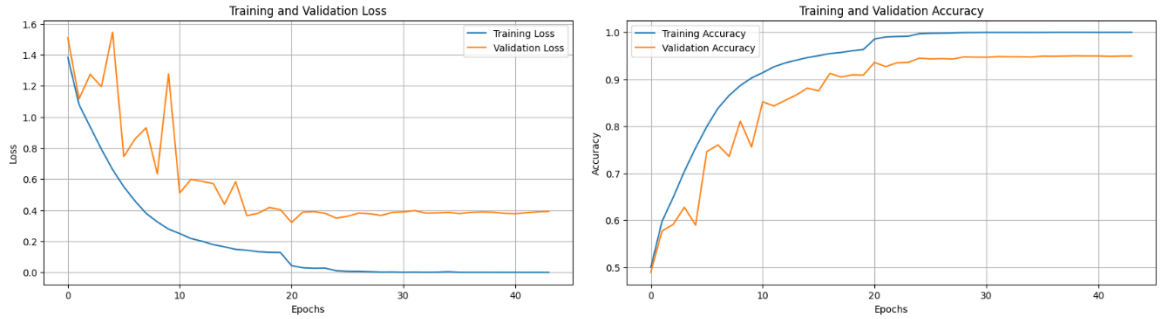


Figure 24: MFCC Resnet Training And Validation Loss and Accuracy

Figure 25: MFCC InceptionNet Training And Validation Loss and Accuracy

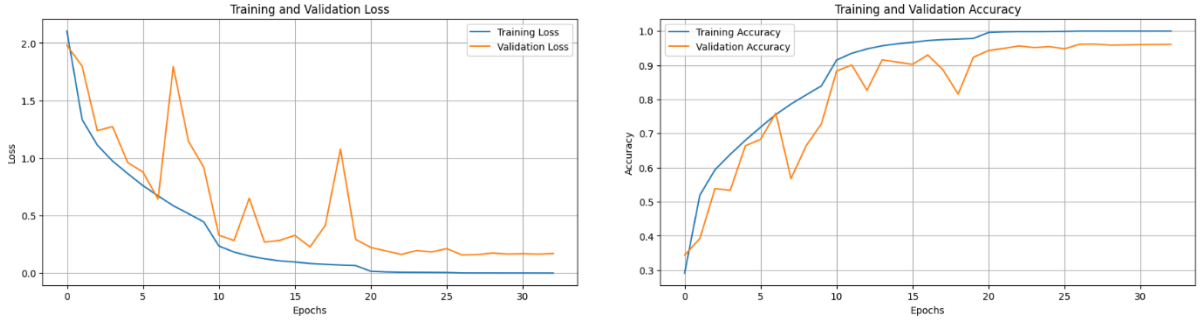


Figure 26: Mel InceptionNet Training And Validation Loss and Accuracy

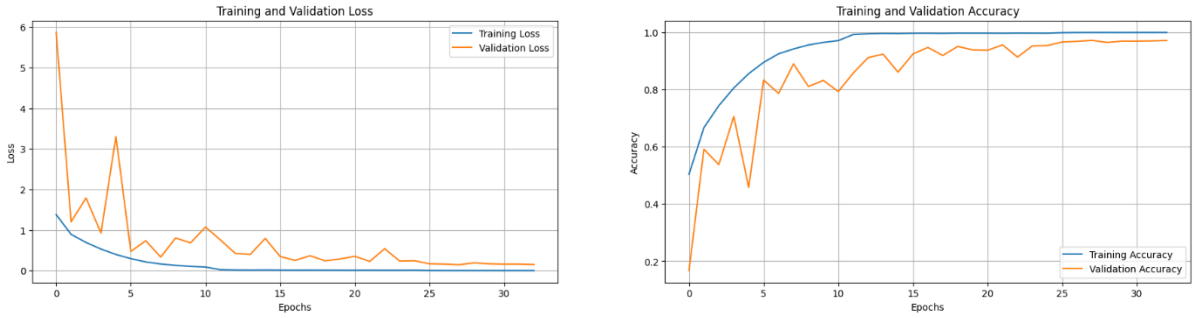


Figure 27: Mel ResNet Training And Validation Loss and Accuracy

The code files are included in the Appendix Material.