A black and purple background

Description automatically generated

A purple and blue waves on a black background

Description automatically generated

Word Count: XXXX

Date: DD/MM/YYYY

KV6003 – Individual Computing Project

Panagiotis Theodorou – W20020633

Supervisor: Dr Hua Mao

2nd Marker: Dr Nic Whitton

A Comparative Analysis of Machine Learning and Deep Learning Models

Speech Emotion Recognition (S.E.R)

**2023 - 2024**

Table of Contents

[1 Introduction 00](#_Toc161174001)

[1.1 Overview 00](#_Toc161174002)

[1.2 Objectives 00](#_Toc161174003)

[1.3 Scope and Significance 00](#_Toc161174004)

[2 Literature Review 00](#_Toc161174005)

[2.1 Studies of Models in Speech Emotion Recognition 00](#_Toc161174007)

[2.1.1 Deep Learning 00](#_Toc161174008)

[2.1.2 Machine Learning 00](#_Toc161174009)

[2.1.3 Ensemble Learning 00](#_Toc161174010)

[2.2 Challenges and Limitations 00](#_Toc161174011)

[2.2.1 Datasets 00](#_Toc161174012)

[2.2.1.1 CREMAD 00](#_Toc161174013)

[2.2.1.2 RAVDESS 00](#_Toc161174014)

[2.2.1.3 SAVEE 00](#_Toc161174015)

[2.2.1.4 EMODB 00](#_Toc161174016)

[2.2.2 Emotions 00](#_Toc161174017)

[2.2.3 Authenticity 00](#_Toc161174018)

[2.2.4 Cultural and Gender biases 00](#_Toc161174019)

[2.2.5 Subjectivity and Variability 00](#_Toc161174020)

[2.2.6 Overfitting 00](#_Toc161174021)

[3 Methodology 00](#_Toc161174022)

[3.1 Feature Extraction 00](#_Toc161174023)

[3.2 Augmentation 00](#_Toc161174024)

[3.3 Performance 00](#_Toc161174025)

[3.3.1 Deep Learning Models – Vector Features 00](#_Toc161174026)

[3.3.1.1 ANN 00](#_Toc161174027)

[3.3.1.2 CNN 00](#_Toc161174028)

[3.3.1.3 MLP 00](#_Toc161174029)

[3.3.1.4 ResNet 00](#_Toc161174030)

[3.3.1.5 RNN+LSTM 00](#_Toc161174031)

[3.3.2 Machine Learning Models – Vector Features 00](#_Toc161174032)

[3.3.2.1 GMM 00](#_Toc161174033)

[3.3.2.2 HMM 00](#_Toc161174034)

[3.3.2.3 KNN 00](#_Toc161174035)

[3.3.2.4 RF 00](#_Toc161174036)

[3.3.2.5 SVM 00](#_Toc161174037)

[3.3.3 Deep Learning – Image Features 00](#_Toc161174038)

[3.3.3.1 CNN 00](#_Toc161174039)

[3.3.3.2 DenseNet 00](#_Toc161174040)

[3.3.3.3 Inception 00](#_Toc161174041)

[3.3.3.4 ResNet 00](#_Toc161174042)

[3.3.3.5 VGG 00](#_Toc161174043)

[3.3.4 Ensemble Models 00](#_Toc161174044)

[3.3.4.1 Lorem Ipsum 00](#_Toc161174045)

[3.3.4.2 Lorem Ipsum 00](#_Toc161174046)

[3.3.4.3 Lorem Ipsum 00](#_Toc161174047)

[3.3.4.4 Lorem Ipsum 00](#_Toc161174048)

[3.4 Performance 00](#_Toc161174049)

[3.4.1 Datasets & Features 00](#_Toc161174050)

[3.5 Model Selection 00](#_Toc161174051)

[4 Own Approach 00](#_Toc161174052)

[4.1 Lorem Ipsum 00](#_Toc161174053)

[4.2 Lorem Ipsum 00](#_Toc161174054)

[4.3 Lorem Ipsum 00](#_Toc161174055)

[4.4 Lorem Ipsum 00](#_Toc161174056)

[4.5 Lorem Ipsum 00](#_Toc161174057)

[5 Implementation 00](#_Toc161174058)

[5.1 Lorem Ipsum 00](#_Toc161174059)

[5.2 Lorem Ipsum 00](#_Toc161174060)

[5.3 Lorem Ipsum 00](#_Toc161174061)

[5.4 Lorem Ipsum 00](#_Toc161174062)

[5.5 Lorem Ipsum 00](#_Toc161174063)

[6 Results 00](#_Toc161174064)

[6.1 Lorem Ipsum 00](#_Toc161174065)

[6.2 Lorem Ipsum 00](#_Toc161174066)

[6.3 Lorem Ipsum 00](#_Toc161174067)

[6.4 Lorem Ipsum 00](#_Toc161174068)

[6.5 Lorem Ipsum 00](#_Toc161174069)

[7 Discussion 00](#_Toc161174070)

[7.1 Lorem Ipsum 00](#_Toc161174071)

[7.2 Lorem Ipsum 00](#_Toc161174072)

[7.3 Lorem Ipsum 00](#_Toc161174073)

[7.4 Lorem Ipsum 00](#_Toc161174074)

[7.5 Lorem Ipsum 00](#_Toc161174075)

[8 Conclusion 00](#_Toc161174076)

[8.1 Lorem Ipsum 00](#_Toc161174077)

[8.2 Lorem Ipsum 00](#_Toc161174078)

[8.3 Lorem Ipsum 00](#_Toc161174079)

[8.4 Lorem Ipsum 00](#_Toc161174080)

[8.5 Lorem Ipsum 00](#_Toc161174081)

**Title: Speech Emotion Recognition Using Machine Learning and Deep Learning Models**

**Abstract:**

Provide a concise summary of your report, including the purpose, methodology, key findings, and implications of your research.

# Introduction

## Overview

Speech Emotion Recognition (SER) utilizing Artificial Intelligence (AI) signifies a revolutionary approach in understanding human emotions, enhancing the interaction between humans and machines. The integration of AI, particularly through machine learning and deep learning techniques, plays a pivotal role in advancing SER technologies. These AI methodologies have substantially elevated the accuracy and efficiency of emotion recognition from speech, moving beyond traditional methods that often struggled with the complexity and subtlety of human emotional expression. The application of these advanced AI techniques allows for a more nuanced and sophisticated analysis of speech patterns, capturing the intricate variations in tone, pitch, and tempo that characterize different emotional states (Kerkeni et al., 2019; Chandrasekar, Chapaneri, & Jayaswal, 2014).

The development of SER technologies, however, is not without its challenges. One of the main difficulties lies in accurately capturing the subtle emotional cues present in diverse speech patterns. The variability of emotional expression across different languages and cultures further complicates this task. To address these challenges, AI-driven approaches have been employed, including the use of machine learning algorithms for feature extraction and classification. These methods have shown promise in overcoming the limitations posed by the variability of speech and emotional expression, providing more reliable and accurate emotion recognition capabilities (Kerkeni et al., 2019; Wani, Gunawan, Qadri, Kartiwi, & Ambikairajah, 2021).

SER finds its application in a wide range of fields, demonstrating its versatility and broad impact. From customer service bots that can understand and respond to the emotional state of users, to therapeutic and healthcare settings where it can aid in patient care, SER's applications are vast. Educational software can benefit from SER by adapting to the emotional needs of students, while the entertainment industry can use it to create more engaging and responsive experiences. The widespread applicability of SER underscores its potential to revolutionize how we interact with technology, making these interactions more human-like and responsive to our emotional states (Kerkeni et al., 2019; Swain, Routray, & Kabisatpathy, 2018).

In conclusion, the integration of AI in SER represents a significant advancement in our ability to recognize and respond to human emotions through technology. While challenges remain, particularly in dealing with the variability of speech and emotional expression across different contexts, AI-driven solutions offer a promising pathway forward. The applications of SER across various domains highlight its potential to enhance the quality of human-machine interactions, contributing to advancements in fields ranging from healthcare to education and entertainment. As research in this area continues, we can expect further innovations that will expand the capabilities and applications of SER, making it an increasingly integral part of our technological landscape (Chandrasekar, Chapaneri, & Jayaswal, 2014; Wani, Gunawan, Qadri, Kartiwi, & Ambikairajah, 2021).

## Objectives

## Scope and Significance

# Literature Review

## Studies of Models in Speech Emotion Recognition

### Deep Learning Models

Deep Learning models, as explored in the article by Fayek, Lech, and Cavedon (2017), demonstrate significant capabilities in Speech Emotion Recognition (SER). These models can effectively handle both static and dynamic classification problems, making them versatile for various applications. The employment of minimal speech processing and end-to-end deep learning architectures allows for the modelling of intra-utterance dynamics, which is crucial for recognising emotional states from speech. The benefits of using Deep Learning for SER include achieving state-of-the-art results on the IEMOCAP database for speaker-independent SER, demonstrating the models' ability to generalise across different speakers. Furthermore, these models present a simple pipeline and low latency in emotion recognition tasks, which are advantageous for real-time applications. The exploration of feed-forward and recurrent neural network architectures, along with their variants, illuminates their strengths and limitations in handling paralinguistic elements of speech, offering insights into the future development of more sophisticated and efficient SER systems (Fayek, H.M., Lech, M. & Cavedon, L., 2017).

Pandey et al. (2019) explored the use of deep learning techniques in recognizing emotions from speech, highlighting the significance of enabling machines not only to understand the content but also to interpret the emotional nuances of human speech. Unlike traditional machine learning models, which are limited in their ability to capture the complexities of emotional states, deep learning offers a more effective approach by automatically learning high-level features from data.

In their evaluation, Pandey et al. (2019) used "Weighted Accuracy" to compare different deep learning architectures and input features across four basic emotions: Neutral, Angry, Happy, and Sad. Their findings indicated that the best results were obtained using CNN, LSTM, and a hybrid model combining CNN and BLSTM. To ensure the robustness of their findings and mitigate any potential bias from a single dataset, they utilized a five-fold cross-validation approach with two datasets, Emo-DB and IEMOCAP. The hybrid model (CNN+BLSTM) with Mel-Frequency Cepstral Coefficients (MFCC) as the input feature achieved the highest accuracy (82.35%) on the Emo-DB dataset. This model, however, faced challenges in accurately classifying happy emotions, attributed to class imbalances and the arousal similarities between happy and angry utterances. Conversely, for the IEMOCAP dataset, the same hybrid model with a Mel-Spectrogram input feature performed best. Interestingly, it accurately classified "happy" emotions more effectively than others, which Pandey et al. (2019) suggest may be due to the more natural elicitation of emotions in the IEMOCAP dataset, despite an overall accuracy below 50%.

|  |
| --- |
|  |
| ***Fig:*** *Confusion Matrices of Pandey et al. (2019) evaluations of both datasets across 4 emotions in a five-fold validation.* |

### Machine Learning Models

Machine learning models play a pivotal role in the development of Speech Emotion Recognition (SER), which aims to decipher human emotions from spoken language. Traditional machine learning models, distinct from their deep learning counterparts, bring to SER a mix of capabilities and challenges. The efficacy of these models (Support Vector Machines (SVM), Hidden Markov Models (HMM), K-nearest neighbours (KNN) and more) in interpreting emotional nuances in speech is underscored by their specific strengths and limitations. These models' contributions to SER, juxtaposed against their operational drawbacks, illustrate the nuanced landscape of machine learning applications in understanding human affective states through speech.

Support Vector Machines (SVM) are renowned for their effectiveness in classification tasks, including SER. SVMs operate by finding the optimal hyperplane that separates different emotion classes in a feature space. Lin and Wei (2005) utilized SVM in their SER system, demonstrating its capability to classify emotional states with high accuracy (Lin and Wei, 2005). However, the model is primarily binary and struggles with multi-class emotion datasets. It also suffers from long processing times and decreased accuracy in the presence of background noise (Anusha et al., 2021). Despite these drawbacks, the model's high accuracy in binary classifications positions it as a valuable tool in SER where binary emotion classification is sufficient.

Hidden Markov Models (HMM) offer another approach by modelling the temporal sequence of speech features. HMMs are particularly adept at capturing the dynamic nature of speech, making them suitable for SER applications. Aouani and Ayed (2020) demonstrated the use of HMM in SER, achieving impressive recognition rates by capturing the temporal dynamics of speech features (Aouani and Ayed, 2020). However, HMMs face challenges in feature selection, as the selected features may not fully represent the emotional state conveyed in speech. Moreover, the complexity of HMMs can lead to increased computational requirements.

K-nearest neighbors (KNN) is a simpler model that has been applied to SER. Its main advantage lies in its simplicity and interpretability. KNN classifies emotions based on the closest training examples in the feature space. While this model is easy to implement and understand, its accuracy is generally lower compared to more complex models like SVM and HMM. Additionally, KNN's performance is highly dependent on the choice of k and the distance metric used, which may require fine-tuning for optimal performance in SER tasks (Anusha et al., 2021).

Each of these machine learning models brings unique strengths to SER. SVMs are highly effective in binary classifications and can achieve high accuracy levels. HMMs excel in capturing the temporal dynamics of speech, making them suitable for analysing the time-dependent aspects of emotions in speech. Meanwhile, KNN offers simplicity and ease of interpretation, although it may not always achieve the highest accuracy.

However, these models also face limitations. The binary nature of SVMs limits their application in multi-class emotion recognition tasks, and their performance can be significantly affected by noise. HMMs, while powerful, can be complex and computationally demanding. KNN's performance is variable and often inferior to more sophisticated models.

### Ensemble Models

## Challenges and Limitations

### Datasets

When developing a Speech Emotion Recognition model in Artificial Intelligence, one faces several challenges related to the datasets used. The quality of the dataset is paramount; poor audio quality, background noise, or poorly labelled data can significantly hinder model performance. A model can only be as good as the data it learns from, making high-quality datasets essential for accurate emotion classification.

The size of the dataset is another critical factor. Training robust models requires large amounts of data to capture the variability in speech patterns. However, collecting and curating such vast datasets is resource-intensive and often impractical. Small datasets may lead to overfitting, where the model performs well on training data but poorly on unseen data.

Dataset metadata, including information about the speakers (such as age, gender, and native language), recording conditions, and emotional states, is crucial for understanding and contextualizing the data. Lack of detailed metadata can prevent the model from learning nuanced differences in speech that may be influenced by these factors.

Feature extraction is a challenge, as determining which features of the speech are most relevant for emotion recognition is not straightforward. The extraction and selection of features such as pitch, tone, speed, and pauses directly impact the model's ability to learn and classify emotions accurately.

The generalization complexity of datasets refers to their ability to represent real-world variability. Datasets often lack diversity, focusing on specific demographics or languages, which limits the model's ability to generalize across different populations and situations.

Lastly, datasets that do not include facial expressions or other non-verbal cues present a unique challenge. In real-world applications, emotional recognition often relies on a combination of verbal and non-verbal cues. Relying solely on speech data may limit the model's effectiveness in accurately classifying emotions, as it misses out on the rich context provided by facial expressions and body language.

Addressing these challenges requires careful dataset selection, augmentation strategies, and feature engineering to develop a Speech Emotion Recognition model capable of accurately classifying emotions across diverse conditions and populations.

#### CREMA-D

The CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) dataset, created by Cao et al. (2014), is a rich resource developed to enable detailed study and analysis in the field of speech emotion recognition. Comprising 7,442 video clips from 91 actors, including 48 men and 43 women, ranging in age from 20 to 74 years and representing a wide array of ethnic backgrounds, it is designed to reflect the diversity of emotional expression and perception. These actors were recorded expressing six basic emotions: happiness, sadness, anger, fear, disgust, and neutral, with variations in emotional intensity, providing a nuanced spectrum of emotional states for analysis.

Each clip in the dataset features a spoken dialogue designed to fit within a neutral context, allowing the emotional expression to be the focal point, free from biases that might arise from specific situational contexts. This aspect is crucial for creating a controlled environment for emotion recognition tasks, where the emphasis is on the tone, pitch, and modulation of the voice rather than the content of the speech itself.

An innovative aspect of CREMA-D is its approach to labelling emotions. Each clip received emotion and intensity labels not just from the actors or a small group of experts, but from a large crowd-sourced platform involving 2,443 raters. This approach provides a broad, democratically sourced perspective on the emotional content of each clip, contributing to the dataset's reliability and applicability across different demographic and cultural groups.

For researchers focusing specifically on audio-based emotion recognition, the audio files extracted from these clips are of primary interest. The audio dimension of CREMA-D includes a wide range of vocal expressions, offering a valuable dataset for training and testing machine learning and deep learning models. The diversity in the actors' performances, coupled with the varied emotional intensities and the comprehensive labelling, makes CREMA-D a powerful tool for advancing research in speech emotion recognition.

By providing a dataset that not only captures a wide spectrum of emotions and intensities but also represents a diverse cross-section of age, gender, and ethnicity, CREMA-D facilitates a more inclusive understanding of emotional expression and recognition in speech. This inclusivity is key for developing more accurate and universally applicable emotion recognition systems.

#### RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), as provided by Livingstone and Russo (2018), stands out in the field of speech emotion recognition due to its detailed and well-structured approach to capturing a wide range of emotional expressions through both speech and song. This dataset's uniqueness lies in its dual modality, offering both audio and visual emotional expressions, enabling researchers to study the impact of multimodal inputs on emotion recognition systems.

With a total of 24 professional actors (12 male and 12 female) contributing to the dataset, RAVDESS provides a diverse and comprehensive collection of emotional expressions. These actors perform two lexically-matched statements across a variety of emotions—calm, happy, sad, angry, fearful, surprise, and disgust—at two levels of intensity, plus a neutral baseline. This range ensures that the dataset covers a broad spectrum of emotional states, facilitating the development of nuanced emotion recognition models.

Each recording in the dataset is meticulously rated by North American participants for three key attributes: emotional validity, intensity, and genuineness. This rigorous evaluation process ensures the reliability and accuracy of the dataset, making it an invaluable resource for training and benchmarking machine learning models in the field of emotion recognition.

The RAVDESS dataset is not only notable for its size, with 7356 recordings, but also for its accessibility and ethical considerations. It is freely available for use in research, provided that it is properly credited, making it an accessible resource for researchers worldwide. Additionally, the creators of RAVDESS have placed a strong emphasis on ethical considerations, including the informed consent of all participants and the respectful treatment of the data, ensuring that the dataset can be used responsibly in research.

#### SAVEE

The Surrey Audio-Visual Expressed Emotion (SAVEE) Database, created by Jackson and Haq (2014), serves as a foundational resource in the domain of speech emotion recognition, a nuanced area of Artificial Intelligence (AI) that focuses on discerning human emotions through speech. This database is specifically designed to facilitate the development and evaluation of automatic emotion recognition systems. It encompasses audio-visual recordings of four male actors, capturing a spectrum of seven different emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral, resulting in a total of 480 British English utterances. These utterances were meticulously selected from the standard TIMIT corpus to ensure a phonetic balance is maintained across the various emotions represented.

The recording process of the SAVEE database was conducted in a visual media lab equipped with state-of-the-art audio-visual recording technologies. This process placed a strong emphasis on capturing phonetically balanced sentences, incorporating a diverse array of emotional expressions to enhance the dataset's effectiveness for emotion recognition tasks. To validate the quality and applicability of the dataset, ten subjects evaluated the recordings under three different conditions: audio, visual, and audio-visual. These evaluations yielded speaker-independent recognition rates of 61%, 65%, and 84%, respectively, highlighting the database's reliability and potential in developing emotion recognition systems.

The SAVEE database includes a broad spectrum of emotions and high-quality recordings, making it a critical tool for researchers in the emotion recognition field. The dataset's specifications feature recordings from four native English male speakers, aged between 27 to 31 years, thereby ensuring a diversity of vocal characteristics. It includes six basic emotions plus a neutral state, encompassing a comprehensive range of human emotional states across 480 utterances. The focus on high recording quality aims to facilitate precise emotion recognition and analysis.

Given its extensive coverage of emotions, meticulous recording quality, and thorough evaluation, the SAVEE database is indispensable for researchers striving to push the boundaries of speech emotion recognition. It supports not only the exploration of audio and visual modalities in isolation but also fosters research into multimodal approaches, which have demonstrated enhanced accuracy in recognizing human emotions. As such, the SAVEE database significantly contributes to advancing AI's capability to interpret human emotions through speech, paving the way for more intuitive and natural human-computer interactions.

#### EMO-DB

The Emo-DB dataset, developed during a DFG-funded research project between 1997 and 1999 by Burkhardt et al. (2005), is a collection of emotional utterances spoken by actors. These recordings were made in the anechoic chamber of the Technical University Berlin, under the guidance of Prof. Dr. W. Sendlmeier, with contributions from Felix Burkhardt, Miriam Kienast, Astrid Paeschke, and Benjamin Weiss. The dataset includes over 500 utterances, categorized by emotions, which are available for research purposes. Users can filter these utterances based on the speaker, text, and emotion through a web interface, which also provides access to syllable labels, duration information, intonation contours, and results from various perception tests.

The structure of Emo-DB allows for detailed analysis of emotional speech, offering tools to analyse fundamental frequency, energy, loudness, duration, stress, and rhythm measurements. The dataset is not only a resource for sound files but also for label files including syllable and phone labels, along with results from perception tests assessing emotion recognition, naturalness evaluation, syllable stress, and emotional intensity. This comprehensive setup aids researchers in understanding how emotional expressions are perceived and processed.

Access to Emo-DB is facilitated through its website, where researchers can download audio and label files for analysis, ensuring the source is correctly cited. The dataset's coding scheme includes a naming convention that integrates the speaker's number, text code, and emotion, providing a systematic approach to data organization. Emo-DB covers a wide range of emotions such as happiness, sadness, anger, fear, and neutral expressions, making it a valuable tool for the study of speech emotion recognition.

The creation and availability of Emo-DB mark a significant contribution to the field of speech emotion recognition, enabling detailed studies of emotional expression in speech. Its structured format, comprehensive content, and accessibility make it a key resource for researchers aiming to analyse and understand emotional expressions in speech, offering a foundation for the development and testing of machine learning and deep learning models in speech emotion recognition.

### Emotions

The diversity and complexity of different and various range of emotions have a significant impact and impose a great obstacle for any model that tries to classify any emotions from an audio dataset. The human complexity of conveying an emotion from diverse ranges of speech pace, linguistics, dialects, and nuances, and the overall combination of them that can be used for similar emotional responses that sound like the direct opposite, seems to be a more improbable mission to dissect and classify it to its true emotion class.

### Authenticity

### Cultural and Gender biases

### Subjectivity and Variability

### Overfitting

NOTE: (PAINDEY ET AL.) The “Conclusion” wraps up by reinforcing the review’s scope, which scrutinized various feature inputs such as Magnitude Spectrogram, Log-Mel Spectrogram, and MFCCs against different architectures to deduce the optimal feature-architecture synergy. Experiments on Emo-DB and IEMOCAP revealed a preference for the Log-Mel Spectrogram feature when paired with the CNN+LSTM architecture. The study acknowledges the challenge of overfitting, especially with limited data as observed with Emo-DB, and details the incorporation of regularization strategies like Dropout and Batch Normalization to mitigate this issue. The article stands firm on the potential of deep learning in SER, underlining the significance of model and feature selection tailored to the complexities of emotional speech data.

# Methodology

## Feature Extraction

## Augmentation

## Model Constructions

### Deep Learning Models – Vector Features

#### ANN

#### CNN

#### MLP

#### ResNet

#### RNN+LSTM

#### Transformer

### Machine Learning Models – Vector Features

#### GMM

#### HMM

#### KNN

#### RF

#### SVM

### Deep Learning Models – Image Features

#### CNN

#### DenseNet

#### Inception

#### ResNet

#### VGG

### Ensemble Models

#### Lorem Ipsum 1

#### Lorem Ipsum 2

#### Lorem Ipsum 3

#### Lorem Ipsum 4

#### Lorem Ipsum 5

## Performance

### Datasets & Features

Vector Features

* Models tend to perform better on EMO-DB and SAVEE datasets compared to CREMA-D and RAVDESS when using Mel-Spectrograms.
* ANN, MLP, and SVM generally show better accuracy across the datasets when using Mel-Spectrograms.
* There is a noticeable drop in performance for most models when using MFCC features compared to Mel-Spectrograms, especially on the CREMA-D and EMO-DB datasets.
* HMM and GMM models tend to perform less effectively than other models across both feature types and datasets.
* The best performing models with Mel-Spectrogram features on the SAVEE dataset are ANN and SVM, achieving 0.81 accuracy, macro, and weighted averages.

**Mel-Spectrogram Vector Metrics with Emphasis on F1 Macro Average:**

1. **CREMA-D:**
   * **MLP** stands out with the highest F1 Macro Average at 0.53.
   * **ANN** and **SVM** are close contenders with Macro Averages of 0.47 and 0.48, respectively.
   * **HMM** shows significantly lower performance with a Macro Average of 0.14.
2. **EMO-DB:**
   * **MLP** leads again with the highest F1 Macro Average at 0.8.
   * **ANN** and **SVM** are strong performers as well, both with a Macro Average of 0.79.
   * **HMM** lags behind considerably with a Macro Average of 0.12.
3. **RAVDESS:**
   * **GMM** surprisingly has the highest F1 Macro Average at 0.67, which is a unique occurrence across the datasets.
   * **ResNet** also performs well with a Macro Average of 0.67.
   * **HMM** has the lowest Macro Average at 0.15.
4. **SAVEE:**
   * **ANN** has the best F1 Macro Average at 0.8.
   * **MLP** and **SVM** also show good performance with Macro Averages of 0.73 and 0.69, respectively.
   * **HMM** again has the lowest Macro Average at 0.11.

**MFCC Vector Metrics with Emphasis on F1 Macro Average:**

1. **CREMA-D:**
   * All models perform moderately, with the highest F1 Macro Average being 0.44 for **MLP**.
   * **ANN** and **RF** have similar Macro Averages of 0.43.
   * **HMM** has the lowest Macro Average at 0.2.
2. **EMO-DB:**
   * **ANN** leads with an F1 Macro Average of 0.71.
   * **CNN** and **SVM** follow closely with Macro Averages of 0.7 and 0.67, respectively.
   * **HMM** and **GMM** perform poorly on this dataset with Macro Averages of 0.16 and 0.27, respectively.
3. **RAVDESS:**
   * **ANN** and **ResNet** show the best Macro Averages of 0.57 and 0.56, respectively.
   * Most other models hover around the 0.5 mark for the Macro Average.
   * **HMM** shows the lowest performance with a Macro Average of 0.1.
4. **SAVEE:**
   * **ANN** and **CNN** lead with Macro Averages of 0.55 and 0.59, respectively.
   * The performance is more evenly distributed among the models in this dataset with MFCC features.
   * **HMM** again scores the lowest with a Macro Average of 0.14.

**Overall Analysis:**

When using the F1 Macro Average as the key metric:

* **Mel-Spectrogram Features**: MLP consistently emerges as the top performer across most datasets, indicating its robustness and ability to handle the detailed information captured in Mel-Spectrogram features. ANN and SVM also show strong performance, making them suitable choices for tasks requiring high F1 Macro Averages. HMM consistently shows low performance, suggesting that it struggles with Mel-Spectrogram features for emotion recognition tasks.
* **MFCC Features**: The performance drop when using MFCC features is noticeable compared to Mel-Spectrograms, but ANN still leads in most cases, with respectable F1 Macro Averages, particularly on the EMO-DB dataset. Other models like CNN and ResNet also perform well, though not as consistently as ANN.

In conclusion, while accuracy is a straightforward metric, the F1 Macro Average provides a more nuanced view, especially in imbalanced datasets. It can significantly alter the perception of model performance, highlighting the importance of considering multiple metrics when evaluating machine learning models. MLP stands out with Mel-Spectrograms, and ANN does well with both feature types, showing overall robustness in handling emotion recognition tasks.

|  |
| --- |
|  |
| Fig: Comparisons of all 4 datasets F1 Macro-Average Performance Metrics, in Deep Learning and Machine Learning models, using Mel-Spectrogram as a Vector Feature |

|  |
| --- |
|  |
| Fig: Comparisons of all 4 datasets F1 Macro-Average Performance Metrics, in Deep Learning and Machine Learning models, using MFCC as a Vector Feature |

|  |
| --- |
|  |
| Fig: F1-Score Macro-Average Performance Comparisons of Each Deep Learning Model, from each Dataset, relative to each Feature Vector (MFCC vs Mel-Spectrogram) |

|  |
| --- |
|  |
| Fig: F1-Score Macro-Average Performance Comparisons of Each Machine Learning Model, from each Dataset, relative to each Feature Vector (MFCC vs Mel-Spectrogram) |

### Images

### General Observations:

### Model Performance Across Resolutions: Higher resolutions generally lead to better model performance, evident in the increased F1 Macro Average scores as we move from 32x32 to 256x256 resolution. This suggests that higher resolution images capture more detailed information that is beneficial for emotion recognition tasks.

### MFCC vs. Mel-Spectrogram Features: The performance varies significantly between using MFCC and Mel-Spectrogram features, with Mel-Spectrogram features generally leading to better performance across most datasets and resolutions. This could indicate that Mel-Spectrograms provide a richer representation of the audio data for emotion recognition tasks.

### Model Consistency: CNN, DenseNet, and ResNet models show relatively consistent improvement as the resolution increases, while VGG and Inception models display mixed results across different datasets and feature types.

### Detailed Analysis by Dataset and Resolution:

### 32x32 Resolution

### Best Performing Model: CNN with Mel-Spectrogram on the EMODB dataset, achieving an F1 Macro Average of 0.56.

### Worst Performing Model: Several models on the SAVEE dataset using MFCC features, with F1 Macro Averages as low as 0.12 (ResNet).

### 64x64 Resolution

### Improvement Noted: Significant improvements across all datasets when comparing to 32x32 resolution, especially for the CNN model on the RAVDESS dataset using Mel-Spectrograms (F1 Macro Average: 0.62).

### MFCC Feature Note: The CNN model shows a notable improvement using MFCC features on the RAVDESS dataset (F1 Macro Average: 0.59).

### 128x128 Resolution

### Best Performance Jump: The EMODB dataset shows a substantial improvement, with the CNN model reaching an F1 Macro Average of 0.71 using Mel-Spectrograms.

### MFCC Notable Performance: The DenseNet model using MFCC features on the EMODB dataset improved drastically to an F1 Macro Average of 0.6.

### 256x256 Resolution

### Highest Performances: Notable performances include the ResNet model on the CREMA-D dataset using Mel-Spectrograms (F1 Macro Average: 0.58) and the DenseNet model on the EMODB dataset using both feature types (F1 Macro Average: 0.68 with MFCC).

### Mixed Results for MFCC: The MFCC features showed mixed results at this resolution, with notable declines in performance for some models (e.g., CNN on CREMA-D with an F1 Macro Average of 0.05).

### Conclusion:

### The analysis reveals a clear trend where increasing image resolution generally improves model performance for speech emotion recognition tasks. Mel-Spectrogram features consistently outperform MFCC features, suggesting they capture more relevant information for distinguishing between different emotions in audio data. Additionally, models like CNN, DenseNet, and ResNet show promising adaptability and performance improvement with increased resolution, indicating their potential for future research and application in speech emotion recognition tasks.

|  |  |
| --- | --- |
|  |  |
| *Fig: Comparison charts of Mel-Spectrogram and MFCC images, independently, using F1-Score Macro Average metric on 32x32 resolution images* | |

|  |
| --- |
|  |
| *Fig: Comparisons of all model’s F1-Score Macro Average, on all Datasets, against Mel-Spectrogram and MFCC together, on 32x32 resolution* |

|  |  |
| --- | --- |
|  |  |
| *Fig: Comparison charts of Mel-Spectrogram and MFCC images, independently, using F1-Score Macro Average metric on 64x64 resolution images* | |

|  |
| --- |
|  |
| *Fig: Comparisons of all model’s F1-Score Macro Average, on all Datasets, against Mel-Spectrogram and MFCC together, on 64x64 resolution* |

|  |  |
| --- | --- |
|  |  |
| *Fig: Comparison charts of Mel-Spectrogram and MFCC images, independently, using F1-Score Macro Average metric on 128x128 resolution images* | |

|  |
| --- |
|  |
| *Fig: Comparisons of all model’s F1-Score Macro Average, on all Datasets, against Mel-Spectrogram and MFCC together, on 128x128 resolution* |

|  |  |
| --- | --- |
|  |  |
| *Fig: Comparison charts of Mel-Spectrogram and MFCC images, independently, using F1-Score Macro Average metric on 256x256 resolution images* | |

|  |
| --- |
|  |
| *Fig: Comparisons of all model’s F1-Score Macro Average, on all Datasets, against Mel-Spectrogram and MFCC together, on 256x256 resolution* |

## Model Selection

Considering the overall performance across different resolutions, feature types (Mel-Spectrograms and MFCCs), and datasets, the **CNN (Convolutional Neural Network)** model emerges as the best-performing model with the most significant potential for achieving higher quality through further feature enhancement and hyperparameter tweaking. Here are key reasons why the CNN model stands out:

1. **Consistency Across Resolutions and Datasets:** The CNN model consistently ranks among the top performers across various resolutions (32x32 to 256x256) and datasets (CREMA-D, EMODB, RAVDESS, SAVEE). This consistency is a strong indicator of the model's robustness and adaptability to different types of input data and resolutions.
2. **High Performance with Mel-Spectrograms:** The CNN model particularly shines when utilizing Mel-Spectrogram features, achieving the highest F1 Macro Average scores in several cases. For instance, it achieved notable F1 scores on the EMODB dataset at higher resolutions (e.g., 0.71 at 128x128). This suggests that CNNs are highly effective at extracting relevant features from complex image representations of audio data for emotion recognition tasks.
3. **Room for Improvement and Adaptation:** Given the model's architectural flexibility, CNNs can significantly benefit from additional features, advanced preprocessing techniques, and hyperparameter optimization. CNNs are known for their capacity to handle higher-dimensional data, making them suitable for integrating multiple feature types beyond Mel-Spectrograms and MFCCs, potentially leading to further improvements in performance.
4. **Potential for Feature Enhancement:** The inherent design of CNNs allows for the exploration of deeper and more complex network architectures. By applying techniques such as transfer learning, fine-tuning, and layer augmentation, there's substantial scope to enhance feature extraction and learning capabilities, thereby improving emotion recognition accuracy.
5. **Hyperparameter Tweaking:** CNNs offer a wide range of hyperparameters that can be adjusted, including the number of layers, filter sizes, and learning rates. Systematic hyperparameter tuning, possibly through automated approaches like grid search or Bayesian optimization, could unlock higher performance levels.

In summary, the CNN model exhibits a strong foundation for speech emotion recognition tasks across different conditions. Its overall performance, coupled with its architectural flexibility, makes it a prime candidate for further research and development efforts aimed at enhancing its capabilities through additional features, hyperparameter tweaking, and advanced preprocessing techniques.

**Model: ResNet**

* **Rationale:**
  + **Vector Inputs:** For vector inputs, ResNet shows commendable performance across the datasets, particularly excelling in the RAVDESS dataset with Mel-Spectrogram features.
  + **Image Inputs:** When it comes to image inputs, ResNet consistently stands out, especially at higher resolutions (128x128, 256x256), across multiple datasets. Its performance notably improves with image resolution, suggesting its effectiveness in capturing complex features within images.
  + **Versatility:** ResNet demonstrates a versatile capacity to handle both vector and image inputs effectively, making it a strong candidate for exploring hybrid models that combine vector and image features.

### Dataset Selection

**1. EMO-DB:**

* **Vector Inputs:** Shows one of the highest accuracies across models for both Mel-Spectrogram and MFCC features, indicating the dataset's clarity and distinguishability for emotion recognition.
* **Image Inputs:** Performance remains strong, especially at higher resolutions, highlighting the dataset's compatibility with convolutional architectures and its potential for yielding high accuracy in emotion recognition tasks.

**2. SAVEE:**

* **Vector Inputs:** Similar to EMO-DB, SAVEE exhibits high model accuracies for vector inputs, especially with Mel-Spectrogram features, making it one of the top-performing datasets in this category.
* **Image Inputs:** While there's a variability in model performance at different resolutions, SAVEE still shows promising results, particularly at 256x256 resolution, indicating the dataset's rich information content for image-based feature extraction and recognition.

### Summary and Justification

* **Model (ResNet):** Chosen for its strong performance and adaptability across both vector and image inputs, particularly at higher resolutions. Its architectural depth enables it to capture complex patterns in the data, making it ideal for future explorations into combining vector and image features.
* **Datasets (EMO-DB and SAVEE):** Selected based on their consistent high performance across both vector and image input features. These datasets not only show strong results with current models but also offer a balanced challenge for exploring advanced hybrid models due to their diverse and clear emotional expressions.

This selection strategy aims to leverage the strengths of ResNet in handling complex image features while focusing on datasets that have demonstrated clear and distinguishable emotional features across both input types. The next step involves exploring the combination of vector and image features within ResNet for EMO-DB and SAVEE datasets to push the boundaries of Speech Emotion Recognition further.

# Own Approach

## Lorem Ipsum 1

## Lorem Ipsum 2

## Lorem Ipsum 3

## Lorem Ipsum 4

## Lorem Ipsum 5

# Implementation

## Lorem Ipsum 1

## Lorem Ipsum 2

## Lorem Ipsum 3

## Lorem Ipsum 4

## Lorem Ipsum 5

# Results

## Lorem Ipsum 1

## Lorem Ipsum 2

## Lorem Ipsum 3

## Lorem Ipsum 4

## Lorem Ipsum 5

# Discussion

## Lorem Ipsum 1

## Lorem Ipsum 2

## Lorem Ipsum 3

## Lorem Ipsum 4

## Lorem Ipsum 5

# Conclusion

## Lorem Ipsum 1

## Lorem Ipsum 2

## Lorem Ipsum 3

## Lorem Ipsum 4

## Lorem Ipsum 5

**References:**

* + Chandrasekar, P., Chapaneri, S. and Jayaswal, D., 2014, April. Automatic speech emotion recognition: A survey. In 2014 international conference on circuits, systems, communication and information technology applications (CSCITA) (pp. 341-346). IEEE.
  + Wani, T.M., Gunawan, T.S., Qadri, S.A.A., Kartiwi, M. and Ambikairajah, E., 2021. A comprehensive review of speech emotion recognition systems. IEEE access, 9, pp.47795-47814.
  + Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M.A. and Cleder, C., 2019. Automatic speech emotion recognition using machine learning.
  + Swain, M., Routray, A. and Kabisatpathy, P., 2018. Databases, features and classifiers for speech emotion recognition: a review. International Journal of Speech Technology, 21, pp.93-120.
  + Fayek, H.M., Lech, M. and Cavedon, L., 2017. Evaluating deep learning architectures for speech emotion recognition. Neural Networks, 92, pp.60-68.
  + Pandey, S.K., Shekhawat, H.S. and Prasanna, S.M., 2019, April. Deep learning techniques for speech emotion recognition: A review. In 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA) (pp. 1-6). IEEE.
  + Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A. and Verma, R., 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing, 5(4), pp.377-390.
  + Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F. and Weiss, B., 2005, September. A database of German emotional speech. In Interspeech (Vol. 5, pp. 1517-1520).
  + Livingstone, S.R. and Russo, F.A., 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), p.e0196391.

**Acknowledgments:**

* Acknowledge any individuals, organizations, or resources that contributed to your research.

**Appendices:**

* Include any additional material such as code snippets, detailed model architectures, and supplementary information.