R

A black and white rectangle

Description automatically generatedA black and white rectangle

Description automatically generated

**Using RGB-Encoded Audio Features for Convolutional Models**

**Speech Emotion Recognition (S.E.R)**

***KV6003 – Individual Computing Project***

***Panagiotis Theodorou – W20020633***

***Supervisor: Dr Hua Mao***

***2nd Marker: Dr Nic Whitton***

***Word Count: XXXX***

***Date: DD/MM/YYYY***

**2023 - 2024**

**DECLARATION OF AUTHORSHIP**

**DECLARATIONS**

I declare the following:

(1) that the material contained in this dissertation is the end result of my own work and that due acknowledgement has been given in the bibliography and references to **ALL** sources be they printed, electronic or personal.

(2) the Word Count of this Dissertation is ...................................

(3) that unless this dissertation has been confirmed as confidential, I agree to an entire electronic copy or sections of the dissertation to being placed on the eLearning Portal (Blackboard), if deemed appropriate, to allow future students the opportunity to see examples of past dissertations. I understand that if displayed on eLearning Portal it would be made available for no longer than five years and that students would be able to print off copies or download.

(4) I agree to my dissertation being submitted to a plagiarism detection service, where it will be stored in a database and compared against work submitted from this or any other School or from other institutions using the service.

In the event of the service detecting a high degree of similarity between content within the service this will be reported back to my supervisor and 2nd marker, who may decide to undertake further investigation that may ultimately lead to disciplinary actions, should instances of plagiarism be detected.

(5) I have read the Northumbria University/Engineering and Environment Policy Statement on Ethics in Research and Consultancy and I confirm that ethical issues have been considered, evaluated and appropriately addressed in this research.

SIGNED:

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 Input Features 00](#_Toc161174022)

[3.1 Experimentation 00](#_Toc161174024)

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

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

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

[3.2 Performance 00](#_Toc161174049)

[3.2.1 Datasets & Features 00](#_Toc161174050)

[3.2.1.1 Vector Features 00](#_Toc161174027)

[3.2.1.2 Image Features 00](#_Toc161174027)

[3.3 Model Selection 00](#_Toc161174051)

[3.3.1 Model Selection Justification: CNN Family Models 00](#_Toc161174050)

[3.3.2 Datasets Selection Rationale: EMODB & RAVDESS 00](#_Toc161174050)

[3.3.3 Optimisation 00](#_Toc161174050)

[4 Methodology 00](#_Toc161174052)

[4.1 Data Split and Structure 00](#_Toc161174053)

[4.2 Audio Feature Extraction and Conversion 00](#_Toc161174054)

[4.3 Re-Imagined Datasets 00](#_Toc161174055)

[4.4 Augmentation 00](#_Toc161174056)

[4.5 Classification Models 00](#_Toc161174057)

[4.6 Genetic Algorithm and Optimisation 00](#_Toc161174057)

[4.6.1 Hyperparameters 00](#_Toc161174008)

[4.6.1 Layers & Weights 00](#_Toc161174008)

[5 Results 00](#_Toc161174058)

[5.1 Frozen Layers 00](#_Toc161174059)

[5.2 Unfrozen Layers 00](#_Toc161174060)

[5.3 Best 00](#_Toc161174061)

[5.4 Lorem Ipsum 00](#_Toc161174062)

[5.5 Lorem Ipsum 00](#_Toc161174063)

[6 Discussion 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 Conclusion 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)

**Title: Speech Emotion Recognition: Using RGB-Encoded Audio Features for Convolutional Architectural 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.

# Input Features

## Experimentation

### Deep Learning Models – Vector Features

### Machine Learning Models – Vector Features

### Deep Learning Models – Image Features

## Performance

### Datasets & Features

#### Vector Features

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

In assessing the effectiveness of various artificial intelligence models for Speech Emotion Recognition (SER), a detailed analysis was conducted using vector input features of Mel-Spectrogram and MFCC across four datasets: CREMA-D, EMO-DB, RAVDESS, and SAVEE. The F1-Score "Macro Average" served as a primary metric for this evaluation, given its relevance in reflecting model performance across diverse emotional expressions within these datasets.

The findings indicate a general trend where models tend to exhibit superior performance on the EMO-DB and SAVEE datasets when utilizing Mel-Spectrogram features. Artificial Neural Networks (ANN), Multi-Layer Perceptron (MLP), and Support Vector Machines (SVM) demonstrated enhanced accuracy across these datasets. This trend suggests that Mel-Spectrogram features, which offer a detailed representation of the spectral texture of sounds, align well with the models' capabilities in discerning nuanced emotional states from speech.

A noticeable performance decline was observed for most models when employing MFCC features, especially pronounced in the CREMA-D and EMO-DB datasets. This reduction could be attributed to the inherent differences in the feature representations, where Mel-Spectrograms provide a granular view that may capture emotional nuances more effectively compared to the potentially more abstracted or compressed information conveyed by MFCCs.

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

Among the models evaluated, Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) consistently underperformed across both feature types and all datasets. This outcome could highlight limitations in these models' ability to leverage the dynamic, temporal aspects of speech essential for accurate emotion recognition, underlining the importance of model selection in conjunction with appropriate feature utilization for SER tasks.

Delving deeper into the Mel-Spectrogram vector metrics with an emphasis on the F1 Macro Average reveals distinct patterns. For instance, in the CREMA-D dataset, MLP distinguished itself with the highest F1 Macro Average, suggesting its effectiveness in leveraging the intricate details captured by Mel-Spectrogram features. On the contrary, HMM displayed significantly lower performance, underscoring its struggles with processing Mel-Spectrogram features for emotion recognition tasks.

Similar observations were made across the other datasets, with MLP consistently emerging as a top performer, showcasing its robustness in handling Mel-Spectrogram features effectively. ANN and SVM also showed commendable performance, positioning themselves as viable alternatives for SER applications requiring high F1 Macro Averages.

Transitioning to the analysis of MFCC vector metrics, while a general performance dip was noted compared to Mel-Spectrograms, ANN managed to maintain a leading position in several instances. This resilience underscores ANN's adaptability to different auditory feature representations, a critical capability for SER systems.

Through this comparative analysis, the crucial role of feature selection in SER becomes evident, with Mel-Spectrogram features proving particularly advantageous for enhancing model performance. Furthermore, the reliance on the F1 Macro Average as a key metric offers a nuanced perspective on model efficacy, particularly valuable in the context of imbalanced datasets. The standout performance of MLP with Mel-Spectrograms and the overall resilience of ANN across different feature types highlight their potential utility in the ongoing development of SER technologies. This analysis not only informs the selection of models and features for SER but also lays the groundwork for future advancements in the field.

#### Image Features

|  |  |  |
| --- | --- | --- |
|  | Mel-Spectrograms | MFCCs |
| CREMA-D |  |  |
|  |  |  |
| EMO-DB |  |  |
|  |  |  |
| RAVDESS |  |  |
|  |  |  |
| SAVEE |  |  |
| Fig: Comparison charts of Mel-Spectrogram and MFCC images, independently, using F1-Score Macro Average metric on all resolutions, for each dataset across all 5 deep learning models. | | |

Evaluating model performance in Speech Emotion Recognition (SER) through image input features, particularly Mel-Spectrograms and MFCCs at varied resolutions (32x32, 64x64, 128x128, 256x256), provides key insights into the influence of image resolution and feature type across four datasets: CREMA-D, EMODB, RAVDESS, and SAVEE. The study focuses on the application of deep learning models to discern trends that could inform future SER methodologies.

A notable trend is the correlation between increased image resolution and enhanced model performance. Higher resolutions capture more detailed information, crucial for the accurate recognition of emotions from audio data. This trend underscores the importance of detailed visual representations in distinguishing complex emotional states, with Mel-Spectrogram features consistently outperforming MFCC features across most datasets and resolutions. Mel-Spectrograms, with their rich time-frequency details, offer a comprehensive view of the audio data's emotional content, suggesting their superiority over the potentially more abstract MFCC features in SER tasks.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **32x32** |  |  | **64x64** |  |
|  |  |  |  |  |
| **128x128** |  |  | **256x256** |  |
| ***Fig:*** *Comparison charts of Mel-Spectrogram and MFCC images,, using F1-Score Macro Average metric on all resolutions independently, across all models, with all datasets together.* | | | | |

The analysis across resolutions reveals:

* At **32x32**, CNN models using Mel-Spectrograms particularly on the EMODB dataset, exhibit superior F1 Macro Averages. This resolution also highlights the challenges faced by models using MFCC features on the SAVEE dataset, indicating the limitations of lower-resolution images and MFCC features in capturing emotional nuances.
* The shift to **64x64** resolution marks a significant improvement in model performance across all datasets, with CNN models using Mel-Spectrograms on the RAVDESS dataset showing noteworthy enhancements. This resolution begins to unveil the potential utility of MFCC features, with improved performance in models like CNN on the RAVDESS dataset.
* At **128x128**, the performance jump is particularly evident on the EMODB dataset with CNN models using Mel-Spectrograms. DenseNet models employing MFCC features on this dataset also experience substantial improvements, showcasing the advantages of increased resolution for both feature sets.
* The highest performances are observed at the **256x256** resolution, with models such as ResNet on the CREMA-D dataset using Mel-Spectrograms and DenseNet on the EMODB dataset for both features achieving remarkable F1 Macro Averages. However, the performance of MFCC features at this resolution is mixed, with some models experiencing performance drops, highlighting the complex interplay between feature type, model architecture, and resolution in SER tasks.

The steady improvement in performance with increased resolution across CNN, DenseNet, and ResNet models emphasizes their efficacy in leveraging high-resolution data for emotion recognition. Conversely, VGG and Inception models exhibit varied performances, indicating a more nuanced applicability in SER endeavours based on the dataset and feature type context.

This detailed examination illuminates the critical roles of image resolution and feature type in SER model performance, offering a foundation for optimizing input feature representation and model architecture in emotion recognition from speech. It also highlights the potential for future research aimed at refining the selection and application of image-based features in SER.

## Selection

### Model Selection Justification: CNN Family Models

In our efforts to advance Speech Emotion Recognition (SER) technology, selecting the right model architecture is crucial. After extensive evaluations, we have chosen to focus on the CNN family, including CNN, DenseNet, and ResNet, due to their robust performance across various tests. These models have demonstrated remarkable consistency and adaptability across different datasets, particularly in the EMO-DB and RAVDESS datasets. They effectively manage varied data structures—from the straightforward arrangement of EMO-DB to the more complex, actor-wise hierarchical organization in RAVDESS. Their ability to maintain high performance across these diverse data arrangements is indicative of their strong generalization capabilities, which are essential for the varied dataset requirements in SER.

These models are particularly adept at processing Mel-Spectrogram features, essential for capturing the nuanced emotional states depicted in speech. This capability is crucial for datasets like RAVDESS, which features a broad spectrum of emotional expressions across different actors. The superior handling of Mel-Spectrogram inputs by CNN, DenseNet, and ResNet across various resolutions underscores their effectiveness in extracting detailed emotional cues, which is pivotal for accurate emotion recognition.

The architectural flexibility of the CNN family also significantly contributes to their selection. These models support deep and complex network structures that are advantageous for the evolving needs of sophisticated SER tasks. For instance, DenseNet's design promotes efficient feature propagation and reuse across its layers, ensuring that essential information is preserved and utilized effectively throughout the network. This feature is particularly beneficial for SER, where capturing every emotional nuance can greatly enhance recognition accuracy. Similarly, ResNet’s ability to address the vanishing gradient problem allows for the construction of deeper networks that learn complex patterns more effectively, without the degradation of performance over depth.

By incorporating the CNN family into our SER analysis, we are not only utilizing the distinct advantages of each model but also exploring the potential of hybrid architectures and ensemble methods. These approaches can integrate the strengths of CNN, DenseNet, and ResNet to achieve superior performance. This strategy ensures that our SER system is robust, adaptable, and capable of handling the complex nuances of human emotions in speech effectively, thus paving the way for significant advancements in the field.

### Dataset Selection Rationale: EMODB & RAVDESS

In the domain of Speech Emotion Recognition (SER), the choice of datasets is as critical as the selection of the model architecture. For our ongoing research, we have strategically chosen EMO-DB and RAVDESS as our primary datasets. These selections are made based on their distinct characteristics, which together provide a comprehensive overview of the model's capabilities in varied contexts.

**EMO-DB**: This dataset is favoured primarily for its high accuracy and clarity, which make it an excellent benchmark for evaluating model performance and feature extraction capabilities. EMO-DB is known for its straightforward file structure, which allows researchers to focus intensively on model tuning and feature extraction methodologies. The simplicity of the dataset structure reduces the complexity typically involved in navigating through more layered data organizations. This characteristic makes EMO-DB an ideal choice for initial model validations and fine-tuning, providing clear insights into the effectiveness of different architectural nuances and feature handling strategies.

**RAVDESS**: In contrast to EMO-DB, RAVDESS introduces a higher level of complexity with its hierarchical and actor-wise organization. This structure challenges the model to maintain high performance across a more diverse and voluminous dataset. Such complexity is crucial for evaluating the model’s scalability and robustness under varied and challenging conditions. Additionally, RAVDESS is characterized by its diverse emotional expressions, with multiple actors portraying a range of emotions. This diversity offers a comprehensive canvas for assessing the model's ability to generalize across different vocal characteristics and emotional intensities, thus providing a robust test of the model’s adaptability and accuracy in real-world scenarios.

### Optimisation

As we progress in enhancing our Speech Emotion Recognition (SER) systems, several innovative experiments and analytical approaches are planned to refine and advance our models. These initiatives are aimed at both augmenting the datasets and optimizing model performance through a series of structured experimental setups. To assess the impact of image data augmentation on model performance, an experiment will be designed. This will involve a comparative analysis of the baseline CNN performance with and against an augmented dataset version. Techniques such as rotation, scaling, and flipping could be explored to determine their efficacy in enhancing the robustness and generalization capabilities of our models.

An automated hyperparameter optimization process will be implemented, utilizing advanced techniques like Genetic Algorithm optimization. The objective here is to systematically explore and identify the optimal configuration of CNN parameters that maximizes performance specifically on the EMO-DB and RAVDESS datasets. This step is crucial for ensuring our models are not only accurate but also efficient and scalable.

Another experimental avenue involves testing the SER models on a reduced emotion spectrum. Instead of the broader range presented in EMO-DB’s 7 and RAVDESS’s 8 emotion categories, we will explore the model's performance with only 4 classes. This will provide insights into the model's effectiveness in a more generalized approach to emotion recognition, which could be beneficial for applications requiring less granularity in emotional distinction.

# Methodology

## Data Split and Structure

|  |
| --- |
|  |
| *Fig:* The Audio dataset being organised into emotion classes subfolders, then converted into images. |

## Audio Feature Extraction and Conversion

|  |
| --- |
|  |
| *Fig: Diagram of how audio files are transformed into image, using feature extraction to extract precisely 3 features, and concatenating them towards each of the Red, Green, and Blue channels (RGB) of the image, to create a unique visual representation of that audio, into a single image.* |

## Re-Imagined Datasets

|  |  |  |  |
| --- | --- | --- | --- |
| Red | Green | Blue | **OUTPUT** |
| **Red** | Chroma | Mel-Spectrogram | MFCC | *CH\_ME\_MF* |
| **Green** | MFCC | Chroma | Mel-Spectrogram | *MF\_CH\_ME* |
| **Blue** | Mel-Spectrogram | MFCC | Chroma | *ME\_MF\_CH* |
| **OUTPUT** | *CH\_MF\_ME* | *ME\_CH\_MF* | *MF\_ME\_CH* |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | CH\_ME\_MF | CH\_MF\_ME | | A close-up of a blue and pink background  Description automatically generated | A green and yellow striped background  Description automatically generated | | MF\_CH\_ME | *MF\_ME\_CH* | | A red and green striped background  Description automatically generated | A red and blue striped background  Description automatically generated with medium confidence | | ME\_CH\_MF | *ME\_MF\_CH* | | A blue and green striped background  Description automatically generated | A green and blue striped background  Description automatically generated | |
| *Fig:* Different variations of 3 Features concatenated into an image from the same audio file, by mapping each feature into an RGB channel. |

## Augmentation

|  |
| --- |
|  |
| *Fig:* Using Pre-Trained model, with the Top Layer removed, to create Feature Maps, and clustering them before one-hot encoding and concatenating the original feature maps, with the clustered labels, to create an Augmented Feature Set for an improved training input. |

**1. Feature Extraction**

**The model used is pre-loaded with weights trained on ImageNet. By setting include\_top=False, the network is configured without its final classification layers, allowing it to serve as a feature extractor.**

**The output from the pre-trained model. feeds into a GlobalAveragePooling2D layer, which reduces each feature map's dimensions from the convolutional layers to a single vector per map. This reduction simplifies the dataset and focuses on the most salient features.**

**2. Clustering as Augmentation**

**To enhance the model's ability to differentiate between similar emotions, KMeans clustering is applied to the extracted features. This technique identifies clusters in the high-dimensional feature space, potentially highlighting subtle differences between images.**

**Cluster labels are one-hot encoded to transform them into a binary format suitable for concatenation with the original feature vectors. This enriched feature set aims to provide the classifier with additional contextual information.**

## Classification Models

The project harnesses the capabilities of several sophisticated CNN architectures to classify emotions from images. Each chosen architecture brings unique advantages crucial for the crucial task of Speech Emotion Recognition (SER). The combination of these models is directed towards enhancing the robustness and accuracy of the emotion classification task, each bringing strengths that complement the others in handling the diversity of emotions represented in the datasets.

* **DenseNet121**

DenseNet121 is notable for its feature reuse capability, which significantly reduces the number of parameters while maintaining depth and accuracy. For SER using audio spectrograms, this efficiency helps in capturing nuanced patterns from audio data efficiently, supporting better emotion recognition with fewer computational resources.

* **EfficientNetB0**

EfficientNetB0 employs a systematic scaling of all dimensions—depth, width, and resolution—which ensures that performance scales effectively with network size increases. In SER, this feature is advantageous when dealing with spectrograms of varying resolutions, facilitating consistent performance across different audio quality levels without losing accuracy.

* **ResNet50**

ResNet50 addresses the vanishing gradient problem through its residual connections, allowing the training of deeper network architectures. For SER, this ability is crucial for learning complex emotional cues from deeper layers of audio spectrograms, ensuring robust performance even with deeper network structures.

* **VGG16**

## The VGG16 architecture, known for its uniform configuration, often serves as a baseline for comparative studies and simplifies the design process of neural networks. In the context of SER, its straightforward layer stacking is particularly effective for feature extraction from audio spectrograms, essential for accurate emotion detection.

## Genetic Algorithm & Optimisation

### Hyperparameters

def initialize\_population(pop\_size):  
 population = []  
 for \_ in range(pop\_size):  
 individual = {  
 'learning\_rate': 10 \*\* np.random.uniform(-5, -1),  
 'batch\_size': np.random.choice([16, 32, 64, 128]),  
 'dense\_neurons': np.random.choice([64, 128, 256, 512]),  
 'activation': np.random.choice(['relu', 'tanh', 'sigmoid', 'leaky\_relu', 'elu']),  
 'dropout\_rate': np.random.uniform(0.01, 0.5),  
 'n\_clusters': np.random.randint(2, 20),  
 'unfreeze': random.choice([True, False]), # Randomly decide to unfreeze or not  
 'layers\_to\_unfreeze': np.random.randint(0, 5) # Random number of layers to unfreeze  
 }  
 population.append(individual)  
 return population

POPULATION\_SIZE = 10  
NUM\_GENERATIONS = 20  
NUM\_PARENTS = 5

### Layers & Weights

# Results

## Frozen Layers

|  |  |
| --- | --- |
| **EMODB** | **RAVDESS** |
|  |  |
|  |  |

## Unfrozen Layers

|  |  |
| --- | --- |
| **EMODB** | **RAVDESS** |
|  |  |
|  |  |

## Best

|  |  |
| --- | --- |
| **EMODB** | |
|  |  |
|  |  |
| ***Fig:*** | |

|  |  |
| --- | --- |
| **RAVDESS** | |
|  |  |
|  |  |
| ***Fig:*** | |

## Per-Variant

|  |  |
| --- | --- |
| **CH\_ME\_MF** | |
|  |  |
| **CH\_MF\_ME** | |
|  |  |
| **ME\_CH\_MF** | |
|  |  |
| **ME\_MF\_CH** | |
|  |  |
| **MF\_CH\_ME** | |
|  |  |
| **MF\_ME\_CH** | |
|  |  |
| ***Fig:*** | |

## Lorem Ipsum 5

# Discussion

* Models confuse Angry for Happy
* Emotion Classes as a Spectrum
  + Main emotions as main classes
  + Other emotions as sub classes
* ResNet 4 class ME\_MF\_CH vs CH\_ME\_MF
* VGG 4 class ME\_CH\_MF vs MF\_ME\_CH
* RAVDESS underperforming
  + Songs contrasting for classifiers

## 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.
  + Aouani, H. and Ayed, Y.B., 2020. Speech emotion recognition with deep learning. Procedia Computer Science, 176, pp.251-260.
  + Lin, Y.L. and Wei, G., 2005, August. Speech emotion recognition based on HMM and SVM. In 2005 international conference on machine learning and cybernetics (Vol. 8, pp. 4898-4901). IEEE.
  + Anusha, R., Subhashini, P., Jyothi, D., Harshitha, P., Sushma, J. and Mukesh, N., 2021, June. Speech emotion recognition using machine learning. In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1608-1612). 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.
  + Jackson, P. and Haq, S. (2014) *Surrey Audio-Visual Expressed Emotion (SAVEE) database*, *Surrey Audio-visual expressed emotion (SAVEE) database*. Available at: http://kahlan.eps.surrey.ac.uk/savee/. https://openresearch.surrey.ac.uk/esploro/outputs/journalArticle/Surrey-audio-visual-expressed-emotion-savee-database/99635364402346

**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.