# Investigating Speech Emotion Recognition Models: Comparative Analysis and Strategies for Accuracy Improvement

**Anonymous Authors**[1]

## Abstract

This paper explores the realm of speech emotion recognition (SER), investigating the influence of data augmentation, model architectures, and dataset characteristics on classification accuracy. We review existing research on SER, highlighting the shift from probabilistic models to deep learning architectures and the challenges posed by the subjective nature of emotions. Our study introduces CNN and parallel CNN-RNN models for SER, examining their performance across various datasets and augmentation techniques. Our results demonstrate that data augmentation techniques, particularly pitch manipulation, significantly improve classification accuracy, particularly within the CNN framework. Additionally, interpretation techniques such as Grad-CAM provide insights into the models' decision-making processes, revealing patterns in audio features associated with different emotions. Finally, we scrutinize the impact of gender bias and emotion intensity on model performance, revealing that models exhibit higher classification accuracy under conditions of heightened emotion intensity and when the speaker is female.

## 1. Introduction

In the realm of human communication, emotions play a fundamental role, serving as the essence of our interactions and profoundly influencing the dynamics of relationships, decision-making, and overall well-being. Verbal communication comprises two key components: linguistic and paralinguistic information. The former concerns the grammar, syntax and literal meaning of words; the latter refers to non-verbal aspects of speech, including tone of voice, pitch, volume, rythm and pauses, which convey additional information beyond the literal meaning. In these details, emotions can be detected even if the words used are neutral.

Speech-based Emotion Recognition (SER) has emerged as a new, expansive, and challenging problem in recent decades. The growing interest around this topic is due to its potential applications. For example, in the field of Human-Computer Interactions (HCI), SER models can enhance users experiences: devices, such as voice assistants and smart technologies, can adapt their responses in real-time based on users' emotional states, providing more empathetic and responsive interactions. Moreover, SER models are finding utility across several sectors, from medical applications (Li et al., 2021) to customer service (Li & Lin, 2021).

Thanks to deep learning, we are able to build SER models that can successfully identify emotions by extracting progressively more complicated and abstract features from audio data.

This article is structured as follows: Section 2 presents an overview of the research on speech emotion recognition; Section 3 defines the datasets and the preprocessing used; Section 4 presents the architectures of the Neural Networks (NN) proposed; in Section 5 and Section 6 the training methods and the numerical results are presented, respectively; the interpretation is given in Section 7; limitations and further research are presented in section 8 and conclusions are provided in Section 9.

## 2. Related Work

Initially, SER was based on probabilistic models such as Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM). However, in recent times, deep learning has replaced this first approach and become dominant, relying on Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Deep Neural Networks (DNNs), and Long Short-Term Memory (LSTM) models (Wani et al., 2021).

Audio signals cannot be fed directly into a model. Careful pre-processing is needed to make sure that they are homogeneous, clear and informative (Lee et al., 2019). The rielaboration of audio data through feature extraction is crucial in determining the efficacy of the classification model (Abdulmohsin et al., 2021). Several features can be extracted such

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

as Mel Frequencies Cepstral Coefficients (MFCCs), Zero Crossing rate (ZCR), Root Mean Square (RMS), Chromagram or Mel spectrogram.

The complex and subjective nature of emotions poses significant challenges in identifying an optimal model (Berdos et al., 2022). Moreover, the field of SER models is new and constantly changing (de Lope & Graña, 2023). As a consequence, a wide literature exists that tries to tackle the problem of emotion classification and propose innovative solutions.

In (Ottoni et al., 2023), the authors merge four different datasets, all consisting of audio files of sentences recorded by actors simulating different emotions and, in some cases, also different level of intensities. They consider various combinations of optimizers, augmentation techniques and feature extraction, with both a CNN (with either 2, 4, 6 or 8 blocks) and a CNN combined with an LSTM model. The highest test accuracy (97.37%) is reached when using RAVDESS+TESS+SAVEE datasets, Adam as optimiser with learning rate of 0.001, stretch data augmentation, MFCCs as input and four CNN blocks plus LSTM.

In (Mountzouris et al., 2023) 6 DNNs are proposed for classifying audio from the RAVDESS and SAVEE datasets, using MFCCs as input: a DNN, a Simple Deep Neural Network (SDNN), LSTM, LSTM with an attention mechanism, a CNN and its version with an attention mechanism. The last one results in being the best model with a test accuracy of 74% on SAVEE and 77% on RAVDESS. The choice to include the attention mechanism enables the model to focus mainly on the parts of the audio data that contain more emotional information.

Since spectrograms carry multiple pieces of information of different nature, (Feng et al., 2017) proposed a hybrid architecture resulting from two parallel streams, a CNN and a bidirectional RNN, to classify music genres. The model ensures robust extraction of both spatial and temporal information and reaches a test accuracy of 90.2%.

The authors of (Nguyen et al., 2023) build a parallel deep learning SENet, CNN block and transformer with a Multi-head attention architecture using noise as data augmentation. They achieve a test accuracy of 82.67% on RAVDESS dataset.

(Gupta & Choubey, 2021) implements a shallow CNN and compares the accuracy according to different shapes of the input spectrograms. They reach the highest test accuracy (82.99%) on RAVDESS when using an input dimension of 224x224 and a grey scale (1 channel).

The huge number of parameters makes deep learning models prone to overfitting. This issue may also be exacerbated by clean audio sample. Hence, data augmentation techniques can play a crucial role in increasing the ability of generalisation (Nguyen et al., 2023). Common augmentation techniques with audio data are stretching, pitch, noise, shift, change volume, cropping and resampling (Abayomi-Alli et al., 2022). (Qasem et al., 2023) proposes a new technique, called spectrogram flipping, which consists in flipping horizontally an audio signal and then converting it into a different but equally informative spectrogram. These methods do not change the label of the original data point. However, new augmentation techniques (Wei et al., 2020) perform a linear or non-linear mixing of two sample points and also generate a soft label by performing the same mixing. This means that the new data point will not exactly belong to a certain category.

The contribution of this work relies in the exploration of data augmentation techniques and their efficacy using different datasets and deep learning models. Furthermore, this project investigates the effectiveness of the proposed model architectures with respect to different levels of intensity of the emotions and conditioning on the gender of the speaker.

## 3. Data preparation

### 3.1. Datasets

Two primary datasets are considered: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Surrey Audio-Visual Expressed Emotion (SAVEE).

The RAVDESS dataset contains 7,356 files encompassing speech, songs, and video recordings. Twenty-four professional actors, evenly split between genders, delivered two lexically-matched sentences in a neutral North American accent, either spoken or sung. From this dataset, we selected 1,440 speech files and 1,012 song files. The speech files cover eight emotional expressions: calm, happy, sad, angry, fear, surprise, and disgust; while the songs lack surprise and disgust emotions. Additionally, each sentence is articulated with two intensity levels, except for the neutral emotion, and for each intensity level it is repeated twice.

On the other hand, the SAVEE dataset consists of speech recordings by four male actors, comprising 480 repetitions across seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. These recordings feature sentences drawn from the phonetically-balanced TIMIT corpus, ensuring uniformity across emotional contexts.

We are going to evaluate the performance of the models presented in the next section in five distinct scenarios. First, we are going to assess their accuracy on the speech and song files from RAVDESS, both separately (Speech and Song) and jointly (RAVDESS). Next, we will consider the combined dataset comprising SAVEE and the speech recordings from RAVDESS (S+S). Finally, we are going to examine

the performance on the merger of all available data (All). The distribution of the emotions is showed in table 1.

*Table 1.* Distributions of emotions for each dataset

| | Speech | Song | RAVDESS | S+S | All |
|---|---|---|---|---|---|
| Sad | 192 | 184 | 376 | 252 | 436 |
| Happy | 192 | 184 | 376 | 252 | 436 |
| Calm | 192 | 184 | 376 | 0 | 0 |
| Anger | 192 | 184 | 376 | 252 | 436 |
| Fear | 192 | 184 | 376 | 252 | 436 |
| Disgust | 192 | 0 | 0 | 252 | 0 |
| Surprise | 192 | 0 | 0 | 252 | 0 |
| Neutral | 96 | 92 | 188 | 216 | 308 |

Moreover, for the RAVDESS dataset, the accuracy on the test set is evaluated conditioning on the gender of the speaker and on the intensity of the recorded emotion to investigate the presence of any gender bias in the classification performance.

### 3.2. Preprocessing

Each dataset undergoes a standardised preprocess: audio files are imported at a sampling rate of 16 kHz and silence is removed fixing a threshold of 25 decibels. To ensure uniformity across data points, a centered padding is applied, extending or truncating the duration to match the $90^{\text{th}}$ percentile of the distribution of the trimmed duration of the audio files. At this point, a random split allocates 70% of the data to the training set, 20% to the test set and the remaining 10% to the validation set.

### 3.3. Features extraction

Audio files cannot be directly fed into a model architecture; they need to be converted into another format through a process of features extraction (Swain et al., 2018). Among the most commonly used features in literature are Mel spectrograms and MFCCs.

Audio data can be described by the individual frequencies of the signal and their intensity. This is achieved computing a discrete Fourier transform (DFT) to obtain the spectrum. To provide an even more informative representation of the signal, multiple DFTs can be applied to stack together the resulting spectra, producing a spectrogram: a two-dimensional image, where the x-axis represents time, the y-axis represents frequency, and the color of each pixel denotes the amplitude, usually measured in decibels (dB). This can be converted into a Mel scale, which aligns more closely with human perception of sounds, resulting in a Mel spectrogram (Zaman et al., 2023).

Also, based on the mel scale, MFCCs are coefficients that collectively reproduce a mel-frequency cepstrum (MFC), a representation of the power spectrum of a signal through a non-linear transformation of the log power spectrum. They are a nonlinear spectrum of a spectrum and represent the shape of the vocal tract (Kumbhar & Bhandari, 2019).

In a preliminary analysis, we observed that MFCCs lead consistently to better results in terms of accuracy. For this reason, we decided to consider them, and not other features, as inputs to our models.

### 3.4. Data augmentation

A limitation of the models used in the field of SER is the great number of parameters that make them prone to overfitting (Shah Fahad et al., 2021). To address this concern initially, data augmentation techniques such as adding noise or altering pitch are applied to the training set, effectively tripling its original size. In details, we augmented the data in four different ways, applying one augmentation at a time, except for the last case: first we introduced noise, then we increased the pitch (High), afterwards we lowered the pitch (Low) and, finally, we duplicated the training set once with a higher pitch and once with a lower one (Mix). Evaluating the effectiveness of this approach and quantifying the resulting improvement involves comparing the accuracy levels achieved with and without data augmentation.

## 4. The architecture

The MFCC coefficients are presented as a grid of numbers, whose shape depends on the choice of parameters, such as the number of coefficients to return: a preliminary analysis showed that the a valid choice for such parameter is 30 MFCCs. The other parameters are the default ones of the librosa library. For instance, in the case of speech files in RAVDESS, the final shape of the MFCCs is $30 \times 72$.

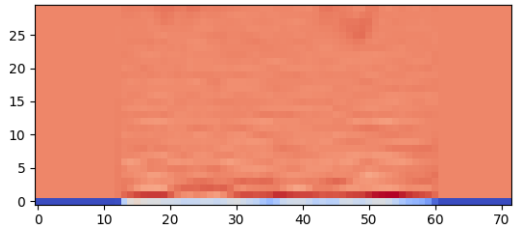


*Figure 1.* Example of the first MFCC in the training set of speech with shape $30 \times 72$.

Two main classes of models are considered: CNN and RNN. The former is usually composed of several convolutional blocks. Each block contains a convolutional layer, which utilises convolutional filters to detect patterns and features within grid-like input data, and a pooling layer to perform a downsampling of the data. The presence of multiple layers

and the introduction of non-linear activation functions allows the model to progressively extract increasingly abstract and complex features (Goodfellow et al., 2016).

While CNN models are able to detect spatial information, RNN models are also considered for their ability in detecting temporal relationships within the data. Indeed, it is reasonable to assume that the frequencies of an audio signal, which are then rielaborated through MFCCs, are time-dependent. RNN models are characterized by recurrent connections in the architecture that allow previous information to persist over time in the form of a hidden state, which acts as memory, representing the network's internal state at a given time step and encoding information about previous inputs.
A CNN and a parallel CNN-RNN (PCR) architectures are discussed here to address the SER problem.

### 4.1. CNN

The CNN model we have considered (figure 7) is composed by 3 convolutional blocks. These layers have 64, 64 and 128 filters, respectively; kernel size is $5 \times 5$ in the first one and $3 \times 3$ in the others; no padding is performed, an L2 penalty with a rate of 0.0001 is considered to mitigate overfitting, the HeNormal initializer is set and a ReLU (Rectified Linear Unit) activation function is used. A batch normalization layer follows the convolutional one and an increasing rate of dropout (20%, 30% and 40%) is introduced. Following, the output of the third convolutional block is flattened and directly fed into a dense layer with 128 neurons and with the same initializer, regularizer and activation function listed above. A final dropout with rate 0.6 preceeds the last classification layer with as many neurons as the number of emotions we aim at classifying and a softmax activation function to return a probability distribution across all classes.

### 4.2. PCR

To take into account also the time-dependencies in the input, a parallel CNN-RNN architecture is discussed (figure 8). The input is taken as the transpose of the MFCC, so in the case of speech files in RAVDESS, the shape of the input would become $72 \times 30$.

The CNN branch is composed of four convolutional blocks. The first block presents a convolutional layer with 128 filters, kernel size $3 \times 3$, stride of 1, ELU (Exponential Linear Unit) as activation function and padding to preserve the dimensions of the input. This is followed by a batch normalization layer, a max-pooling layer with both pool size and stride of $2 \times 2$ and, finally, a 20% dropout layer. The other 3 blocks are completely identical, except for the pool size and the stride of the max-pooling layer which are set to be $4 \times 4$. The output of the last convolutional block is flattened and, after a dropout layer, fed into a dense layer with 64 neurons. The last dropout layer preceeds the final dense layer with a softmax activation function.

The RNN branch has a quite simple structure: it is composed of a Recurrent Gate Units (GRU) layer with regularizer and initializer as in the CNN branch and a layer with 60% chances of neurons dropout. At this point, the outputs of both the CNN and the RNN branch are concatenated and fed into a dense layer with 128 neurons. A last dropout layer is applied, before the final classification layer.
Although an increased number of parameters, compared to the single-branch CNN, could potentially lead to overfitting, the introduction of an RNN branch tries to enhance the performance capabilities of our architecture, by taking into account not only spatial information, but also the likewise crucial temporal features hidden in the input.
Similar to LSTMs, GRUs preserve long-term dependencies in sequential data and handle the vanishing gradient problem that can occur with standard RNNs, where the information from earlier steps is repeatedly multiplied by small activation values during back-propagation, causing the gradients to become increasingly smaller. Both LSTMs and GRUs use gates to regulate the flow of information. GRUs have two gates, an update and a reset gate, whereas LSTMs have three gates, an input, a forget and an output gate. The simplified architecture of GRUs results in fewer parameters, compared to LSTMs, and therefore reduced complexity, meaning that they are less prone to overfitting and have reduced computational costs. (Zhang et al., 2021)

During our preliminary analysis, we found that the LSTM resulted in more overfitting, possibly due to the relatively smaller size of our datasets, and took longer to train compared to the GRU. Given the importance of minimising overfitting, and our limited computational resources, we opted for a GRU over an LSTM.

## 5. Training methods

The models we present in this work are trained using Adam as optimizer with an initial learning rate of 0.001. The learning rate is made flexible by allowing for reduction when the validation loss does not improve after a specified number of epochs. Moreover, each model is trained 3 times for each combination of dataset and data augmentation technique and the metric considered is the average accuracy on the test set over the three runs.

As already mentioned, overfitting may represent a great issue. To address this problem, an L2 penalty and several dropout layers are incorporated into the architectures. Moreover, data augmentation techniques are performed to increase the size of the training set, aiming at increasing the ability of the models to generalise to the unseen data point in the test set.

## 6. Numerical results

The performance of the CNN and PCR models on different datasets and augmentation techniques is summarized in Tables 2 and 3, respectively.

*Table 2.* Average accuracy test of the CNN (in percentage)

| | No aug | Noise | High | Low | Mix |
|---|---|---|---|---|---|
| SPEECH (8 labels) | 71% | 76% | 78% | 77% | 78% |
| SONG (6 labels) | 83% | 90% | 91% | 91% | 93% |
| RAVDESS (6 labels) | 76% | 85% | 84% | 83% | 86% |
| S+S (7 labels) | 55% | 68% | 68% | 69% | 69% |
| ALL (5 labels) | 69% | 80% | 81% | 81% | 82% |

The performance of the CNN model allows us to demonstrate improvements on how data augmentation leads to an increase in accuracy on all the available datasets. Notably, the song dataset consistently outperforms speech datasets: this fact could be driven by the presence of less emotions labels compared to the RAVDESS speech dataset and, perhaps, to the ability of songs to transmit emotions more effectively than speech. On the other hand, the combined dataset of RAVDESS and SAVEE exhibits lower accuracy, possibly due to the merging of SAVEE data, that are clearly different with respect to RAVDESS data in terms of audio quality. Lastly, we notice that the augmentation mix of low pitch and high pitch consistently yields the highest improvements in accuracy across all scenarios.

*Table 3.* Average accuracy test of the PCR (in percentage)

| | No aug | Noise | High | Low | Mix |
|---|---|---|---|---|---|
| SPEECH (8 labels) | 76% | 72% | 75% | 75% | 76% |
| SONG (6 labels) | 89% | 88% | 90% | 86% | 89% |
| RAVDESS (6 labels) | 80% | 77% | 81% | 79% | 81% |
| S+S (7 labels) | 73% | 71% | 71% | 71% | 73% |
| ALL (5 labels) | 82% | 82% | 83% | 82% | 83% |

Contrary to the CNN model, augmentation with PCR does not lead to significant improvements. The song dataset continues to exhibit higher accuracy compared to speech datasets, while the combined dataset of RAVDESS and SAVEE still lags behind, but not as much as with the CNN. Similar to this latter, the augmentation mix of low and high pitch consistently demonstrates the highest improvements in accuracy.

### 6.1. Overfitting

We now consider the performances of both models in the case of mixed augmentation, as it is the technique that improved the accuracies the most. The plots of the loss and the accuracy for both validation and training set of the CNN and the PCR are presented in figure 2 and 3, respectively. Both the loss and the accuracy of the CNN are stable, even though the gap between training and validation set, narrower for the PCR, may be an indicator of overfitting. Moreover, it seems that the PCR is learning the training set at a slower pace, determining an accuracy curve with a less steep initial slope.

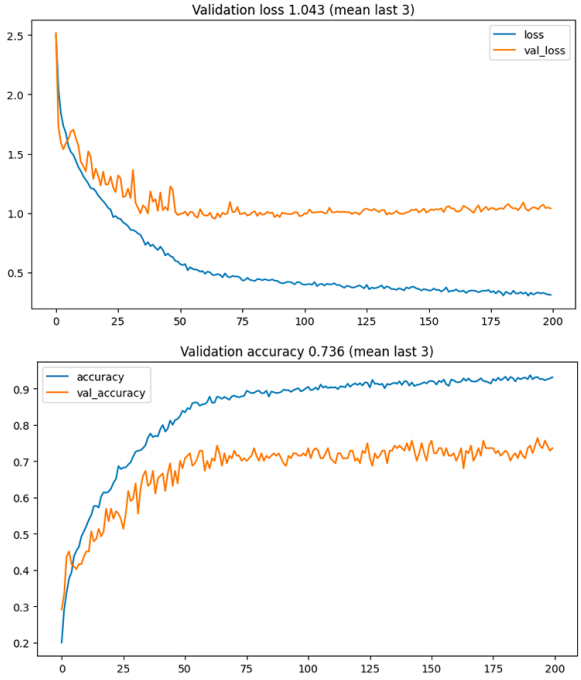


*Figure 2.* Accuracy and loss per each epoch of the CNN

Moreover, the confusion matrices for both cases are reported in figure 4. We saw in the previous section that the CNN and the PCR reached 78% and 76% of accuracy on the test, respectively. As a consequence, the confusion matrices present a marked diagonal, with the exception of neutral emotion, which is both more difficult to identify and also underepresents in the dataset.

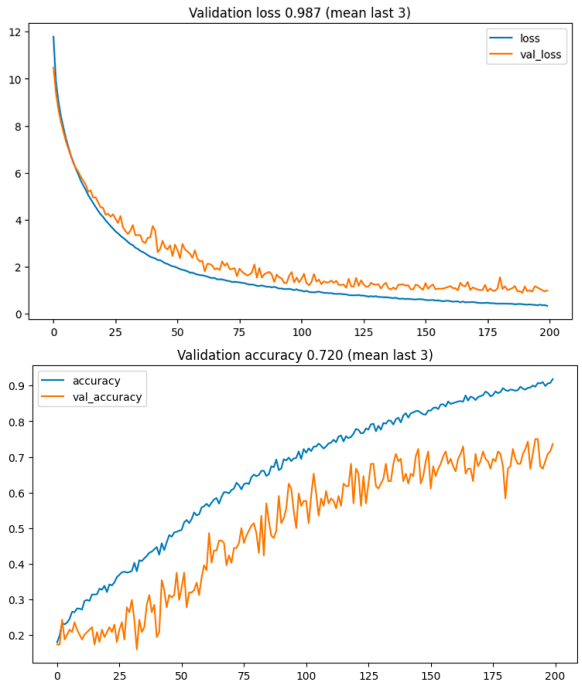


*Figure 3.* Accuracy and loss per each epoch of the PCR

### 6.2. Gender and intensity

In this section we focus our analysis on the RAVDESS Speech dataset, since it provides further information about the intensity and the gender of the speaker for each audio file. In fact, we splitted the test set by emotion intensity and by gender.

*Table 4.* Test accuracies by gender and intensity

| | CNN accuracy | Parallel accuracy |
|---|---|---|
| General | 79% | 77% |
| Female | 86% | 85% |
| Male | 72% | 69% |
| Low intensity | 74% | 73% |
| High intensity | 86% | 83% |

Table 4 shows how the CNN achieved a higher overall accuracy compared to the PCR. When conditioning the accuracy by the gender, both models recorded higher accuracies for female speakers. Moreover, it is interesting to see how both models showcased a higher ability in discerning high-intensity emotions.

These findings highlights the subtle interplay between gender and emotion intensity in SER models, with both architectures, demonstrating superior performances for female speakers and high-intensity emotions.

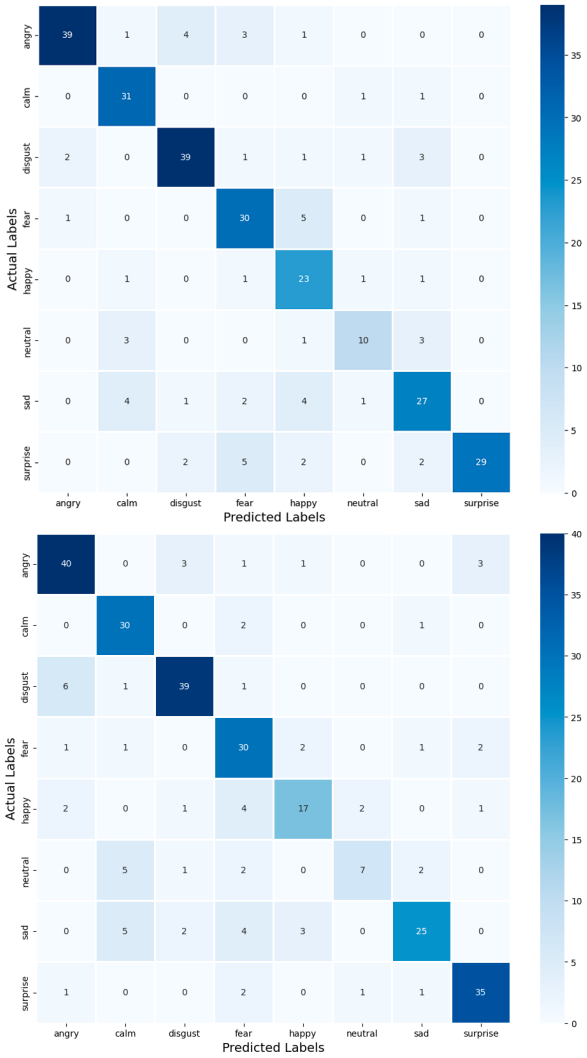


*Figure 4.* Confusion matrix of the CNN (top) and of the PCR (bottom)

## 7. Interpretation

### 7.1. Model explainability

To comprehend the decision-making process of the CNN for classification, we employed the Grad-CAM technique. Grad-CAM, which stands for Gradient-weighted Class Activation Mapping, is a technique used to understand which regions of an input image are important for the network's prediction of a particular class (Selvaraju et al., 2019). This technique works by decoding the importance of each feature map for a specific class by analyzing gradients in the last convolutional layer. Grad-CAM generates a heatmap that highlights the crucial regions of an image, thereby providing insights into the predictions, aiding debugging, and enhancing performance.

Our focus was on the final max-pooling layer preceding the flatten layer, which has a shape of $1 \times 7$. Given the architecture's design, this vector can be interpreted as a representation of the original audio file, segmented into seven parts. By using the Grad-CAM, we can discern which of these seven segments significantly influenced the classification decision.

Analyzing the Grad-CAMs of 15 observation of the training set for each emotion (see figure 9 in Appendix B), some patterns emerge. For instance, "Sad" often draws from the entire audio file, while "Neutral" typically focuses on either the beginning or the end, and, interestingly, in some cases there is not any part of the sentence relevant for the labeling 6. "Happiness" frequently has relevant parts towards the end of the audio. "Calm" is often associated with the latter part or the middle of the audio file. "Anger" shows a more varied pattern, which potentially makes it one of the labels most difficult to classify by models. "Fear" mirrors "Anger" but often has more relevance at the beginning. Lastly, "Disgust" and "Surprise" take information from the entire sequence, particularly the end.

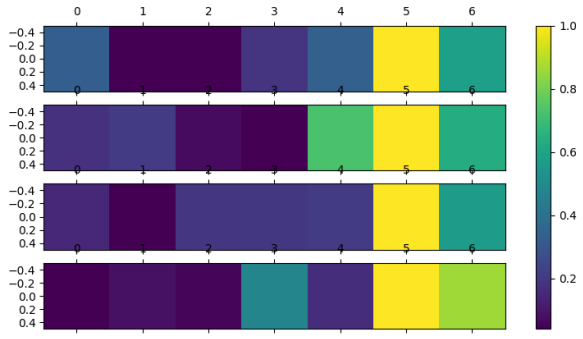

*Figure 5.* Gradcam of 4 observations of the training set with label "Surprise"

It is particularly intriguing to observe how the Grad-CAMs for "Surprise" (Figure 5) predominantly concentrate on the final segment of the audio file: in our interpretation, this could be related to the fact that typically, the last part of a sentence encapsulates the emphasis of a question.

Our analysis aims to explore the relevance and the location of specific vocal features in identifying emotions, such as vocal inflections, variations in pitch or pauses. In our opinion, these variations, even if subtle, may be the ones that truly allow us to understand the emotional status of people we interact with on a daily basis.

The model performance can also be evaluated with greater details according to its ability of classifying each emotion. In general terms, precision is a metric that reflects the accuracy of the model's positive predictions; recall quantify its ability to correctly identify all positive instances, and

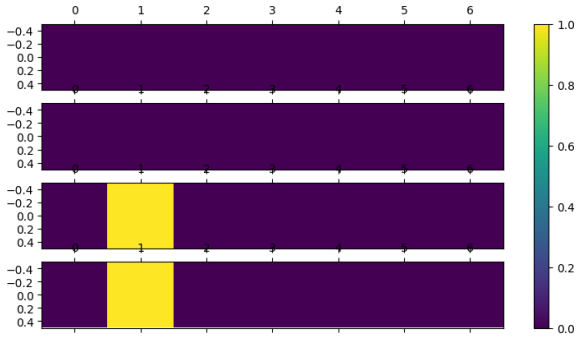

*Figure 6.* Gradcam of 4 observations of the training set with label "Neutral"

the F1-score is the harmonic mean of precision and recall, offering a summary metric of the two. For instance, in Table 5, which refers to the CNN, we observe that the model achieves high precision and recall for emotions like anger and surprise, indicating its effectiveness in correctly identifying these states. However, for emotions like happiness and calmness, while precision is relatively lower, recall remains moderate, suggesting that the model classifies audio files as "Happy" or "Calm" more times than it should. Similarly, Table 6 represents the classification report for the PCR model. Comparing the two tables, we can discern how each model performs across different emotional categories. For instance, the PCR model tends to have higher precision and recall for disgust and surprise compared to the CNN model, but it struggles more with emotions like happiness and sadness.

*Table 5.* Classification report of the CNN

| Emotion | Precision | Recall | F1-Score |
|---|---|---|---|
| Angry | 83% | 90% | 86% |
| Calm | 75% | 91% | 82% |
| Disgust | 90% | 74% | 81% |
| Fear | 79% | 81% | 80% |
| Happy | 69% | 81% | 75% |
| Neutral | 76% | 76% | 76% |
| Sad | 81% | 67% | 73% |
| Surprise | 89% | 85% | 87% |

### 7.2. Investigation of gender bias and intensity

Emotion intensity or the gender of the speaker may play an important role in terms of how accurately an emotion is identified. Both the CNN and the PCR exhibit an interesting behaviour, achieving higher accuracies when conditioning on female speakers and on high-intensity emotions. The enhanced ability at recognizing stronger emotions may be due to the fact that these types of audio files show clearer acustic features, whereas less intense emotions are characterised by

*Table 6.* Classification report of the PCR

| Emotion | Precision | Recall | F1-Score |
|---|---|---|---|
| Angry | 81% | 79% | 80% |
| Calm | 76% | 79% | 78% |
| Disgust | 93% | 79% | 85% |
| Fear | 72% | 76% | 74% |
| Happy | 51% | 78% | 62% |
| Neutral | 71% | 59% | 65% |
| Sad | 81% | 64% | 71% |
| Surprise | 81% | 85% | 83% |

more subtle attributes. On the other hand, gender bias may be explained by physical differences that lead to different pitches and tones in voices, which, in turn, affect the sound properties of the speech. It is important to keep these biases in mind when designing and training SER models to ensure their efficacy accross genders.

## 8. Limitations and further research

While our study provides valuable insights into SER models, it also has its limitations and areas for further research. One limitation is the reliance on synthetic datasets, namely RAVDESS and SAVEE, which may not fully capture the complexity and variability of real-world emotional expressions. Additionally, our focus on a specific set of augmentation techniques and model architectures may overlook potentially effective approaches not explored in this study. Despite efforts to mitigate overfitting through regularization techniques such as dropout and L2 penalty, the deep learning models utilized in this study may still be susceptible to overfitting, particularly when trained on relatively small datasets. Further exploration into advanced regularization methods or strategies to increase dataset diversity could help address this limitation and improve the generalization capabilities of the models.

Future research could explore the impact of additional factors, such as cultural differences in emotional expression, linguistic variations, and context-specific cues, on SER model performance. Moreover, investigating the transferability of SER models across different languages and cultures would be valuable for creating more universally applicable models. Additionally, incorporating multimodal data, such as facial expressions and physiological signals, could enhance the accuracy and robustness of SER systems. Furthermore, exploring novel techniques for data augmentation, model interpretability, and handling class imbalances could further improve the effectiveness of SER models in real-world applications. Overall, addressing these limitations and expanding the scope of research could lead to more robust and inclusive SER systems.

## 9. Conclusion

In this paper, we investigated the realm of Speech Emotion Recognition by exploring the influence of data augmentation, model architectures, and dataset characteristics on classification accuracy. We reviewed existing research in the field, highlighting the transition from probabilistic models to deep learning architectures, and the challenges posed by the subjective nature of emotions.

Our study introduced CNN and parallel CNN-RNN models for SER and examined their performance across various datasets and augmentation techniques. We found that data augmentation techniques, particularly pitch manipulation, significantly improved classification accuracy, particularly within the CNN framework. Additionally, interpretation techniques such as Grad-CAM provided insights into the models' decision-making processes, revealing patterns in audio features associated with different emotions.

Furthermore, we scrutinized the impact of gender bias and emotion intensity on model performance, revealing that models exhibited higher classification accuracy under conditions of heightened emotion intensity and when the speaker was a female.

Overall, our findings contribute to a better understanding of SER models and provide valuable insights for improving their performance in real-world applications. By addressing limitations and exploring further research avenues, we can continue to advance the field of SER and develop more robust and inclusive models for understanding and interpreting human emotions in speech.

## References


Abayomi-Alli, O. O., Damaševičius, R., Qazi, A., Adedoyin-Olowe, M., and Misra, S. Data augmentation and deep learning methods in sound classification: A systematic review. *Electronics*, 11(22), 2022. ISSN 2079-9292. doi: 10.3390/electronics11223795. URL https://www.mdpi.com/2079-9292/11/22/3795.

Abdulmohsin, H. A., Abdul wahab, H. B., and Abdul hossen, A. M. J. A new proposed statistical feature extraction method in speech emotion recognition. *Computers Electrical Engineering*, 93:107172, 2021. ISSN 0045-7906. doi: https://doi.org/10.1016/j.compeleceng.2021.107172. URL https://www.sciencedirect.com/science/article/pii/S0045790621001749.

Berdos, P. J. B., Saligumba, J. O., Deveza, K. P., and Estrada, J. E. Discovering the optimal setup for speech emotion recognition model incorporating different cnn architectures. In *2022 IEEE 14th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management*

*(HNICEM)*, pp. 1–5, 2022. doi: 10.1109/HNICEM57413.2022.10109279.

de Lope, J. and Graña, M. An ongoing review of speech emotion recognition. *Neurocomputing*, 528:1–11, 2023. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.01.002. URL https://www.sciencedirect.com/science/article/pii/S0925231223000103.

Feng, L., Liu, S., and Yao, J. Music genre classification with paralleling recurrent convolutional neural network, 2017.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Gupta, R. and Choubey, A. Shallow over deep neural networks: A empirical analysis for human emotion classification using audio data. 05 2021. ISBN 978-3-030-76735-8. doi: 10.1007/978-3-030-76736-5_13.

Kumbhar, H. S. and Bhandari, S. U. Speech emotion recognition using mfcc features and lstm network. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pp. 1–3, 2019. doi: 10.1109/ICCUBEA47591.2019.9129067.

Lee, K. H., Kyun Choi, H., Jang, B. T., and Kim, D. H. A study on speech emotion recognition using a deep neural network. In *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 1162–1165, 2019. doi: 10.1109/ICTC46691.2019.8939830.

Li, H.-C., Pan, T., Lee, M.-H., and Chiu, H.-W. Make patient consultation warmer: A clinical application for speech emotion recognition. *Applied Sciences*, 11(11), 2021. ISSN 2076-3417. doi: 10.3390/app11114782. URL https://www.mdpi.com/2076-3417/11/11/4782.

Li, X. and Lin, R. Speech emotion recognition for power customer service. In *2021 7th International Conference on Computer and Communications (ICCC)*, pp. 514–518, 2021. doi: 10.1109/ICCC54389.2021.9674619.

Mountzouris, K., Perikos, I., and Hatzilygeroudis, I. Speech emotion recognition using convolutional neural networks with attention mechanism. *Electronics*, 12(20), 2023. ISSN 2079-9292. doi: 10.3390/electronics12204376. URL https://www.mdpi.com/2079-9292/12/20/4376.

Nguyen, A. H., Trang, K., Thao, N. G. M., Vuong, B. Q., and Ton-That, L. Speech emotion classification with parallel architecture of deep learning and multi-head attention transformer. In *2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE)*, pp. 1549–1554, 2023. doi: 10.23919/SICE59929.2023.10354088.

Ottoni, L. T. C., Ottoni, A. L. C., and Cerqueira, J. d. J. F. A deep learning approach for speech emotion recognition optimization using meta-learning. *Electronics*, 12(23), 2023. ISSN 2079-9292. doi: 10.3390/electronics12234859. URL https://www.mdpi.com/2079-9292/12/23/4859.

Qasem, M. M. Z., Salwani, M. D., and Samy, S. A.-N. Spectrogram flipping: a new technique for audio augmentation. *Journal of Theoretical and Applied Information Technology*, 101(11):4433–4447, 2023. URL https://www.jatit.org/volumes/Vol101No11/26Vol101No11.pdf.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128 (2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Shah Fahad, M., Ranjan, A., Yadav, J., and Deepak, A. A survey of speech emotion recognition in natural environment. *Digital Signal Processing*, 110:102951, 2021. ISSN 1051-2004. doi: https://doi.org/10.1016/j.dsp.2020.102951. URL https://www.sciencedirect.com/science/article/pii/S1051200420302967.

Swain, M., Routray, A., and Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1):93–120, 2018. doi: 10.1007/s10772-018-9491-z. URL https://doi.org/10.1007/s10772-018-9491-z.

Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., and Ambikairajah, E. A comprehensive review of speech emotion recognition systems. *IEEE Access*, 9:47795–47814, 2021. doi: 10.1109/ACCESS.2021.3068045.

Wei, S., Zou, S., Liao, F., and weimin lang. A comparison on data augmentation methods based on deep learning for audio classification. *Journal of Physics: Conference Series*, 1453(1):012085, jan 2020. doi: 10.1088/1742-6596/1453/1/012085. URL https://dx.doi.org/10.1088/1742-6596/1453/1/012085.

Zaman, K., Sah, M., Direkoglu, C., and Unoki, M. A survey of audio classification using deep learning. *IEEE Access*,

11:106620–106649, 2023. doi: 10.1109/ACCESS.2023.3318015.

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. *Dive into Deep Learning*. CRC Press, 2021.

## A. Architectures

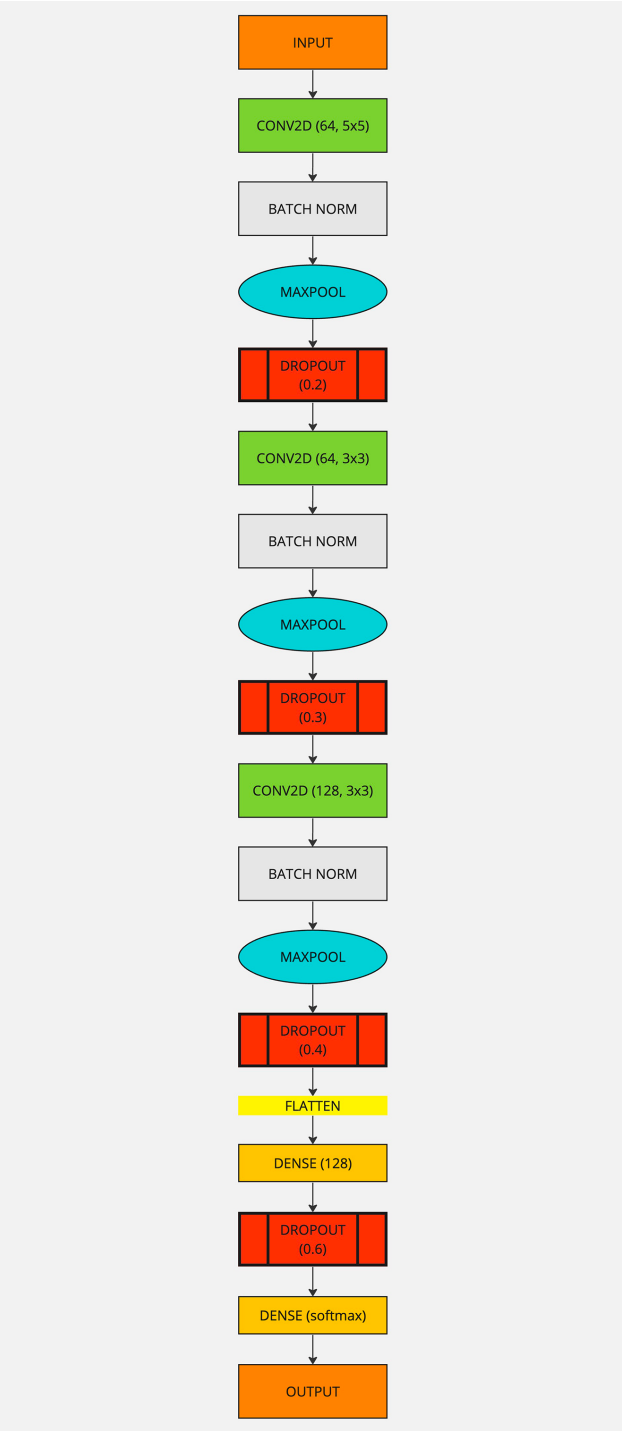


*Figure 7.* Architecture of the CNN.

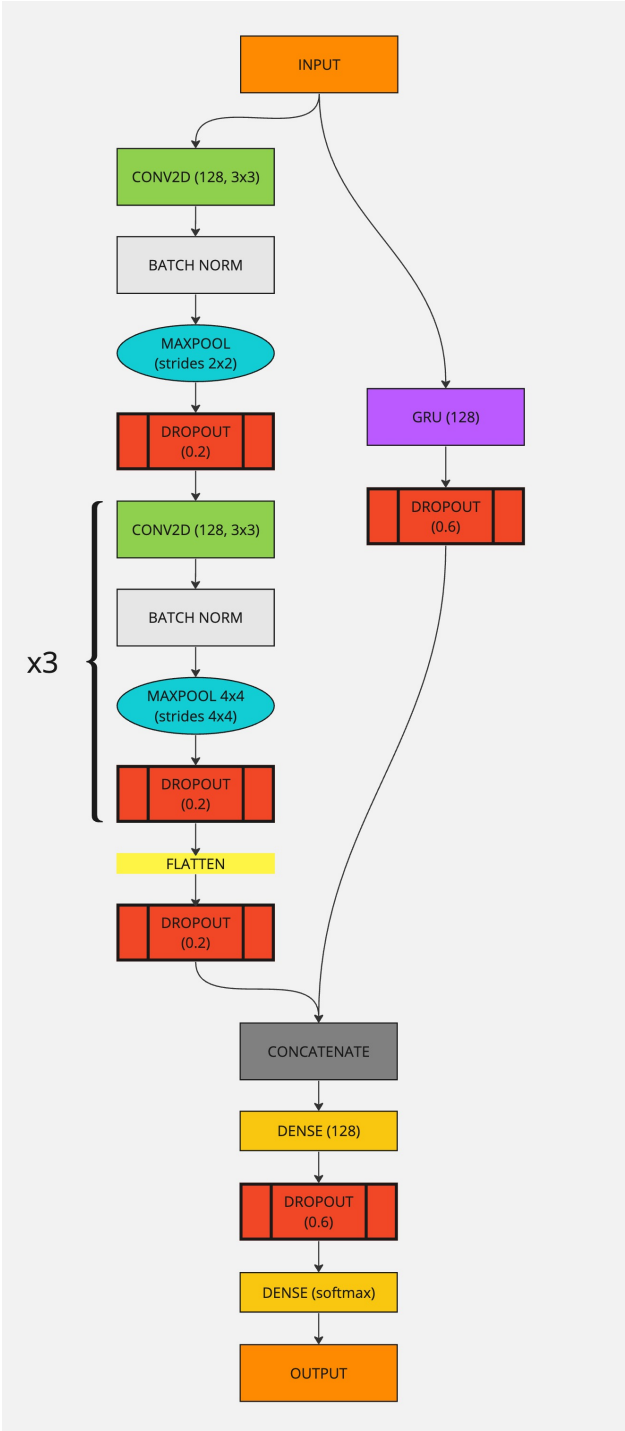


*Figure 8.* Architecture of the PCR.

## B. Grad-CAMs of 15 observation for each emotion

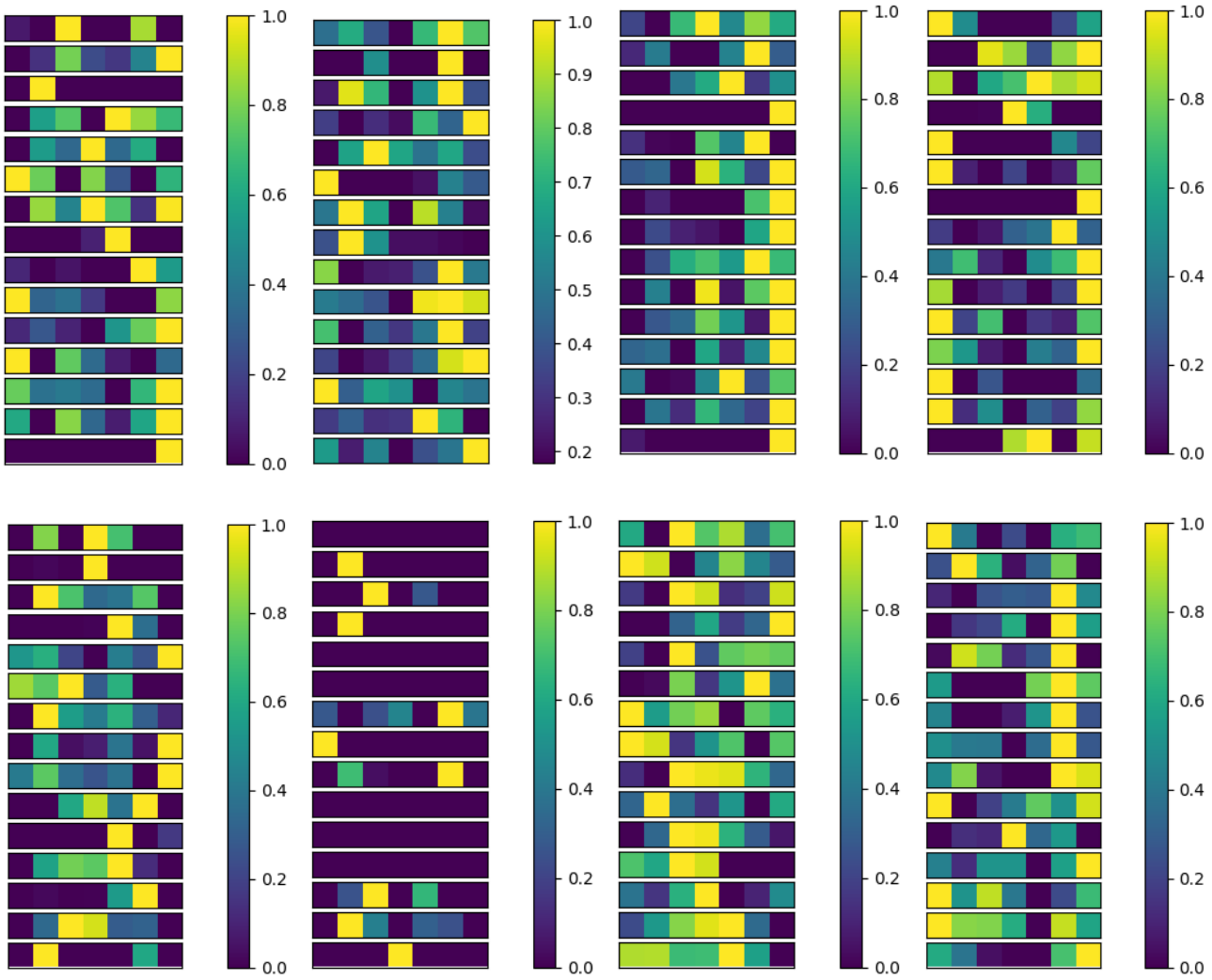

*Figure 9.* Grad-CAMs of 15 observation for each emotion: in the order from left to right, angry, disgust, calm, fear, happy, neutral, sad and surprise

## C. Statement about individual contributions

Every member of the team made equal contributions to both the code writing and report production.