Coventry University

Faculty of Engineering, Environment and Computing

BSc Computer Science

"Comparative Analysis on Single Classification Models Using SVM and CNN Algorithms for Speech Emotion Recognition Systems in Vehicles"

Sofia Alexandra Valente
Student ID: 8897758

6001CEM Declaration of Originality

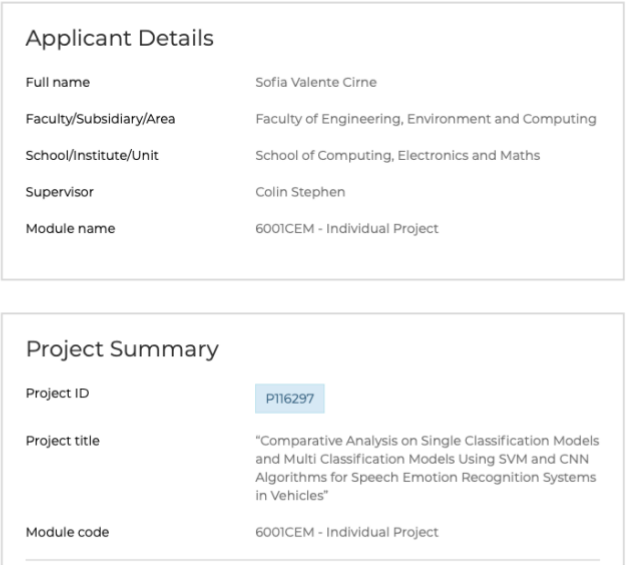
I declare that this project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

Statement of copyright

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialize products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information, please see www.coventry.ac.uk/ipr or contact ipr@coventry.ac.uk.

Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below.



Applicant Details

Full name	Sofia Valente Cirne
Faculty/Subsidiary/Area	Faculty of Engineering, Environment and Computing
School/Institute/Unit	School of Computing, Electronics and Maths
Supervisor	Colin Stephen
Module name	6001CEM - Individual Project

Project Summary

Project ID	P116297
Project title	"Comparative Analysis on Single Classification Models and Multi Classification Models Using SVM and CNN Algorithms for Speech Emotion Recognition Systems in Vehicles"
Module code	6001CEM - Individual Project

Figure 1 - Ethics Application: P116297

Signature:

Date: 22/11/21

Table of Contents

Table of Figures

Table of Tables

1. Abstract

Speech Emotion Recognition (SER) is essential for human-computer interaction as autonomous vehicles enter the industry. Human-vehicle collaboration is imperative not only for safety, but for user experience as well. SER data can assist the assessment of the driver's capabilities and behaviour in almost real-time. Therefore, comparative analysis of data, features and models assists the understanding of the process to select the most precise one. Thus, in this study the implications of machine learning techniques such data augmentation, conventional classification and artificial neural networks models are reviewed and assessed to later be compared and evaluated in order to reach the best fit when predicting emotions through speech signal in a driving environment.

Datasets such as IEMOCAP, TESS, CREMAD, SAVEE, Emo-DB were examined for speech feature extraction. The features extracted were Mel Frequency Cepstral Coefficients (MFCC), Chroma energy and Logarithmic Mel Spectrogram (LMS) to accommodate the data inputs requirements of the 1D and 2D CNN models. Several audio augmentation tools were used, and their efficacy tested. The raw data reached higher accuracies when implemented in a CNN 1D context. However, the noise injection outperformed it when applied to SVM. Later, audio classification models, SVM, LSTM and CNN, were considered individually and combined, which achieve accuracies between approximately 37% and 67%, with the CNN 2D+LSTM surpassing the rest.

2. Acknowledgments

During the development of this project and dissertation, I received immeasurable support from the people around me.

Firstly, I would like to thank my Family and Friends for the encouragement, kindness, and patience in this challenging phase of my life. A special thank you to Mother Anabela Valente for the guidance in the early stages of my dissertation and Wallef Borges and Donata Izekor for embarking in this adventure with me and never giving up.

I would also like to thank all my Lectures for providing the tools, skills, and an environment where I could learn and progress to successfully complete this project. Additionally, to Peter Every and Collin Stephen for the guidance and advice when completing this project.

Lastly, to Coventry University for the opportunity to develop myself, not only as a professional, but also as a person.

3. Introduction

As technology advances in the world, more advanced techniques are required for human-machine interaction. As a sub-set of artificial intelligence (AI), Machine Learning algorithms mimic the human learning processes through the identifications of patrons in large amounts of data. This predicts the outcome of a specific situation based on that data and is applied is several industry sectors like finance, entertainment, advertising, health care and transportation.

As society moves to a more connected world, transportation and commuting occupy a significant part of life. However, driving is often associated with negative emotional states such as anger, frustration, anxiety, or fear, (Underwood et all., 1999). Due to emotions being so critical for daily functioning such as decision making, critical thinking and communication, a negative emotional state can impact the driver and trigger a lack of control or focus, travel delays and potential accidents. Thus, several methods can be applied for recognising emotion according to the type of data collected from the vehicle: facial expression, speech, and physiological signals.

Speech Emotion Recognition (SER) uses the collected data almost in real time from speech signals to predict the emotion expressed. This method recognises patterns in speech by assessing pitch, speed, intonation, and other features without a need for context. Additionally, more data can be collected for this method as the sensor is of easy implementation and can be tested in most levels of vehicle autonomy, which vary from zero to five, being five completely autonomous. Thus, Machine Learning and Speech Emotion Recognition together provide a prediction mechanism to make driving more predictable and less dangerous.

In this dissertation the lack of industry benchmarks when predicting the efficacy of a model is addressed, (Zepf et all., 2020), by analysing, reviewing, and comparing the datasets used to train the models, data augmentation strategies and the SER models with the most efficient accuracy. This comparison is imperative to eliminate any biases and errors when training and testing these systems. While there has been previous research on SER algorithms, none has focused in comparing augmented data with deep learning such as CNN and LSTM and supervised models, SVM.

To prove the methodology proposed, the design and comparison of different tools and models was done. The data augmentation tools were applied in the early stages of the process and the models trained afterwards using four datasets of actors demonstrating different emotion in an audio format and tested on different samples. In the literature review a more detailed overview is considered, addressing the consequences of SER in driving scenarios, while referring to several researches revelling its benefits and implementation. The methodology presents the tools and models that were trained and tested with. The research incorporates six main trained models. However, four data augmentation tools were applied for two of the main 6, SVM and CNN 1D, making a total of 14 models compared. In the results chapter, the results of the comparison are observed and in the discussion section the results are their implication are discussed, along with the limitations of the study.

4. Literature Review

The influence of emotional states has been considerably researched, especially in a driving context, as negative emotions impact cognitive processes, communication, and performance. This can have implications not only to the driver, but also to others. Thus, assessing the emotional state is imperative to detect stress, anger, and anxiety/fear in drivers to identify potential threats while driving and prevent long-term health issues caused by high levels of negative emotions during long periods of time, (Underwood et all., 1999). As adaptive assistance and automotive systems advance, more systems are design to predict dangerous situations based on the emotional state.

There is a considerable amount of research on emotion recognition in vehicles, especially since machine learning started to integrate most systems. However, as there is an absence of comparative research with industry benchmarks, particularly when using speech signals, (Zepf et all., 2020). this study aims to review the data, tools, and models for an unbiased and more uniform research. This research compares classifications systems of supervised (SVM) and deep learning (CNN and LSTM) models. Therefore, the models are compared individually and combined to accurately analyse their influence in systems.

4.1. Emotion

To classify and identify emotions, it is required to define them first. This definition is both important and difficult, especially when building a system that aims to identify them based on data. According to Cabanac et all (2002), and Schacter et all (2011), found that emotion is any conscious experience characterized by any mental activity with high intensity and a high level of pleasurable or unpleasurable sensations which are then translated into motor, visceral and cognitive components. However, further research has determined that emotions are also related to temperament, mood, personality, motivation, and disposition. Though, a scientific consensus is yet to be reached about the subject. Emotions are associated with a complex state that can be expressed mentally, physiologically, and physically, (Cabanac et all.,2002), making it possible to distinguish patterns and recognised them.

4.2. Categorization of emotions

Emotions are categorized based mainly on two approaches: categorical and dimensional. The first emotions are defined due to a discrete number of classes. Ekman (1992) clarifies that each emotion acts as a discrete category. He also differentiates between primary and secondary emotions. He identified six primary emotions: anger, disgust, fear, happiness, sadness, and surprise.

The second method describes emotions as a combination of psychological dimensions. These are represented in axes. Even though, Wundt in 1897 proposed three dimensions: strain/relaxation, pleasurable/unpleasurable and arousing/subduing, (Sarprasatham, 2015), other researchers proposed less with variations.

However, in the context of speech emotion recognition, the categorical method is preferred, (Koolagudi et all., 2012), distinguishing between the primary emotions from Ekman’s model plus neutral as these have distinctive physiological characteristics that easily measured.

4.3 Sensory modalities

Human emotions are expressed physiologically in a diverse number of ways. Nonverbal communication can be expressed by facial expression, biological signals, and the characteristics of speech without relying on content, (Zepf et all., 2020). Facial expression is widely used as camera equipment is of easy access, thus, making data collection easier. However, some algorithms rely on face 3D images. This increases the accuracy of the models, but resources are less available. Physiological signals are less used as they require to be in contact with the driver and are more costly to implement. In speech, data is collected either through microphones or video feed, which increases the learning rate of the model as more examples are available. Therefore, in this study speech signals are used.

4.4. Speech

Speech is a complex wave signal that contains information in a sound wave format. It holds several types of information such as the content of the message, the speaker, language, and emotion. As humans we comprehend the message by perceiving the underlying information using multi-modal clues such as emotions and cultural background, along with the phonetic information and the other factors, (Koolagudi et all., 2012).

Speech Emotion Recognition (SER) has emerged as one of the important speech research areas, especially due to real-time application between humans and machines as nonverbal communication carries important information from machine's perspective.

4.5. Datasets

Before the models are implemented and feature extraction applied, a database must be selected. An accurate representation of emotions through audio signals is imperative to train the proposed systems. The most used databases for speech signals in literature are IEMOCAP, Emo-DB, SAVEE, TESS and CREMAD. All of them rely on simulated emotional data carried out by actors in a non-driving environment, (Zepf et all., 2020), (Jackson et all., 2011), (Pichora-Fuller et all., 2010) and (Cao et all., 2014). The Data was collected in improvised and scripted sessions with a range of actors, female, and male, containing recordings 6 emotional states: anger, anxiety/fear, boredom, disgust, happiness, neutral and sadness, (Abbaschian, 2021).

Zepf et all, (2020) also identified other datasets, such as CIAIR Corpus, UTDrive DB Portable and UTDrive Classic. This data was collected in-vehicle in a more natural way, without the help of actors. However, these data sets are reduced and in controlled situations. Due to this and the lack of access, they were not used.

4.6. Features

The features extracted from the audio data have a massive impact on the model as these are the characteristics the algorithm will base its prediction on. These carefully shaped can increase the recognition rate and performance. Akçay and Oğuz (2020) categorise TEO features as ideal for stress and anger recognition. Yet, Zeng et all (2020) implemented several feature extractions approaches, reaching an understanding of how it affects the results. Thus, finding Mel Frequency Cepstral Coefficients (MFCC) more efficient, especially when combined with SVM. However, in addition to MFCCs, Zhao et all., (2019), relied on Log Mel Spectrogram (LMS) to extract the features, while Reakaa et all (2021) also incorporated in their analysis the Chroma feature in combination with LMS. Even though other methods of feature extraction would be suitable, due to the 2D models, CNN 2D and CNN 2D – LSTM, these three features were implemented across all the models.

4.7. Models

Özseven (2019) started by centring his research on supervised algorithms such as SVM, KNN and MLP, concluding that SVM approaches reach higher accuracy results. However, with the introduction of deep learning algorithms in emotion recognition systems, CNN and LSTM approaches started to emerge in literature for emotion recognition. The CNN technique is extensively applied since its design accommodates the process of multidimensional data structures in multiple arrays or tensors, and LSTM due to its ability to process to not only process single data points, but also long sequences of data. More recently, this algorithm has only been used in hybrid models, such as CNN+LSTM, as a layer for higher-level feature extraction, (Chen et all. 2018). CNN algorithms can also be characterized according to the dimensions of the data, 1D, 2D and 3D, which can also be implemented alongside hybrid system, (Zhao, 2019). As so, Christy et all (2020) made a broader comparison not by just including linear regression, decision trees and random forests, but also CNN algorithms to classify and predict from speech signals.

Though CNN algorithms are the most used techniques, some authors argue RNN has a greater impact on the system then CNN, Yao et al. (2020). RNN algorithms are mostly used for sequential data as they have the ability to remember and process sequences of input data, where each sample is connected to the previous one, due to its internal memory. Hence making a good candidate for speech recognition algorithms. Even though supervised and deep learning models are the most enforced techniques, some authors, such as Latif et all (2020), started to analyse unsupervised deep learning techniques in the context of speech recognition.

Techniques like GAN mimic and learn data distributions for augmenting feature learning and generation in speech recognition. However, this method has shown lower performance rates. Latif et all (2020) proposes a new approach that addresses the issue by applying a mixup more effectively. Unsupervised deep learning techniques might have some potential, however, more research on the model’s mechanisms and its possibilities in speech recognition is necessary.

Even though many algorithms have been built, tested, and compared, different architectures in CNN algorithms have not, especially when combined with other dimensions and algorithms, LSTM. Influenced by Szeged et all (2016), and their research, as well as Zepf et all (2020) and Ismail Fawaz et all (2020), this research not only compares individual and multiple classification models, but also the data collection method. Additionally, the project creates a benchmark for some of the most popular used algorithms, SVM and CNN, and databases for the automotive industry.

5. Methodology

To recognise emotion through speech, Machine Learning was applied to a set of audio files of several actors demonstrating six different emotions: anger, disgust, anxiety/fear, happiness, sadness, and neutral. After the audio is process into digital, the data is pre-processed and data augmentation features are applied, such as noise, pitch, shift and speed&pitch. After, the data is trained with both SVM and CNN 1D to observe the impact these features have on the accuracy of the models. Feature extraction is the applied to the data, obtaining the speech features, MFCC, LMP and Chroma. The features are then used to train several models individually, SVM, CNN 1D, CNN 2D and LSTM, and then combined, CNN 1D – LSTM and CNN 2D – LSTM.

Later, the models are compared for accuracy and performance, as they are commonly used for similar situations in emotion recognition. The process is displayed on a Jupyter notebook, using python. The methods used rely mainly on TensorFlow and SKLearn.

5.1. Dataset

This project integrates four different datasets: Emo-DB, SAVEE, TESS and CREMAD. These datasets were chosen primarily due to the quality of the records. Initially, only SAVEE and TESS were implemented, however, these lacked diversity as 4 male actors were using in SAVEE and only 2 female actors in TESS. Emo-DB and CREMAD were added later since they have extensive utterances and diverse actors, with CREMAD incorporating audios from different cultural backgrounds, genders, and accents. Thus, making these additions essential for a better accuracy. Table 1 represents a complete description of each database and where it was downloaded.

Name	Path	Organization	Utterances	Description
Emo-DB	http://emodb.bilderbar.info/start.html	Technical University of Berlin	535	10 actors: 5 female, 5 male
SAVEE	http://kahlan.eps.surrey.ac.uk/savee/	University of Surrey	480	4 male actors
TESS	https://tspace.library.utoronto.ca/handle/1807/24487	University of Toronto	200	2 female actors
CREMAD	https://github.com/CheyneyComputerScience/CREMA-D, Cao et all., (2014)	University of Pennsylvania	7442	91 actors: 48 male, 43 female

Table 1 - Dataset's description

5.2. Data Analysis

Before the data is pre-processed, an analysis is completed to verify and understand the information from each individual dataset. The knowledge of their organization, labels and ratio of each emotion label are fundamental for processing the data in a way that can be studied by the models. To assistance with this, distribution graphs were designed for each dataset, figure 2 to 5.

Most of the datasets expresses mainly uniform emotion label rations. Both SAVEE and CREMAD stating the same number of labels of each emotion, 60 and 1271, with the exception if the neutral status. This emotion counts double the samples, 120, for SAVEE and 1087 for CREMAD, figure 3 and 5. However, TESS remain uniform throughout the whole set of emotion, with 400 samples each, figure 4. Emo-DB is the only one that doesn't display uniformity, incorporating 127 utterances of the emotion anger, 81 of boredom, 79 of neutral, 71 of happiness, 69 of anxiety/fear, 62 of sadness and 46 of disgust, figure 2.

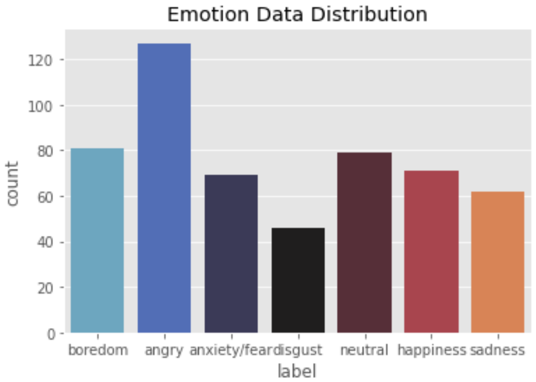


Figure 2 - Emo-DB Distribution Graph

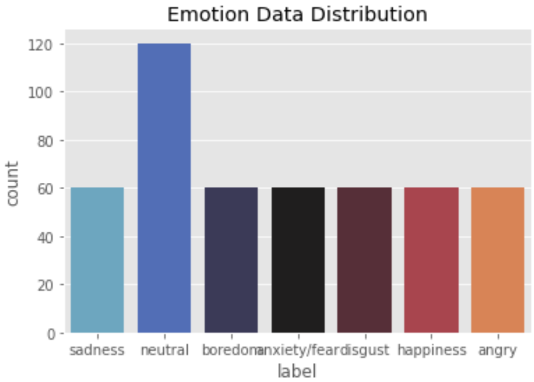


Figure 3 - SAVEE Distribution Graph

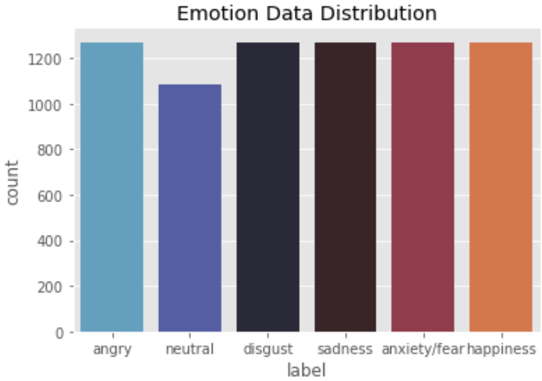


Figure 4 - CREMAD Distribution Graph

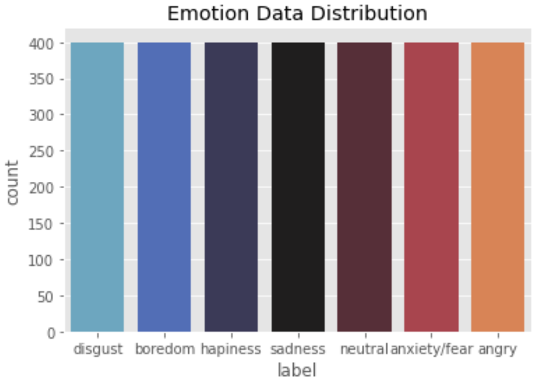


Figure 5 - TESS Distribution Graph

5.3. Data pre-processing

5.3.1 Data augmentation

Data augmentation allows to increase the diversity of the dataset by creating synthetic data samples thru adding small perturbations in the raw data, while preventing overfitting. Overfitting is a common problem in Machine Learning that emerges when small datasets are paired with large neural networks, allowing the model to learn from statistical noise. This prevents the model to recognise new data, Zhao et al., (2019). In this project overfitting was observed in the early stages. However, due to data augmentation and the combination of larger datasets, such as the TESS and CREMAD datasets.

Augmented data trains the models to be invariant when these patterns are present, while enhancing its ability to generalize. Several techniques can be implemented to achieve this, however only noise, shifting, pitching and speed&pitch were tested, as these are the most used. Even though all these techniques introduce diversity to the sample, only noise injection and shifting add data. Noise injection adds a random value to the processes data from audio, while shifting moves a second either forward or backwards, (Tiwari, 2020). Pitch and speed transform the data instead of adding new information.

5.3.2. Feature Extraction

Feature extraction is a crucial phase in the algorithm as it finds and highlights the most dominating and discriminating characteristics of a signal, while not depending on the speaker or the content. Suitable features mimic the properties and are compacted in groups for processing. These are categorised into continuous or time domain, qualitative, spectral, and TEO (Teagerenergy operator)-based features and each is subcategorised into several techniques, (Ayadi, 2011). However, in this project only continuous and spectral features were used.

Continuous features have been widely used in SER since it includes pitch and energy formats. Even though continuous features include a variety of techniques, only Chroma energy was used in this project due to its output format, 2D. Chroma energy focuses on the twelve pitch attributes of sound as used in western music notation, figure 7. Each Chroma vector describes the energy distribution across these twelve bands regarding the pitch of each, making it an essential tool for content-based retrieval tasks, such as audio matching and version identification, (Müller, 2021).

Moreover, MFCCs and LMS, represented in figure 8 and 6, which belong to the spectral group, are the most relevant and effective approach for speech, (Koduru et all., 2020). This is due to the filtration of sound being shaped by the vocal tract, which determines the output sound. MFCCs take the human perception sensitivity with respect to the frequencies considered. In this research, the first 12 coefficients were extracted. While LMS is generated by the Short-Term Fourier Transform (STFT) of windowed audio. The calculated magnitude is mapped to the Mel-scale to obtain the spectrogram, (Yenigalla et all., 2018).

As researchers haven't reached a consensus in which feature represents the signal best, a combination is often implemented to obtain the most information possible. Though, due to the nature of the 2D models only features that output 2D formats can be considered. This restricts the features implemented in the rest of the models. Thus, for comparison reason the same features must be implemented across all models.

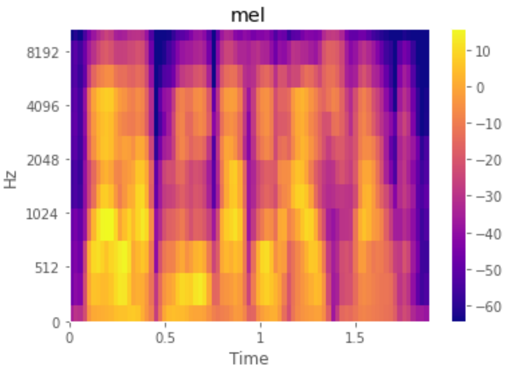


Figure 6 - Spectral Domain, LMS of one audio file (Emo-DB, 12a01Fb.wav)

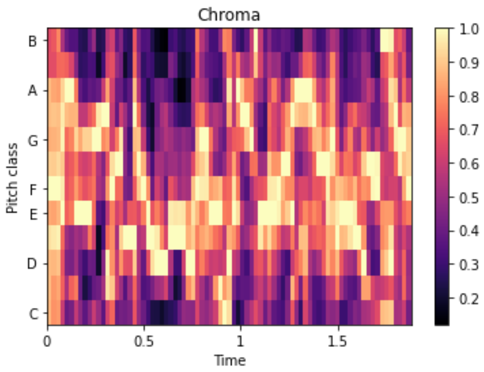


Figure 7 - Time Domain, Chroma Energy of one audio file (Emo-DB, 12a01Fb.wav)

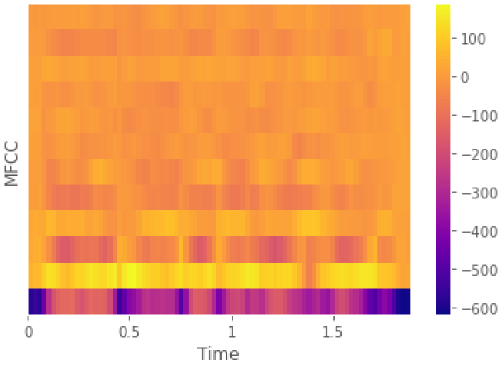


Figure 8 - Spectral Domain, MFCC of one audio file (Emo-DB, 12a01Fb.wav)

5.3.3. Data split

Before training any model, the data was processed to estimate the performance of the algorithms. Two subsets were created based on this dataset split, training and testing. The first subset trained and fitted the model, training set. The second dataset, testing set, is provided to the model, which is used to evaluate the performance of the model by making prediction and comparing them to the expected values. The training and testing set were split with a ratio of 75 to 25%, respectively.

This technique can be used for classification and regression models to improve accuracy and decrease bias when implementing the training and testing parameters, thus making it not only suitable, but also essential for this problem.

5.3.4. Feature scaling

Feature Scaling is a method used to standardize the scope of variables within a dataset bringing the values to the same magnitude, as the original scale variables vary in weight, especially when encountering a dataset with a wide range of features. To contradict the weight issue, there are two main approaches, Standardisation or Z-core normalization, which rescales the data to have a mean of 0 and a deviation of 1 and Normalization that rescales it into a range of [0, 1].

Even though normalization includes different variations, the focus was on Standardisation, which according to Böck et all, (2017), performed better in a context of SER when compared to other methods.

5.3.5. Label encoders

After the Standardisation, the label dataset is required to be encoded. These convert categorical data into numeric for a better understanding on the model's part, where each label has a correspondent unique identifier. After the model is fitted and predictions calculated, the labels from the prediction set are reversed to their respective categorical label to be compared to the original label.

5.4. Models

In earlier efforts to recognize emotions from the speech signal, several models were implemented. The most common being SVM and LSTM. However, with new Machine Learning developments in neural networks, more complex models such as CNNs have been the topic of research recently.

5.4.1. SVM

Support Vector Machines (SVM) is a very efficient binary classification and regression algorithm used for classification and pattern recognition. The method is similar to supervised learning algorithms that involves feature extraction, thus making it widely used for emotion recognition. It constructs a N-dimensional hyperplane, separating different categories of the input data into optimal categories to differentiate values based on a specific characteristic, (Jain et all., 2020).

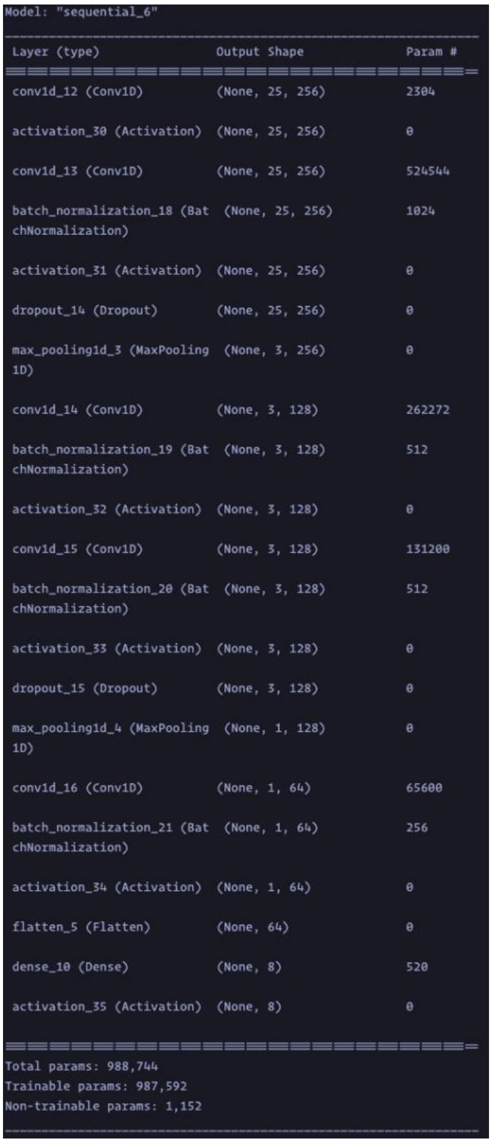
In this research, the SVM was one of the algorithms used to compared data augmentations tools. Therefore, five instances of the same algorithm were processed, SVM with raw data, SVM with noise, SVM with pitch, SVM with shift and SVM with speed&pitch.

5.4.2. CNN

Convolutional neural networks (CNNs) are regularised versions of Multilayer Perceptron (MLP) since they take advantage of the hierarchy in data to construct complex patterns derived from simpler and smaller ones. This is accomplished by using local connectivity and weight sharing to train the data, which reduces the quantity of parameters, therefore; simultaneously reducing the likelihood of overfitting, which is a common problem in Machine Learning.

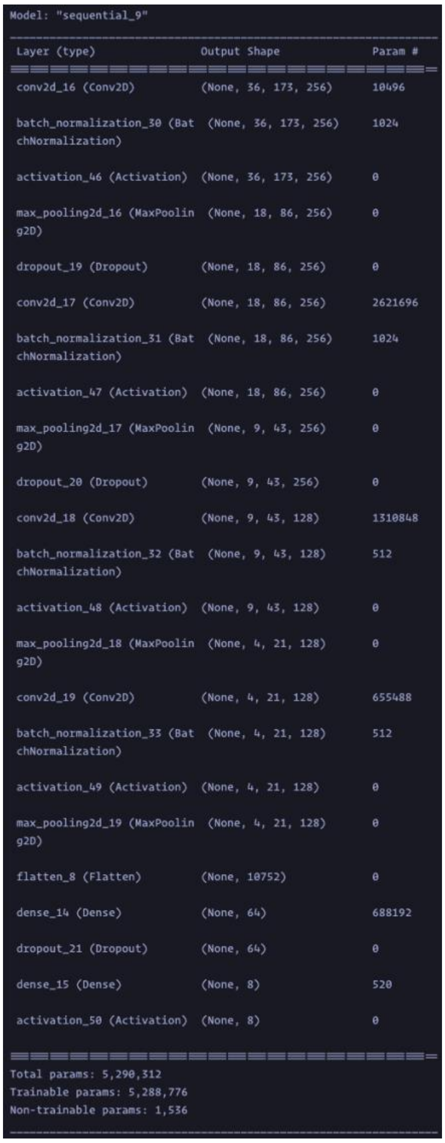
This type of neural networks can be applied according to the complexity input layer, 1D or 2D, and other specifications. 1D CNN is implemented for time series data and relies on a WxC filter. This filter slides across the W direction, where W represents the time and C the multivariate dimension. Figure 9 demonstrates the architecture of the model, where 4 1D Convolutional layers can be observed, followed by 2 1D Max-polling and 6 Activation layers. Dropout layers are also included in this model to avoid overfitting. Lastly a Dense layer is added with 8 neurons for a precise prediction.

2D CNN uses the same dimensions as 1D CNN with the integration of an addition plane of weights. This requires the input to also have an extra layer and is observable in figure 10, where it implements a 4 2D Convolutional layers, 4 2D Max-polling and 5 Activation. Lastly, 2 Dense layers, 4 Batch Normalization, and 3 Dropout layers complement the model.



```
Model: "sequential_6"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d_12 (Conv1D)          (None, 25, 256)           2304

 activation_30 (Activation)  (None, 25, 256)           0

 conv1d_13 (Conv1D)          (None, 25, 256)           524544

 batch_normalization_18 (Bat  (None, 25, 256)          1024
 chNormalization)

 activation_31 (Activation)  (None, 25, 256)           0

 dropout_14 (Dropout)        (None, 25, 256)           0

 max_pooling1d_3 (MaxPooling  (None, 3, 256)           0
 1D)

 conv1d_14 (Conv1D)          (None, 3, 128)            262272

 batch_normalization_19 (Bat  (None, 3, 128)           512
 chNormalization)

 activation_32 (Activation)  (None, 3, 128)            0

 conv1d_15 (Conv1D)          (None, 3, 128)            131200

 batch_normalization_20 (Bat  (None, 3, 128)           512
 chNormalization)

 activation_33 (Activation)  (None, 3, 128)            0

 dropout_15 (Dropout)        (None, 3, 128)            0

 max_pooling1d_4 (MaxPooling  (None, 1, 128)           0
 1D)

 conv1d_16 (Conv1D)          (None, 1, 64)             65600

 batch_normalization_21 (Bat  (None, 1, 64)            256
 chNormalization)

 activation_34 (Activation)  (None, 1, 64)             0

 flatten_5 (Flatten)         (None, 64)                0

 dense_10 (Dense)            (None, 8)                 520

 activation_35 (Activation)  (None, 8)                 0

=================================================================
Total params: 988,744
Trainable params: 987,592
Non-trainable params: 1,152
_________________________________________________________________
```

Figure 9 - 1D CNN (keras.sequential)



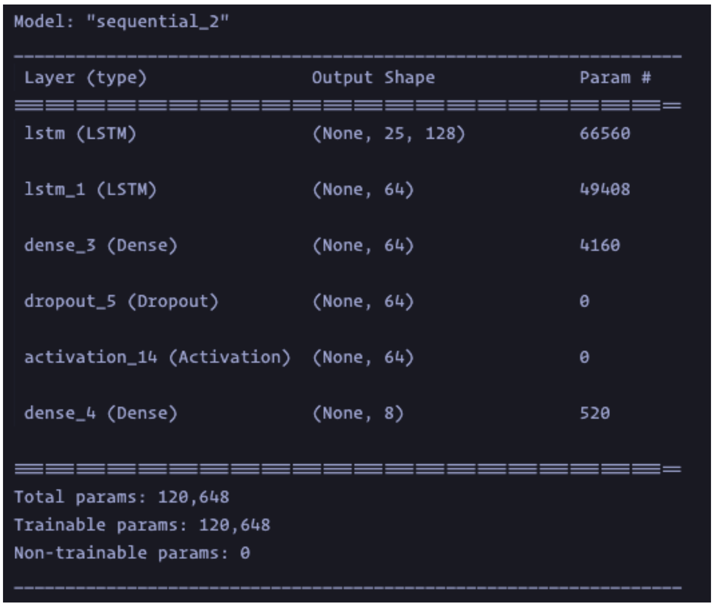
```
Model: "sequential_9"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_16 (Conv2D)          (None, 36, 173, 256)      10496

 batch_normalization_30 (Bat  (None, 36, 173, 256)     1024
 chNormalization)

 activation_46 (Activation)  (None, 36, 173, 256)      0

 max_pooling2d_16 (MaxPoolin  (None, 18, 86, 256)      0
 g2D)

 dropout_19 (Dropout)        (None, 18, 86, 256)       0

 conv2d_17 (Conv2D)          (None, 18, 86, 256)       2621696

 batch_normalization_31 (Bat  (None, 18, 86, 256)      1024
 chNormalization)

 activation_47 (Activation)  (None, 18, 86, 256)       0

 max_pooling2d_17 (MaxPoolin  (None, 9, 43, 256)       0
 g2D)

 dropout_20 (Dropout)        (None, 9, 43, 256)        0

 conv2d_18 (Conv2D)          (None, 9, 43, 128)        1310848

 batch_normalization_32 (Bat  (None, 9, 43, 128)       512
 chNormalization)

 activation_48 (Activation)  (None, 9, 43, 128)        0

 max_pooling2d_18 (MaxPoolin  (None, 4, 21, 128)       0
 g2D)

 conv2d_19 (Conv2D)          (None, 4, 21, 128)        655488

 batch_normalization_33 (Bat  (None, 4, 21, 128)       512
 chNormalization)

 activation_49 (Activation)  (None, 4, 21, 128)        0

 max_pooling2d_19 (MaxPoolin  (None, 4, 21, 128)       0
 g2D)

 flatten_8 (Flatten)         (None, 10752)             0

 dense_14 (Dense)            (None, 64)                688192

 dropout_21 (Dropout)        (None, 64)                0

 dense_15 (Dense)            (None, 8)                 520

 activation_50 (Activation)  (None, 8)                 0

=================================================================
Total params: 5,290,312
Trainable params: 5,288,776
Non-trainable params: 1,536
_________________________________________________________________
```

Figure 10 - 2D CNN (keras.sequential)

5.4.3. LSTM

LSTM is a recurrent neural network (RNN) architecture with multiple hidden layers. This was adopted with the aim of learning the long-term contextual dependencies of sequencies, such as temporal information. The LSTM removes or adds information to the block state using four components: an input gate, an output gate, a forget gate and a cell with a self-recurrent connection.

The LSTM architecture in figure 11 applies 2 LSTM layers, one Activation, one Dropout and 2 Dense layers, one in the middle, after the first LSTM layer and one at the end.

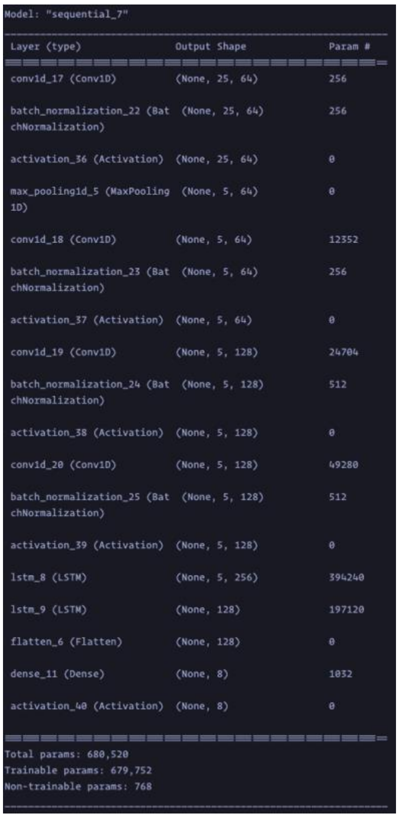


```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 25, 128)           66560

 lstm_1 (LSTM)               (None, 64)                49408

 dense_3 (Dense)             (None, 64)                4160

 dropout_5 (Dropout)         (None, 64)                0

 activation_14 (Activation)  (None, 64)                0

 dense_4 (Dense)             (None, 8)                 520

=================================================================
Total params: 120,648
Trainable params: 120,648
Non-trainable params: 0
_________________________________________________________________
```

Figure 11 - LSTM model (keras.sequential)

5.4.4. CNN 1D + LSTM

This model is a hybrid between CNN 1D and LSTM. Similarly, to the CNN 1D, this architecture starts with a group of 1D CNN, Batch Normalisation, Activation and Max-poling layers, followed by 3 groups of 1D CNN and 2 Activation layers. The size of the kernel of the CNN remains the same throughout the model with a value of 3. After the two layers of LSTM, a Connect layer was included to finalise the model, figure 12.

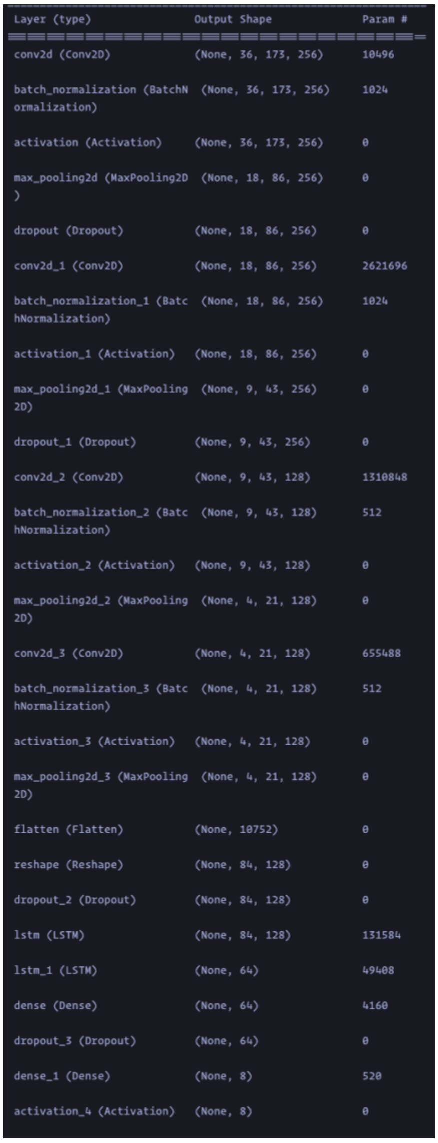


```
Model: "sequential_7"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d_17 (Conv1D)          (None, 25, 64)            256

 batch_normalization_22 (Bat  (None, 25, 64)           256
 chNormalization)

 activation_36 (Activation)  (None, 25, 64)            0

 max_pooling1d_5 (MaxPooling  (None, 5, 64)            0
 1D)

 conv1d_18 (Conv1D)          (None, 5, 64)             12352

 batch_normalization_23 (Bat  (None, 5, 64)            256
 chNormalization)

 activation_37 (Activation)  (None, 5, 64)             0

 conv1d_19 (Conv1D)          (None, 5, 128)            24704

 batch_normalization_24 (Bat  (None, 5, 128)           512
 chNormalization)

 activation_38 (Activation)  (None, 5, 128)            0

 conv1d_20 (Conv1D)          (None, 5, 128)            49280

 batch_normalization_25 (Bat  (None, 5, 128)           512
 chNormalization)

 activation_39 (Activation)  (None, 5, 128)            0

 lstm_8 (LSTM)               (None, 5, 256)            394240

 lstm_9 (LSTM)               (None, 128)               197120

 flatten_6 (Flatten)         (None, 128)               0

 dense_11 (Dense)            (None, 8)                 1032

 activation_40 (Activation)  (None, 8)                 0

=================================================================
Total params: 680,520
Trainable params: 679,752
Non-trainable params: 768
_________________________________________________________________
```

Figure 12 - 1D CNN - LSTM Model (keras.sequential)

5.4.5. CNN 2D + LSTM

Similarly to the previous model, this is also an hybrid architecture. However, the CNN layers are 2D instead of 1D, forming 4 groups of 2D CNN, Batch Normalisation, Activation and Max-polind2D. The first two groups also include a Dropout layer at the bottom of the stack. Following a Connective layer. After the CNN architecture, LSTM was incorporated twice with 2 more Dense layers and a Dropout, figure 13.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 36, 173, 256)      10496

batch_normalization (BatchN  (None, 36, 173, 256)      1024
ormalization)

activation (Activation)      (None, 36, 173, 256)      0

max_pooling2d (MaxPooling2D  (None, 18, 86, 256)       0
)

dropout (Dropout)            (None, 18, 86, 256)       0

conv2d_1 (Conv2D)            (None, 18, 86, 256)       2621696

batch_normalization_1 (Batc  (None, 18, 86, 256)       1024
hNormalization)

activation_1 (Activation)    (None, 18, 86, 256)       0

max_pooling2d_1 (MaxPooling  (None, 9, 43, 256)        0
2D)

dropout_1 (Dropout)          (None, 9, 43, 256)        0

conv2d_2 (Conv2D)            (None, 9, 43, 128)        1310848

batch_normalization_2 (Batc  (None, 9, 43, 128)        512
hNormalization)

activation_2 (Activation)    (None, 9, 43, 128)        0

max_pooling2d_2 (MaxPooling  (None, 4, 21, 128)        0
2D)

conv2d_3 (Conv2D)            (None, 4, 21, 128)        655488

batch_normalization_3 (Batc  (None, 4, 21, 128)        512
hNormalization)

activation_3 (Activation)    (None, 4, 21, 128)        0

max_pooling2d_3 (MaxPooling  (None, 4, 21, 128)        0
2D)

flatten (Flatten)            (None, 10752)             0

reshape (Reshape)            (None, 84, 128)           0

dropout_2 (Dropout)          (None, 84, 128)           0

lstm (LSTM)                  (None, 84, 128)           131584

lstm_1 (LSTM)                (None, 64)                49408

dense (Dense)                (None, 64)                4160

dropout_3 (Dropout)          (None, 64)                0

dense_1 (Dense)              (None, 8)                 520

activation_4 (Activation)    (None, 8)                 0
```

Figure 13 - 2D CNN - LSTM Model (keras.sequential)

5.5. Evaluation criteria

To evaluate the performance of each model, accuracy, loss of prediction of testing and training through time is considerate in a graphical format, along with the confusion matrix, the predictions, and the accuracy score.

6. Results

6.1. SVM

As mentioned before, four data augmentation techniques were tested: noise, shift, pitch, and speed&pitch. Thus, five confusion matrices and models were created, one for each data augmentation techniques, and an additional one with raw data as test control. By the figures from 14 to 18 and their respective accuracies, it is visible that the highest effectiveness rate is 53.14% when noise is injectant into the data

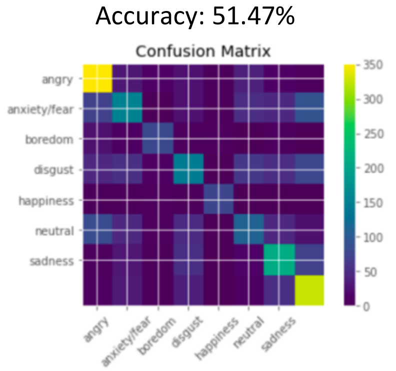


Figure 14 - Confusion Matrix of SVM with raw data

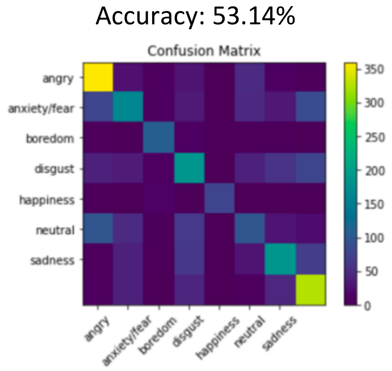


Figure 15 - Confusion Matrix of SVM with noise

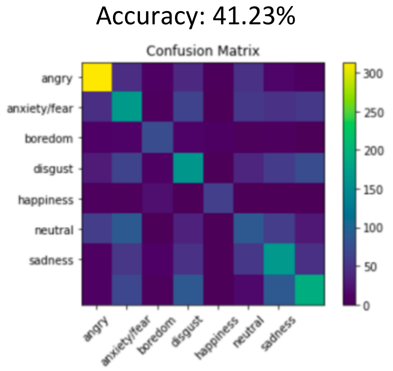


Figure 16 - Confusion Matrix of SVM with shift

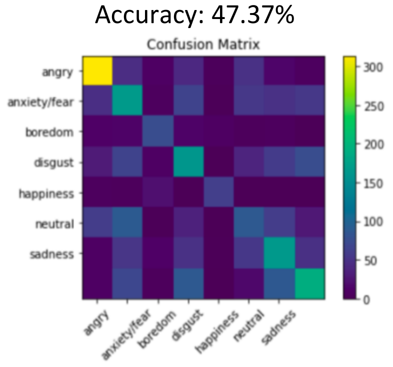


Figure 17 - Confusion Matrix of SVM with raw pitch

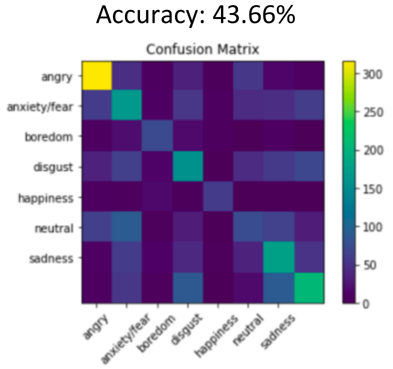


Figure 18 - Confusion Matrix of SVM with speed&pitch

6.2. CNN 1D

6.2.2. Raw data

In this model nothing will be applied to the data, serving as a test control group to compare and isolate the different approaches and their impact on the data. In figure 19, the accuracy of the CNN 1D model is calculated, as well as the loss, obtaining 56.77% and 1.12. Additionally, the confusion matrix was calculated, figure 20, as it is an essential part of a research to serve as a basis of comparison for the data augmentation techniques.

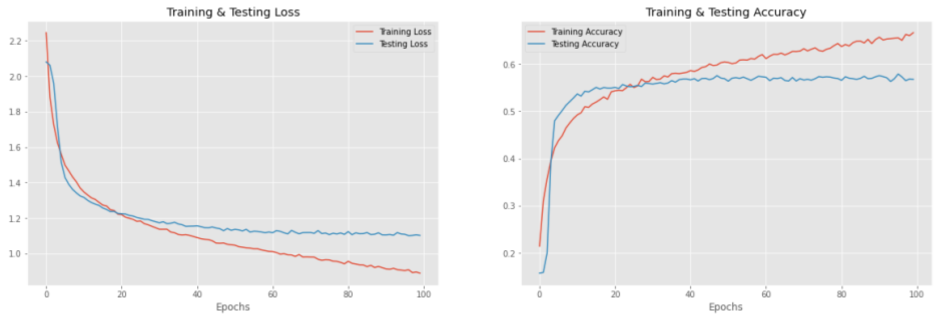


Figure 19 – CNN 1D train/test loss and train/test accuracy of raw data

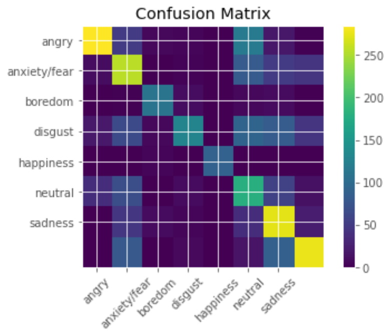


Figure 20 - Confusion Matrix of CNN 1D of raw data

6.2.3. Data with Noise

After training with ~100 epochs, the model reached an average of 46.50% in accuracy, table 2, and a loss of 1.32. In figure 21, the process of the training and testing loss can be compared. Both datasets present a similar curvature shape; however, the testing set expresses better results. Figure 22 demonstrates a Confusion Matrix confirming the feature distribution of the y_test and the predictions.

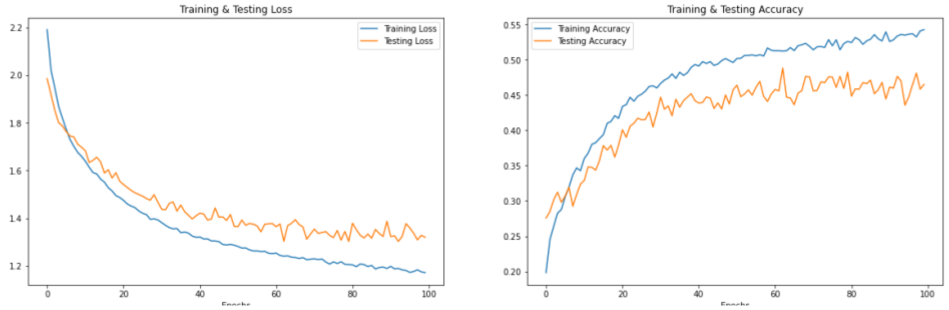


Figure 22 – CNN 1D train/test loss and train/test accuracy of data with Noise

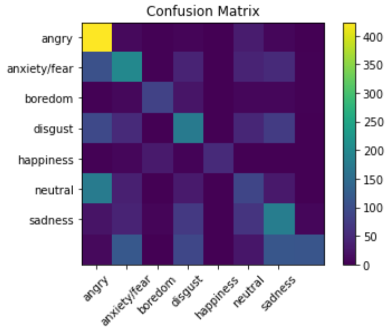


Figure 21 - Confusion Matrix of CNN 1D of data with Noise

6.2.4. Data with Shift

This model, like the previous, applies data training with ~100 epochs. This data augmentations tool coupled with CNN 1D model reached an average of 39.00% and 1.65 in loss, figure 23. Compared to the previous one, the data with shifting didn't reach the expected results, showing a difference of 7.50% of accuracy and 0.33 in loss. This that be verified by the confusion matrix, that shows non-identifiable labels, figure 24.

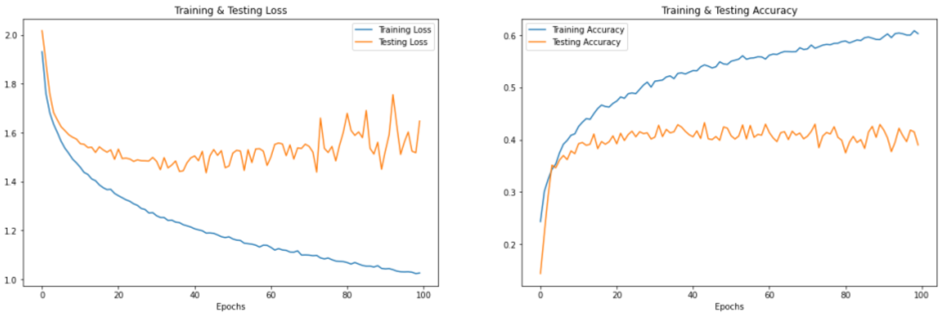


Figure 23 – CNN 1D train/test loss and train/test accuracy of data with Shift

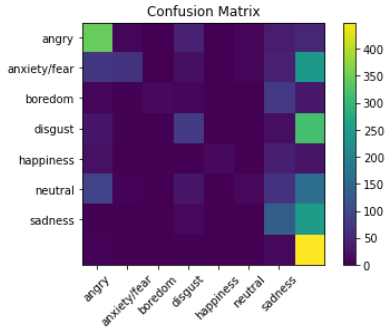


Figure 24 - Confusion Matrix of CNN 1D of data with Shift

6.2.5. Data with Pitch

This technique used in a CNN 1D model for data augmentation was the least effective, performing with 35.67% overall and a loss of 1.83. This is represented in the graphs of figure 25. The lack of effectiveness of the data augmentation technique, pitch, can also be verified in the confusion matrix, figure 26, as there are a lot of features that were not considered, altering the prediction effectiveness of the learning algorithm.

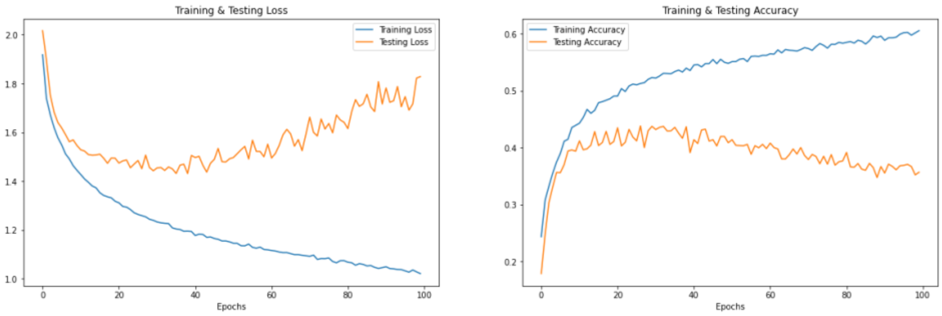


Figure 25 – CNN 1D train/test loss and train/test accuracy of data with Pitch

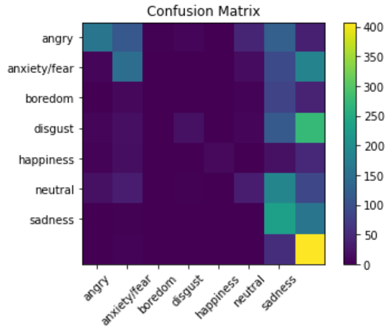


Figure 26 - Confusion Matrix of CNN 1D of data with Pitch

6.2.6. Data with Speed&Pitch

Lastly, CNN 1D using the speed&pitch data augmentation approach was observed. Even though, speed&pitch reached one of the highest accuracies with 46.35% and loss 1.35, figure 27. The efficacy of this is not ideal, deviating itself from the testing set as the model processed. The Confusion Matrix, figure 28, shows an overview of the prediction emotion distribution.

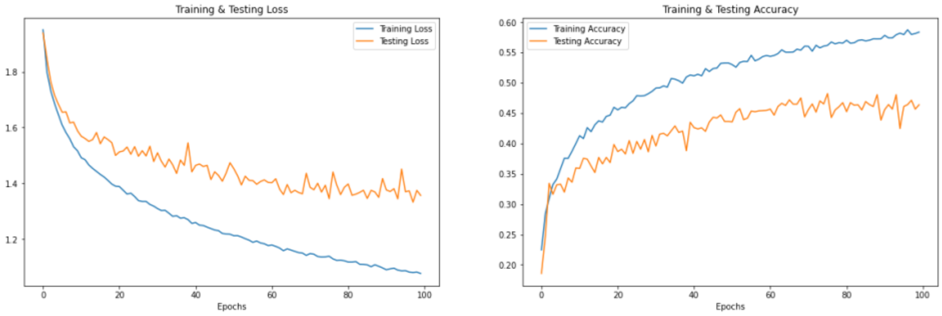


Figure 27 – CNN 1D train/test loss and train/test accuracy of data with Speed&Pitch

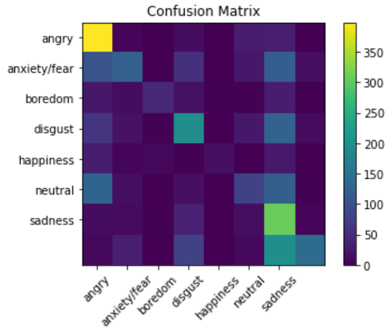


Figure 28 - Confusion Matrix of CNN 1D of data with Speed&Pitch

As observed in the graphs above and summarized in the table 2, the model with the best accuracy is the CNN combination with the noise injection, scoring 46.50%, following speed&pitch and shift, and lastly the pitch method. However, the model with lack of augmentation features in this model reaches higher accuracies. Therefore, the dataset with the raw data is applied in the rest of the models for evaluation.

Model	Accuracy	Loss
CNN 1D with raw data	56.77%	1.12
CNN 1D with Noise	46.50%	1.32
CNN 1D with Shift	39.00%	1.65
CNN 1D with Pitch	35.66%	1.83
CNN 1D with Speed&Pitch	46.35%	1.35

Table 2 - CNN 1D Accuracies and loss

6.3. CNN 2D

The 2D CNN architecture reached an accuracy of 63.48%. The model, similarly to the previous methods, used ~100 epochs, figure 29. The effectiveness of the model is observed in the confusion matrix when comparing it to confusion matrices with higher accuracies such as figure 14 and 20. However, this process took significantly more time than any other.

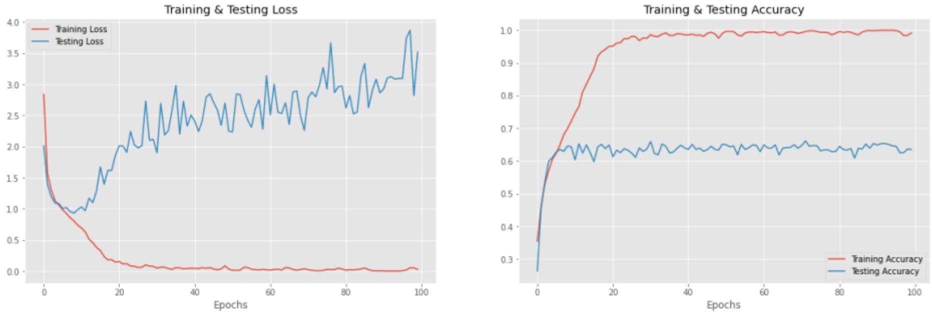


Figure 29 – CNN 2D train/test loss and train/test accuracy of raw data

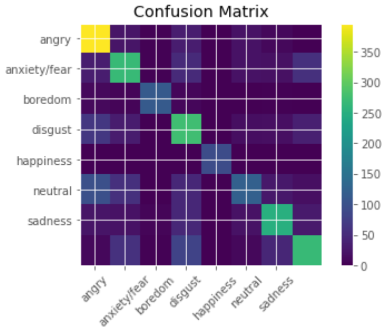


Figure 30 - Confusion Matrix of CNN 2D of raw data

6.4. LSTM

LSTM is an algorithm used when machine learning in speech emotion recognition was still in the begging of its development. LSTM has a simple and fast implementation process, thus, making it a good method for this. However, even though the graphs in figure 31 show a close relation between the test and the training set, the model didn't perform well, obtaining an accuracy of 37.35%. This can be verified with its respecting confusion matrix, figure 32. More recently, this model has been used as hybrid. Although, to eliminate any unknown variable in the comparison, LSTM had to be processed individually.

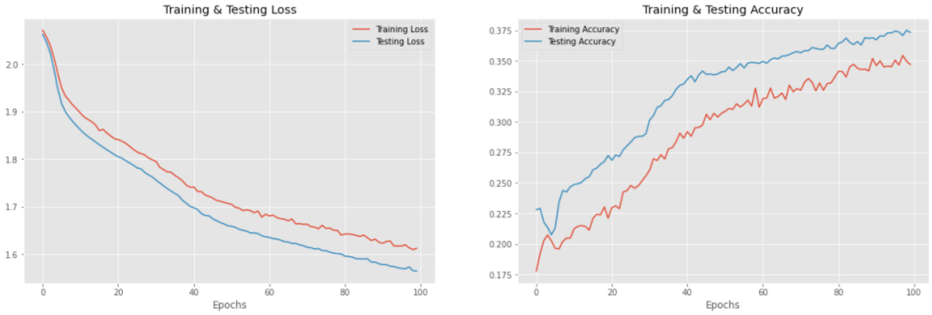


Figure 31 - LSTM train/test loss and train/test accuracy of raw data

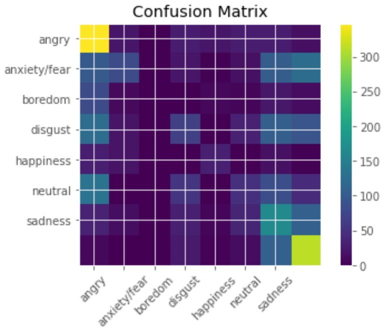


Figure 32 - Confusion Matrix of LSTM of raw data

6.5. CNN 1D + LSTM

As mentioned above, LSTM is used in hybrid system. In this model, LSTM and CNN 1D have been combined to achieve a higher precision rate. Even though figure 33 expresses an identical trajectory in the early stages of processing, they tend to disperse as its processes. This results a lower accuracy than the CNN 1D, 54.28%, but higher than the LSTM. The same effect is observed in the comparison between the training and testing loss and consequently showed in the confusion matrix in figure 34.

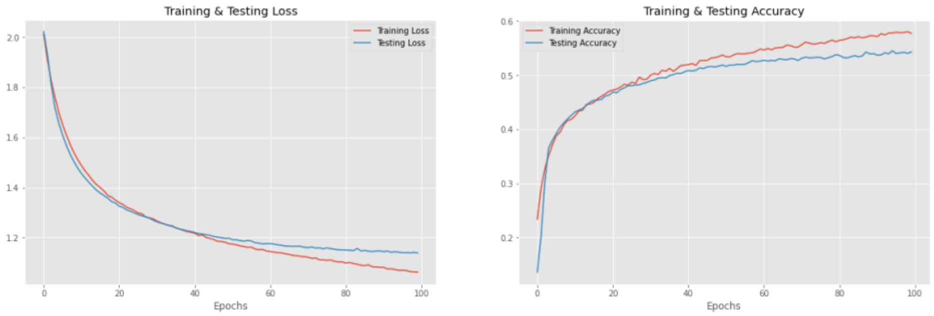


Figure 33 - CNN 1D + LSTM train/test loss and train/test accuracy of raw data

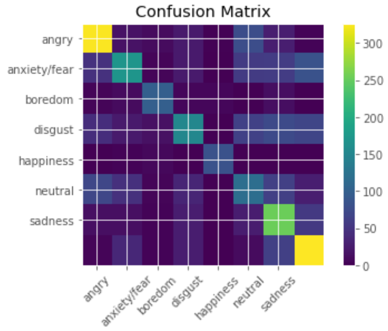


Figure 34 - Confusion Matrix of CNN 1D + LSTM of raw data

6.6. CNN 2D + LSTM

The last model is a hybrid of a CNN 2D and LSTM models. This algorithm reached an accuracy 66.61% here and loss. Figure 35 demonstrates the relation between the loss and accuracy of the training and testing set, which is similar until ~40 epochs. After the training accuracy continues to grow but the testing remains constant with limited growth. The same is observed with the loss.

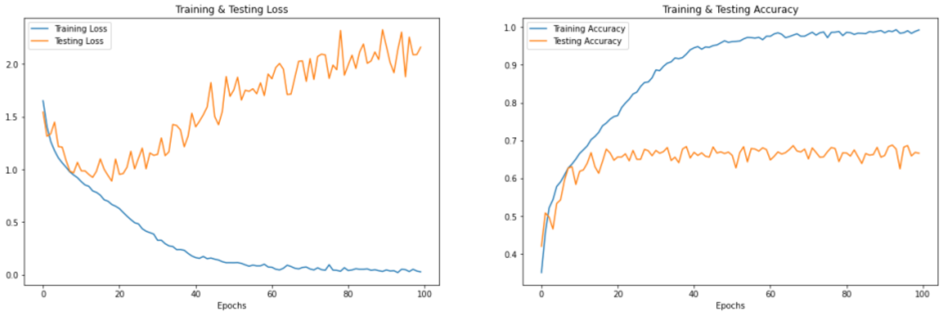


Figure 35 - CNN 2D + LSTM train/test loss and train/test accuracy of raw data

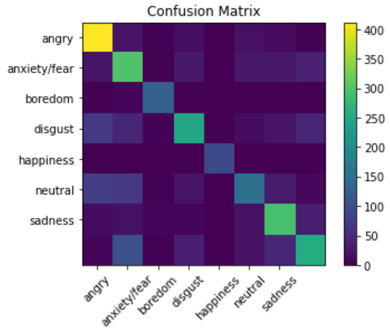


Figure 36 - Confusion Matrix of CNN 2D + LSTM of raw data

7. Discussion

The aim of this research was to demonstrate that machine learning models are able to predict emotion based on speech, while reviewing and comparing them to achieve a more effective manner of reducing inaccuracies and errors in these systems. The training and testing of individual and hybrid models with and without data augmentation techniques, while relying on four reliable audio datasets of actors expressing several emotions, such as anger, disgust, anxiety/fear, happiness, boredom and neutral; with the same transcript. All learning models reached an accuracy of over 50% except for the LSTM after trained. However, when reviewing the data augmentation techniques both from the SVM and the CNN 1D models, it is observed that noise injection is successful at improving the performance of the SVM model by approximately 2% when compared to the raw data. Though, this is not the case when implemented in the CNN 1D model, dropping approximately 10% accuracy relatively to the control data.

Machine learning algorithms has been successful in other fields before such as in facial image recognition in the context of emotion prediction and audio to text recognition. Zhang et all., in 2020, was successful at predicting emotions using image facial recognition using deep neural networks such as CNN models. Passricha and Aggarwal in the same year tested a hybrid system using LSTM and CNNs for speech recognition. Both researches inspired the use of speech in emotion recognition using hybrid systems in this project, which was later confirmed possible by Zhao, (2019) and Chen, (2018) in the context of SER processing. Both researches implemented a hybrid system effectively with a accuracy of 95.89% and 64.74%, respectively. However, Chen experimented with 3D CNNs while Zhao with 2D CNNs. However, due to the vast application methods, it is difficult to compare models as each different techniques and/or datasets for the predictions, making the results inconsistent as they are tailored to succeed in a specific situation. A comparative study demonstrates the efficiency when consistency is applied across the different algorithms.

The results of the research emphasizes the difference between data augmentation tools and original data, as well as hybrid and individual algorithms, when creating a multi classification model for emotion prediction. In contrast with Zhao's research, the model CNN 2D + LSTM in this research did not reach the same outcomes. This was due to the construction of the models. Zhao implemented four learning layers, convolutional layer, batch normalization layer, one exponential linear unit layer and a max polling layer, one LSTM, one connection layer. Even though four learning layers were also implemented in this research, the max polling layer was only used in the first learning group and an additional LSTM implemented. Furthermore, fours datasets were used, instead of one, allowing more data to be processed.

For future work, a comparison of these models with noise injection can be processed for CNNs classification algorithms and the combination of the CNNS and LSTM. More algorithms and architectures such as Inception-ResNet, InceptionTime and Inception-v4 can be compared with the same benchmarks, which should perform better, according with similar studies. Although, Inception-ResNet is commonly applied in image recognition and InceptionTime for time series classification in images. InceptionTime hasn't been experimented in a SER framework.

8. Project Management

To project manage, catalogue, and review this research Notion was chosen. Notion is a software for notetaking, task management, project management using a database and markdown pages for personalisation directed for personal and collective work.

8.1. Time management

The Researched proposal suggested a Gant chart for time management. The division of the tasks and the initial proposed duration of the research can be observed in figure 37. However due to the Pandemic situations the research had to be postponed, having a starting date on the 13th of September.

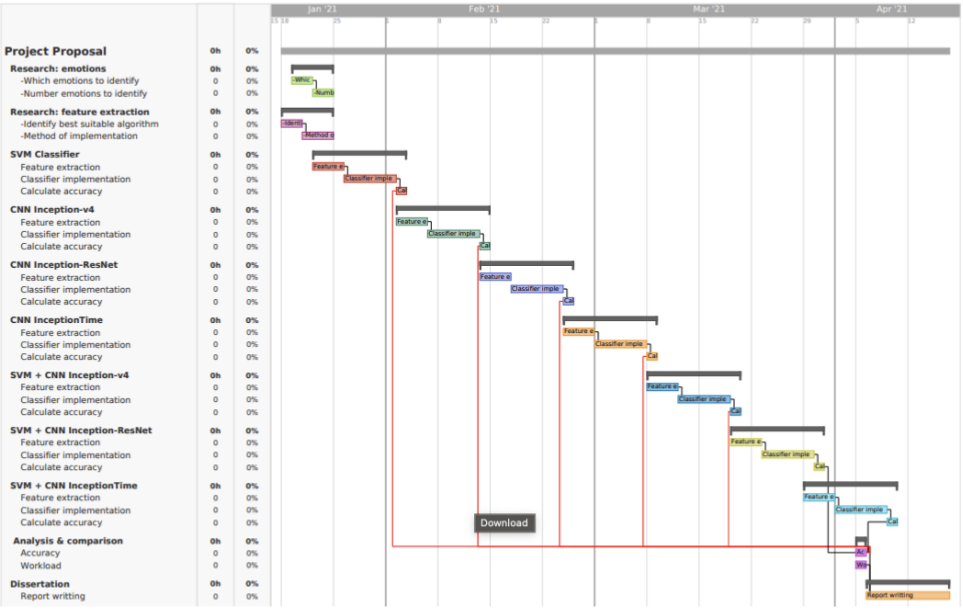
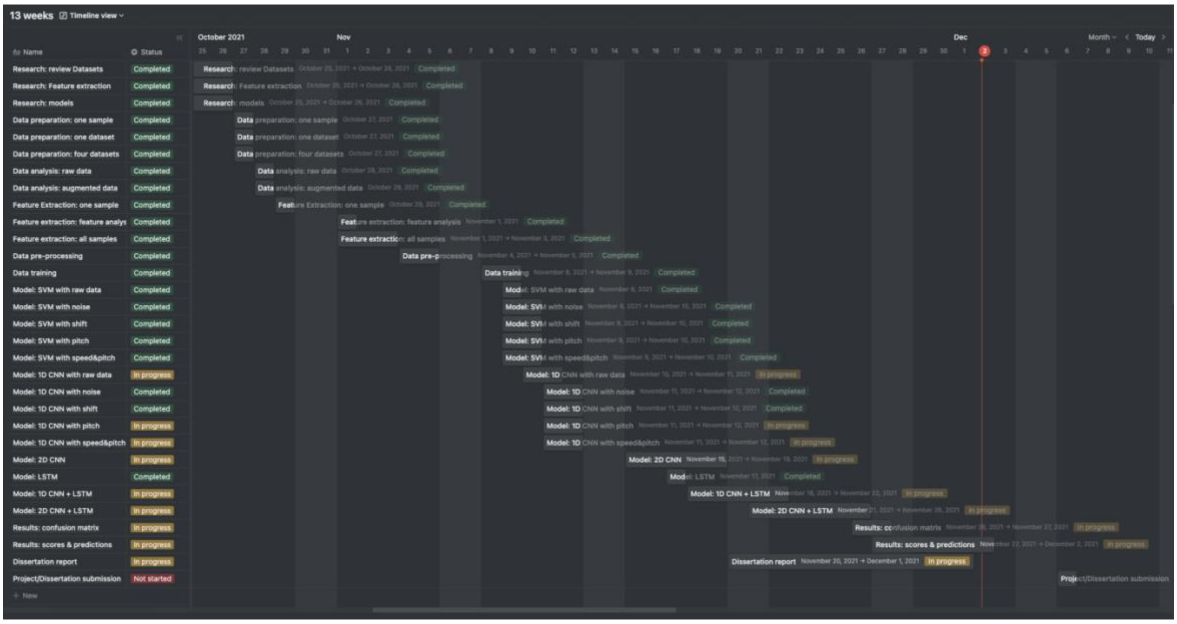


Figure 37 - Proposed Gant chart for research time management

However, after a few weeks my supervisor was reallocated and no substitution occurred, causing delays and adjustments to the project. Therefore, a new Gant chart was designed in Notion, figure 38.

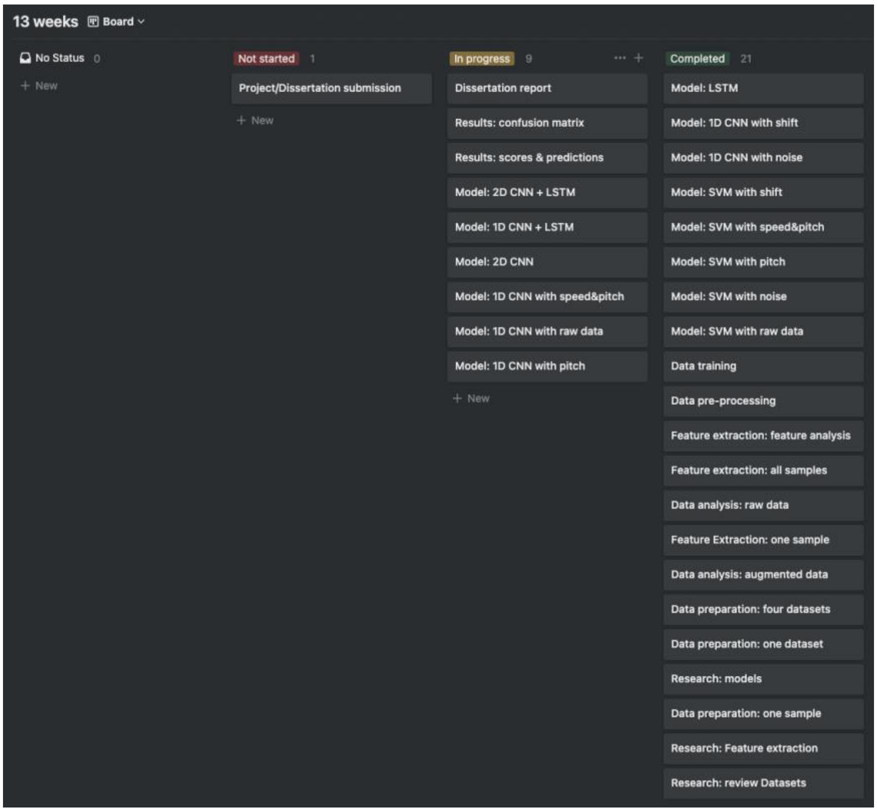


Name	Status	
Research: review Datasets	Completed	October 25, 2021 → October 26, 2021
Research: Feature extraction	Completed	October 25, 2021 → October 26, 2021
Research: models	Completed	October 25, 2021 → October 26, 2021
Data preparation: one sample	Completed	October 27, 2021
Data preparation: one dataset	Completed	October 27, 2021
Data preparation: four datasets	Completed	October 27, 2021
Data analysis: raw data	Completed	October 28, 2021
Data analysis: augmented data	Completed	October 28, 2021
Feature Extraction: one sample	Completed	October 29, 2021
Feature extraction: feature analysis	Completed	November 1, 2021
Feature extraction: all samples	Completed	November 1, 2021 → November 3, 2021
Data pre-processing	Completed	November 4, 2021 → November 5, 2021
Data training	Completed	November 8, 2021 → November 9, 2021
Model: SVM with raw data	Completed	November 9, 2021
Model: SVM with noise	Completed	November 9, 2021 → November 10, 2021
Model: SVM with shift	Completed	November 9, 2021 → November 10, 2021
Model: SVM with pitch	Completed	November 9, 2021 → November 10, 2021
Model: SVM with speed&pitch	Completed	November 9, 2021 → November 10, 2021
Model: 1D CNN with raw data	In progress	November 10, 2021 → November 11, 2021
Model: 1D CNN with noise	Completed	November 11, 2021 → November 12, 2021
Model: 1D CNN with shift	Completed	November 11, 2021 → November 12, 2021
Model: 1D CNN with pitch	In progress	November 11, 2021 → November 12, 2021
Model: 1D CNN with speed&pitch	In progress	November 11, 2021 → November 12, 2021
Model: 2D CNN	In progress	November 15, 2021 → November 19, 2021
Model: LSTM	Completed	November 17, 2021
Model: 1D CNN + LSTM	In progress	November 18, 2021 → November 22, 2021
Model: 2D CNN + LSTM	In progress	November 21, 2021 → November 26, 2021
Results: confusion matrix	In progress	November 26, 2021 → November 27, 2021
Results: scores & predictions	In progress	November 27, 2021 → December 2, 2021
Dissertation report	In progress	November 20, 2021 → December 1, 2021
Project/Dissertation submission	Not started	

+ New

Figure 38 - Reviewed Gant Chart

In addition to the Gant charts, a Kanban board was created to assist with task management within the week, figure 39. Kanban board is included in the AGILE methodology, which is a common task management tool for projects due to the columns mechanism of separating the tasks into "in progress", "completed" and "not started". This was also developed using Notion.



No Status 0	Not started 1	In progress 9	Completed 21
+ New	Project/Dissertation submission	Dissertation report	Model: LSTM
	+ New	Results: confusion matrix	Model: 1D CNN with shift
		Results: scores & predictions	Model: 1D CNN with noise
		Model: 2D CNN + LSTM	Model: SVM with shift
		Model: 1D CNN + LSTM	Model: SVM with speed&pitch
		Model: 2D CNN	Model: SVM with pitch
		Model: 1D CNN with speed&pitch	Model: SVM with noise
		Model: 1D CNN with raw data	Model: SVM with raw data
		Model: 1D CNN with pitch	Data training
		+ New	Data pre-processing
			Feature extraction: feature analysis
			Feature extraction: all samples
			Data analysis: raw data
			Feature Extraction: one sample
			Data analysis: augmented data
			Data preparation: four datasets
			Data preparation: one dataset
			Research: models
			Data preparation: one sample
			Research: Feature extraction
			Research: review Datasets

Figure 39 - Kanban board

8.2. Ethical considerations

As mentioned in the methodology, the use of a dataset is indispensable to train and test the algorithms compared in this research. All datasets are open source and were carried out by educational organizations such as the university of Toronto, Surrey, Berlin, and Pennsylvania, which employed actors for the recordings. Thus, privacy concerns do not apply as all the participants are anonymous and from different countries and cultural backgrounds.

Concerns around secondary use of data are unlikely as no personal information is contained in the audio samples due to the generic nature of the transcripts. Therefore, only knowledge regarding the characteristics of the sound can be extracted.

8.3. Problems

Due to lack of time, support, and overload of work, the proposed research couldn’t be completed as planed in the Gant chart in figure 38. As mentioned above, the supervisor allocated was repositioned at the beginning of the project and not replaced. I was advised to contact Lectures trained on the matter to assist, but unfortunately that wasn’t possible. Therefore, the research and the project management plan had to be reviewed and adapted. This forced the Inception-ResNet, InceptionTime and Inception-v4 architectures to be exclude from the project.

In addition to this research, several other modules and projects were taking place, along with a full-time job. The combination of all of these and issues related to the pandemic, the project had to be postponed. However, this allowed for a successful review, planning and investigation of most of the proposed topics.

9. Conclusion

This research concludes that machine learning was effective at determining the emotions of people. However, a lot of progress is still required to build an algorithm that is reliable in any circumstance. This is imperative due the implications of failure in a driving context. Failures can cause frustration with the driver and consequently interfering with people’s lives and creating accidents, not only for the drivers, but also passengers and other people interacting with the vehicle, instead of preventing them. Therefore, the reliability of such system is crucial.

To achieve this, several biases need to be eliminated as the world moves to a more globalized environment and more diverse cultures interact and rely on these systems. Hence, the creation of databases with diverse accents, tones and speech patterns is imperative for a more diverse patron recognition in SER systems. Thus, SER models created in this research used four datasets from different countries, Canada, US, England, and Germany, coupled with models previously studied and applied in this context to assist the research and create benchmarks for future research.

From analysing the results, a discrepancy of accuracies was detected. First, data augmentation techniques didn’t contribute as expected, with the best accuracy being with the raw data for CNN 1D models, 53.14% accuracy. In contract, noise injection performed better when applied to a SVM model, reaching an average accuracy of 56.77%. Therefore, due to the lack of reliability of data augmentation techniques in the CNN model, the following models did not include any data augmentation. These results could be improved with different model architectures paired with different augmentation tools and more accurate data in real time situations as the ones used were performed by actors in a controlled environment.

In terms of methodology and results, the model that expressed to be more reliable in their predictions was the hybrid model trained without any data augmentation and using MFCC, LMS and Chroma features, with an accuracy of66.61%. Of the several models created, two were hybrid and four original models. It was expected for the hybrid models to archive a better performance, however the original models had to be compared as well to create a control environment without unknown variables. Thus, this concludes that even though there is still research to be carried and changes to be made, comparative studies are essential to determine the most effective methods in predicting SER.

10. References

Underwood, G., Chapman, P., Wright, S. and Crundall, D. (1999) 'Anger while driving.' *Transportation Research Part F: Traffic Psychology and Behaviour* 2, 1, 55–68. DOI: https://doi.org/10.1016/S1369-8478(99)00006-6 [20 November 2021]

Zepf, S. Hernandez, J. Schmitt, A. Minker, W and W. Picard, R. (2020) 'Driver Emotion Recognition for Intelligent Vehicles: A Survey'. ACM Computing Surveys. vol. 53, no 64, pp. 1-30, DOI: 10.1145/3388790 [2 May 2021]

Cabanac M. (2002). 'What is emotion?'. *Behavioural processes*, *60*(2), 69–83. https://doi.org/10.1016/s0376-6357(02)00078-5 [21 November 2021]

Schacter, DL., Gilbert, DT. And Wegner, DM. (2011) 'Psychology (2nd Edition)'. New York: Worth. [21 November 2021]

Paul Ekman (1992) 'An argument for basic emotions', Cognition and Emotion, 6:3-4, 169-200, DOI: 10.1080/02699939208411068 [21 November 2021]

Sarprasatham, M. (2015). 'Emotion Recognition: A Survey'. International Journal of Advanced Research in Computer Science. 3. 14-19. [21 November 2021]

Koolagudi, S. and Rao, K. (2012). 'Emotion recognition from speech: a review.' Int. J. Speech Technol. 15, 2 (June 2012), 99–117. DOI: https://doi.org/10.1007/s10772-011-9125-1 [21 November 2021]

Jackson, P. and ul haq, S. (2011). 'Surrey Audio-Visual Expressed Emotion (SAVEE) database'. University of Surrey. [21 November 2021]

Pichora-Fuller, M. K. and Dupuis, K. (2010) 'Toronto emotional speech set'. University of Toronto, Psychology Department. DOI: https://doi.org/10.5683/SP2/E8H2MF. [21 November 2021]

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE transactions on affective computing*, *5*(4), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244 . [22 November 2021]

Abbaschian, B. J., Sierra-Sosa, D. and Elmaghraby, A. (2021). 'Deep Learning Techniques for Speech Emotion Recognition', Databases to Models. *Sensors (Basel, Switzerland)*, *21*(4), 1249. https://doi.org/10.3390/s21041249.[22 November 2021]

Requardt, A. F., Ihme, K., Wilbrink, M. and Wendemuth, A. (2019) 'Towards affect-aware vehicles for increasing safety and comfort: recognising driver emotions from audio recordings in a realistic driving study', IET Intelligent Transport Systems, 14, (10), p. 1265-1277, DOI: 10.1049/iet-its.2019.0732. [21 November 2021]

Reakaa, S., and Jeganathan, H. (2021). 'Comparison study on speech emotion prediction using machine learning.' Journal of Physics: Conference Series. 1921. 012017. DOI: 10.1088/1742-6596/1921/1/012017. [22 November 2021]

Özseven, T. (2019) "A novel feature selection method for speech emotion recognition", Applied Acoustics, 146, pp. 320-326. DOI: 10.1016/j.apacoust.2018.11.028. [5 December 2020]

Zhao, J. and, Lijiang Chen, X. M. (2019) ‘Speech emotion recognition using deep 1D & 2D CNN LSTM networks.’ Biomedical Signal Processing and Control, Volume 47, Pages 312-323, ISSN 1746-8094, DOI: https://doi.org/10.1016/j.bspc.2018.08.035. [22 November 2021]

Tiwari, U., Soni, M., Chakraborty, R., Panda, A and Kopparapu, S. K. (2020) ‘Multi-Conditioning and Data Augmentation Using Generative Noise Model for Speech Emotion Recognition in Noisy Conditions,’ ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7194-7198, DOI: 10.1109/ICASSP40776.2020.9053581. [22 November 2021]

Ayadi, M. E., Kamel, M. S. and Karray, F. (2011) ‘Survey on speech emotion recognition: Features, classification schemes, and databases.’, Pattern Recognition, V. 44, I. 3, pages 572-587,ISSN 0031-3203, DOI: https://doi.org/10.1016/j.patcog.2010.09.020. [23 November]

Müller, M. (2021) ‘Fundamentals of Music Processing Using Python and Jupyter Notebooks.’, 2nd edition, 495 p., ISBN: 978-3-030-69807-2 [24 November]

Koduru, A., Valiveti, H. and Budati, A. (2020). ‘Feature extraction algorithms to improve the speech emotion recognition rate.’, International Journal of Speech Technology. DOI:10.1007/s10772-020-09672-4. [24 November 2021]

Böck, R., Egorow, O., Siegert, I and Wendemuth, A. (2017). ‘Comparative Study on Normalisation in Emotion Recognition from Speech’. DOI: 10.1007/978-3-319-72038-8_15. [25 November 2021]

Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S., & Vepa, J. (2018). ‘Speech Emotion Recognition Using Spectrogram & Phoneme Embedding.’ *INTERSPEECH*. DOI: DOI:10.21437/Interspeech.2018-1811. [25 November 2021]

Jain, M., Narayan, S., Balaji, P., BharathK., P., Bhowmick, A., Karthik, R., & Muthu, R.K. (2020). Speech Emotion Recognition using Support Vector Machine. *ArXiv, abs/2002.07590*. [25 November 2021]

Zhang, W., He, X. and Lu, W. (2020) "Exploring Discriminative Representations for Image Emotion Recognition with CNNs," in *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 515-523, DOI: 10.1109/TMM.2019.2928998.

Passricha, V. & Aggarwal, R. (2020). A Hybrid of Deep CNN and Bidirectional LSTM for Automatic Speech Recognition. *Journal of Intelligent Systems*, *29*(1), 1261-1274. DOI:https://doi.org/10.1515/jisys-2018-0372

Chen, M. et al. (2018) "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition," in IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440-1444, DOI: 10.1109/LSP.2018.2860246.