University of St Andrews

Exploring uncertainty in the input level in multimodal emotion recognition

Supervised by Dr Juan Ye

Author: Katerina Montoban (Stankova)
Master's Thesis in Artificial Intelligence under Computer Science Departement, 09/2020

Abstract

Emotion Recognition (ER) has been a hot topic already for some time as it is crucial for building truly intelligent systems which will acquire a certain level of emotional intelligence within them. Although, a lot of progress has been achieved in ER (i.e. multimodal integration for ER), uncertainty in Artificial Intelligence (AI) advances is a serious issue which slows down this progress and degrades prediction reliability. Existing techniques for addressing uncertainty are not sufficient to differentiate the type of uncertainty nor to properly address its negative impact on performance. Furthermore, there is a lack of exploratory research on specific types of uncertainty and experimental upgrades of current methods with attempt to address it. Therefore, this project attempts to provide a brief introduction to ER, conduct extensive experiments to explore the effects of uncertainty caused by noise in the input data in multimodal integration (namely Audio-Video emotion recognition (AVER). Secondary objectives is to propose possible solutions to address some of the negative impacts on the performance caused by uncertainty. This will be done by adopting existing codes for multimodal, unimodal both audio and visual baseline architectures, and noise augmentation in the test split. Experiments suggest that selected AVER model is not robust to the noise in the audio nor video input modality. Furthermore, noise in one modality may also have a negative influence on the 'clean' modality. In summary, it is suggested that current AI methods would benefit from additional information manipulation which would result in better uncertainty handling and support the potential of more robust multimodal integration.

Acknowledgements

This project is a combination of my various interests that I wish to explore and gain knowledge about. Thanks to my supervisor, I was able to do so. I am very grateful for that because I learnt even more than I expected, including the 'dark' side of this topic: to be critical and not to accept the accuracy of models as the ultimate truth. Not only was this project very challenging in sense of the complexity and necessary knowledge to gain before even I was able to touch the project's purpose but also because of the pandemic Covid-19. The uncertain times had a negative impact on my progress on this project and my daily life. Nonetheless, my accomplishments would not be possible without the support and cheerful motivation of my friends from around the world.

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence. i, iii, 1, 4–8, 10

AVER Audio-Video emotion recognition. i, iii, vi–x, 2, 8, 13, 27, 28, 30–33, 35, 38–45, 47, 48, 50–58, 60–64, 68–71, 74–87, 89, 90, 93, 94, 96–101

CNN Convolutional Neural Network. 6, 13

CREMA-D Crowd-Sourced Emotinal Multimodal Actors Dataset. iv, v, 9, 18–20

DL Deep Learning. iii, 1, 2, 5, 6, 13, 35

ER Emotion Recognition. i, iii, 2, 4–11, 27, 35

FE Feature Extraction. v, 19, 20

FER Facial expression/emotion recognition. iii, v–x, 2, 6, 7, 11, 16, 17, 27–29, 31–33, 38–42, 45–47, 58, 59, 64–68, 88, 91, 92

HCI Human Computer Interaction. 1, 2

LSTM Long Short Term Memory. 9

MFCCs Mel Frequency Cepstral Coefficients. v, 6, 14, 21, 22

ML Machine Learning. iii, 1, 6, 35

MLP multi layer perceptron. 14

RAVDESS Ryerson Audio-Visual Database of Emotinal Speech and Song. iv, v, 6, 9, 18, 21

ReLU Rectified Linear Unit. 13, 14

SER Speech emotion recognition. iii, v–x, 2, 6, 11, 17, 27–33, 38–42, 49, 50, 59, 60, 71–74, 88, 94, 95

SVM Support Vector Machines. 13

CHAPTER 1

Introduction

Uncertainty in ML and DL - general overview

In the field of AI, ML and DL are widely used techniques to build automated mathematical models to solve tasks across various fields (e.g. in Human Computer Interaction (HCI), data science, biometrics, multimedia data analysis, and bioinformatics) [54], [31] [63]. For example, ML and DL models are suitable for pattern recognition because they have the ability to learn underlying regularities in the observed data and apply this knowledge to new unseen instances [15], [19] [29]. However, models' prediction is an estimation rather than an ultimate truth [15], [19], this means that uncertainty is inherited within the models and it is inevitable. Moreover, there are various kind of uncertainty which influences current advances in AI (see Chapter 2) [15], [36]. Importantly, uncertainty can have either a negative impact on the model's performance [35], [36] or it may enhance its generalisability if it is properly handled [24], [35]. There is no doubt that uncertainty plays an important role in ML techniques [6], [24], [35], [15] and further, that it has in recent years received increasing attention from the scientific community due to the new challenges that it represents [15], [36].

Challenges with uncertainty

Although, probability theory is usually used as the fundamental method for addressing uncertainty in ML and DL techniques it is restricted [15] [19] and fails to fulfil expectations in some areas. For instance, current techniques in general cannot distinguish between different resources of uncertainty nor its levels. [15], [36]. Secondly, models usually report only measures such as an average accuracy and/or in-overall confidence of the models. This may be problematic when such models will be applied to systems for use in a real-life settings because the user will be interested in the reliability of a specific prediction obtained at the real time rather than in the quality of the model in general [32] [15]. These issues are as similarly apparent in medicine as they are in the area of autonomous vehicles or in social robotics where the reliability of the prediction is crucial [15] [42]. Lastly, probabilistic ML models usually rely on strong assumptions about the underlaying distributions [32] which do not always lead to the correct classification even though the reliability of the reported prediction can be high. This shows that it is important to demonstrate what the model does not know because if there are two or more classes with similar output values, then the

model will mathematically prefer the highest one even though it is a wrong class. For instance, Kendall and Gal [35], effectively demonstrated this problem with their DL model for image classification in their study.

Handling uncertainty is highly demanding in HCI because it has a significant impact on the progress in computer ER [42] [36]. The negative effect is not only due to the inherited uncertainty in human emotion and emotional expression in general, [63] but crucially also due to the factors which influence the sensory inputs [36] and others. These factors are discussed in detail in the Chapter 2.

Challenges in multimodal integration

Through research it has been found that different modalities in multimodal architectures have some overlapping and some new, additional information considered crucial for recognition. Here, the overlapping information is important because it represents the potential of multimodal models to address the uncertainty that is caused by environmental conditions such as noise in the input. This means that when one of the modalities is faulty the other remaining sensors could compensate for the loss of the faulty one [49], [54]. Note: the terms modality and a sensor will be used interchangeably throughout this paper.

Nevertheless, the solution is not straightforward as there are a number of flaws in current multimodal architectures which need to be explored and understood before designing a new generation of classical methods that are able to reach the desired, advanced level of ER [4], [33]

Objectives of this project

The key aim of this project is to design and conduct comprehensive experiments to enable the exploration of different conditions of imperfect signals in multimodal integration and to facilitate the design of solutions to address some of the issues caused by uncertainty.

- Primary: to conduct a literature review on topics such as Uncertainty, Unimodal and Multimodal integration and related topics for instance, feature extraction and fusion, and ER Secondly, to understand the impact of imperfect signals in multisensory integration through experiments
- Secondary: to design solutions which address some of the imperfect problems.

Outline of the project

The rest of the project is covered in four following chapters:

Chapter 2 provides the relevant background to the topic and leads the reader though the structured literature review. Firstly, ER the field is introduced in general and then unimodal FER, SER and multimodal AVER various architectures are explored. This is then followed by features fusion of features extracted from individual sensors in the multimodal models. Finally, the general notion of uncertainty is considered, with subsequent

narrowing of focus to that of the aspect of uncertainty explicitly of interest to the research conducted by this study.

Chapter 3 describes the methodology of the architecture and dataset used, the software tools used for the design, parts of details about code implementation and finally, the experimental setup.

Chapter 4 shows the results in written format and also showcases various visualisations.

Chapter 5 builds on the previous chapter and covers the discussion of the findings, their evaluation, limitations of this study, future research and an overall conclusion.

Appendix A This Appendix contains experiments tables with accuracy overview

Appendix B This Appendix contains experiments Confusion Matrices of predicted and actual labels

Appendix C This Appendix contains experiments Bars correct vs incorrect predictions

Appendix D This Appendix contains experiments t-sne plots showing the class discriminability

CHAPTER 2

Literature review

2.1 ER and its importance in AI

Despite the fact that computer ER has received considerable attention from computer scientists, engineers and human computer interaction (HCI) specialists, its progress appears comparatively slow slow due to the complexity of emotions and other uncertainties in the pattern recognition process [54], [36]. Defining emotions is not a trivial task and there is a lack of agreement on this topic. Nonetheless, computer ER is highly desirable because it will allow more natural communication between humans and machines. Furthermore, scientists have presumed that emotions play an important and complex role in other cognition such as memory, perception, learning, decision making and attention [54]. Furthermore, others argue that the capability of emotion recognition is a valuable part of intelligence [53], [22], [48]. Thus, specialists agree with the notion that truly intelligent systems can be built only if a certain level of emotional intelligence within the systems is achieved [54]. This type of machine which can understand human feelings would be beneficial in social context, for example, personalised tytors, to learn user's preferences for future use, monitoring individuals' stress level or support in diagnoses of psychological disorders, in social robots, and job review which are on remote basis [48], [42], [36].

In humans, the expression of emotions is accompanied with physiological changes and certain behaviours. In particular, the multimodal nature of emotion consist of for instance, visual (e.g. facial, expression, gestures, motion), audio modatility (e.g. voice, speech from which is possible to obtain paralinguistic properties), and physiological sensory signals. One of the approaches to describe emotions based on the measurable/observable phenomena from various modalities is by labelling emotions into discrete categories such as fear, happiness, joy, sadness and others. Across the ER research of various modalities, these emotion classes are treated as universal within all individuals and cultures despite some level of deviation (individual differences, cultural habits [26]) [54]. This labelling of emotions is important for specialists in the field because this is the reason why they can approach ER as a pattern recognition task [54].

Datasets for ER

The research community developed various datasets with discrete categories such as static vs dynamic (in visual modality) [2], unimodal or multimodal,

and acted vs wild [16], [54]. In the dynamic dataset the input is video, which represents a continuous number of frames with dependencies. On the other hand, a static dataset represents a single frame. Unimodal datasets contain only one modality such as audio or visual, whereas multimodal datasests contain two or more modalities. 'Acted' datasets attempt to represents non ambiguous emotions which fit into one of the predefined categories. This kind of dataset is suitable for the current project because it has a lower degree of ambiguity in comparison to 'wild' datasets. The 'wild' dataset covers spontaneous or natural expressions from various resources which means that the instances can overlap in classes because of the nature of emotions. This ambiguity is one of the uncertainties in ER [36] which is not the interest of the study. It is important to select the right dataset and to match the needs of a study but one must be aware of the fact that not only are there different types but also that each dataset within the type is created in different conditions [39]. For the purposes of this study, the selected datasets are multimodal (audio-video) acted (namely RAVDESS and Crema-D) which are described in the more detail in the Chapter 3.

Feature Extraction

The raw representation of the data in the datasets has a high degree of complexity and can be computationally infeasible/exhaustive when processed in the raw form. Furthermore, besides the important information the raw form also contains redundant information i.e. noise. Therefore, feature extraction is crucial in ER and helps with all aforementioned challenges. The aim of this process is to extract the best patterns in the data which lead to the correct classification of emotions [28]. Ideally, after feature selection, the input should contain only relevant information linked to the desirable output and the redundant one should be eliminated. However, feature extraction is challenging and in real scenarios, it is impossible to remove uncertainty entirely [62], [43].

There is not a single right way to extract features and indeed, there are a variety of different techniques from which it is possible to choose. However, through the literature review it was found that there are two main approaches of feature extraction techniques - low-level or high-level. Both approaches are possible for any modality or even their combination, however, the specific method and its details depend on the type of modality. In this project the focus is on the input from two sensors - audio and visual, therefore, only these two modalities will be covered further in this chapter.

Low-level features

This kind of feature extraction is also called handcrafted because it require signal preprocessing before it is possible to pass to a classifier or DL architecture for further deeper extraction.

- **Low-level Visual features**
  In it visual modality to obtain handcrafted features the most common focus is either on a specific part of an image such selecting only facial features or without the narrow focus where the whole image is preprocessed the techniques are for example, Histograms of Oriented Gradients [12]or texture analysis which result in histograms of an image [45].

- **Low-level Acoustic features**
  Also in the audio modality the focus can be split between two main different areas from which features are extracted such as static where the overall audio signal is used [18] or dynamic which splits audio into frames and focuses on the temporal links between them [3] (both areas summarised by [39]). The low-level acoustic features attempt to capture characteristics such as spectral and prosodic information represented in an audio signal. Commonly used features are for example, Mel-scaled spectrogram, MFCCs, Spectral centroid, Pitch, Energy and others [39], [28], [60]

High-level features

Deep features are obtained by passing either the raw data or preprocessed features directly to an end-to-end architecture which have the ability to learn intrinsically the most important embeddings [28], [8]. High-level of feature extraction in unimodal ER (SER if the input is related to audio or FER in the case of visual modality) has been widely researched and represents a valuable foundation for multimodal research.

Features extracted from ML or DL models can achieve better results than the handcrafted which are without further extraction [43].

1. ***Unimodal architectures:***
   A couple of examples for both SER and FER are presented below.
   - **High-level Acoustic features forSER**
     - **example 1:** VGGish [25], [17]
       To note, this architecture uses a combination low and high level extraction rather than only high-level because the input for VGGish must be transferred into Log-Mel Spectrograms with other preprocessing steps before possible to be passed into the VGGish model. VGGish model is a pretrained Convolutional Neural Network offered publicly by Google which contains four convolutional layers each with activation and max pooling layer followed by two fully connected layers with also . VGGish but with a final fully connected logits layer was used on segments of the dataset IEMOCAP (audio part) with four classes (angry, happy, sad and neutral), splits in the ratio 8:1:1. They achieved accuracy of 68.7% which has been claimed as the improvement of the state-of-the-art [52].

     - **example 2:** Convolutional Neural Network (CNN) [28]
       In this example the network used for the extraction of higher level features was CNN. Same as the previous example, this is also combined feature extraction because the higher level of features was extracted from low-level features (MFCCs, Mel-scaled spectrograms, Chromagrams, Spectral contrast features and Tonnetz representation) extracted from the dataset RAVDESS of 1440 speech files. The splits were random 80% train vs 20% test set and the achieved result was 71.61% of accuracy.

- **High-level Visual features for FER**
  - **example 1:** CNN for Temporal pooling from Covariance Matrix [1], [2]
    Firstly, the first-order information is captured by CNN from aligned set of images from a video. Then, output from fully connected CNN layer is used for calcuation of Covariance Matrix and then passed to SPD net for temporal pooling to extract temporal deep features. On a wild dynamic dataset AFEW they achieved 32.5%, however, this was only a brief experimentation because the focus was rather a static dataset SFEW.

2. ***Multimodal architectures***

   Nonetheless, as previously mentioned, emotions are not only expressed through one sense but rather their combination. Thus, where it is possible to obtain information from different sensors, multimodality should be preferred over unimodality because the combination of audio and visual information complement each other and can improve models' accuracy in ER [54]. Moreover, multiple sensor inputs can help to minimise the uncertainty caused by a partial input or a noise in the data from one of the modalities. For example, the uncertainty resulted from the noise in the audio input (speech) can be decreased by visual clues (mouth movements) [49]. In addition, having more modalities may prevent a wrong estimation of multimodal classifier due to the wrong input in one of the sensors; in the Ladowska [37] study cameras which recorded a same event from different angles resulted in a various emotion classification. She found that the angle in which is video/photo of face taken matters. Although, having two or more modalities is beneficial for ER the difficulty lies in the fusion of used sensors. Thus, not only some examples of multimodal architectures but also fusion techniques are discussed.

Feature fusion

In the literature various kinds of features fusion are presented. This concatenation of information from different modalities is important and play a crucial role in ER. However, at the same time it is accompanied with various challenges. To note, only two frequently used kinds of features fusion are discussed, however, others nor specific techniques of each kind is beyond this project (for more details see summaries made by [39] or [55]

- *early fusion (see figure 2.1)*
  A preferable method used in multimodal fusion also called the fusion of the feature level. In this tactic features are extracted from each modality and are fused at the early beginning before classification. This resulting high dimensional feature vector is minimised and passed to the classifier for a final prediction. However, early fusion is highly challenging because expects already properly preprocessed extracted features of different modalities in synchronised way which allows to capture some of the existing temporal links between the modalities.

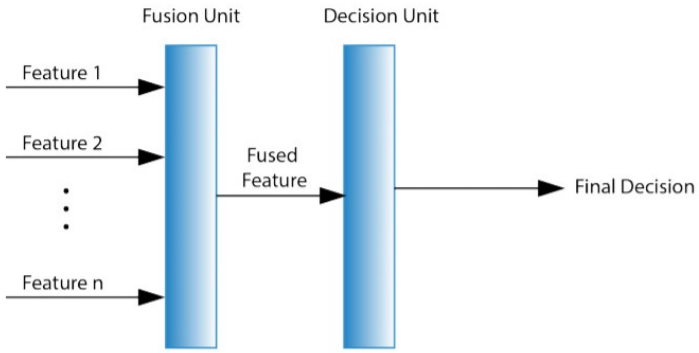

Figure 2.1: Early Fusion (feature level) figure from [55]

- *late fusion (see figure 2.2)*
  In comparison to the early fusion, the concatenation of the features happen after the decision process. Therefore, this method is also known as the decision level fusion. Each modality is processed individually by own part of the model with no existing links to the another modality. Only after the classification of the specific modality the fusion occurs. This fusion is about combining the final decisions obtained from each modality with the aim to derive to the final estimation of the architecture. However, as mentioned there is a lack of interlinks between the modalities during the process and this fact represents a significant drawback of this tactic.

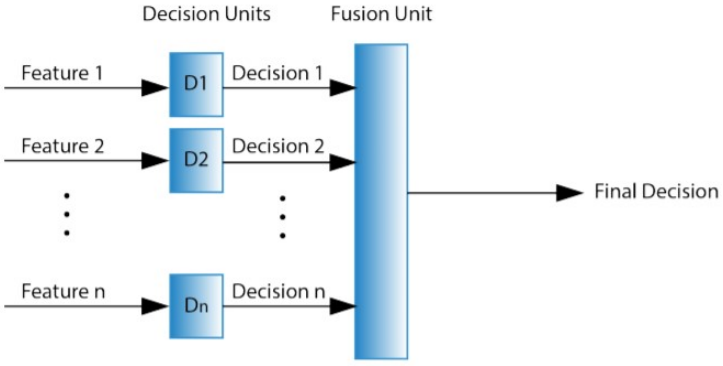

Figure 2.2: Late Fusion (decission level) figure from [55]

- **High-level multimodal features and fusion for AVER**
  - *example 1:* 3D-CNN (visual modality), SoundNet (audio modality), LSTM (for features fusion) [20]
    High level features for visual modality were extracted by 3D-CNN which represents the spatiotemporal correlations between

frames. Whereas audio features were extracted by soundNet. Then, Long Short Term Memory (LSTM) network was used for feature fusion because this method is suitable for temporal nature of the data and can capture the relationship between these two modalities over time. The cross validation was split based on actors into 10-folds. The used datasets were RAVDESS of 1440 samples (speech) and CREMA-D on which the achieved accuracy was 67.7% and the latter 74.0% with the classifier Nearest Neighbour (KNN).

– **example 2:** 3D-CNN (visual modality), 1D-CNN (audio modality), one layer fully connected neural network (fusion) [13]
The dataset used was RAVDESS of 1440 samples (speech) on which the achieved accuracy was 87.33%. However, the splits were random selection of 32 samples for each class which represented test split and the rest was in the train split.

2.2 Uncertainty in input data

There are many sources of uncertainty which have a negative impact on the models' performance (see overview from Ladowska [36]). However, the focus here is narrowed to a single source of uncertainty derived from the first stage of ER – i.e. input data. This kind uncertainty is also known as Aleatoric. Hullermier [15] argues that this uncertainty type is interlinked with others and sometimes it is difficult or impossible to distinguish the boundary between them which makes addressing it difficult. Also, he mentioned that Aleatoric uncertainty is irreducible by adding new knowledge or more data inputs.

Overall, this uncertainty resource is inherited within the data in the sense of noisy or imprecise information retrieved from one of the sensors. Notably, current models are not robust to the noise in the data and even small changes in the modality input can have a fatal influence on the accuracy or credibility of the model [15], [36], [46]. Also, this resource includes the possibility of overlapping classes in a specific area of the input space. In some cases, transferring input into a higher dimensional space can help resolve this overlapping issue; otherwise, uncertainty can lead to an inccorect estimation of the model [15]. The noise in the visual modality can be understood as an issue in cases such as images of faces in low-lighting conditions, poor quality of the image, insufficient visibility of a face (includes also wearing glasses, beard, atypical hairstyles or partial coverage of the face), camera location [37] or intentional masking (sarcasm or hiding real emotions). Then in the case of the audio modality, the uncertainty represents an issue in voice signal, namely background noise, voice distortion or intentional masking) [21], [35], [64], [36]. However, it would be sensible to include also two other resources which are in the author's understanding relevant to noise in the input. Firstly, uncertainty in how meaningful the input is ER (e.g. in visual modality expecting faces rather than a tree). Secondly, a faulty sensor in multimodal integration which does not provide any sensible information and can confuse the model [36].

These factors, which influence the quality of the input data, also determine the quality of the model's performance and its credibility. The high uncertainty

in ER persists despite the fact that current AI techniques have the capability to address or prevent some of the factors. For example, if ER an application provides instructions to the user to capture the necessary angle of the face [36]. However, this is artificially constructed with strict settings which would not be applicable in all scenarios, thus explaining a reason why there is currently an interest in wild datasets.

Cross-modal plasticity - a possible improvement in noise robustness

In a scenario such as when one of the senses is impaired in a living species, a brain is able to reorganise in order to compensate for this loss of information from an impaired modality. It is argued that remaining modalities are reorganised to enhance their performance to minimise this loss. For example, in one study it was found that when touch sense in roundworms was eliminated it enhanced their smell sense and when the touch was returned the higher smell sensitivity disappeared [50]. Also, in humans this plasticity is present, it has been found that people with impaired hearing have enhanced vision [56], whereas individuals with impaired vision have improved hearing [23]. These facts can be used as an inspiration for new ways of pre-processing data before it is passed through an architecture. For example, if there is known impairment in one of the sensors, this information could be passed to the network which would strengthen the connections of the correct modality input with higher importance, thus influencing the final prediction. This could prevent the confusion of the model by the noise resulting from the impaired modality. To mention, crossmodal plasticity was also an inspiration in the [39] dissertation to develop a new generation of models inspired biological systems. The model outperformed existing classical models and also demonstrated robustness with regard to noise in the input. It suggests that communication between modalities is important and should be present earlier than in the fusion level.

2.3 Summary

Uncertainty in ER is a hot topic full of challenges and it appears that without new ways of addressing such an issue, the credibility of the output from such systems will remain relatively low [36]. Furthermore, there is a need for new techniques which can be used to enhance multiple aspects of system credibility, rather than simply the singular aspect of model accuracy. [36], [15].

Throughout the literature review it has been found that accuracy does not always represents a reliable output measure for the effective comparison between studies due to the different ways of reprocessing. Moreover, most of the research is focused on improving accuracy rather than exploring the issue of uncertainty and its associated negative impact on performance and credibility. [36].

Despite of all the listed challenges, the significance of ER models is crucially high for AI and other fields which would benefit from the use of such systems in various applications [36].

Therefore, the contribution of this project to the research community is to provide insights into the uncertainty that can be caused by or attributed to imperfect input data. More specifically, it considers the extent to which

one sensor can compensate for a failing, or at least lack of efficiency, in others – i.e. how well it can deal with the loss of important information elsewhere, or its ability to overwrite noisy with correct sensory input. The expectations or outcome of the research may take several forms when one considers those discussed in the literature review. Firstly, modalities have overlapping information connected to the output, and it is suggested that multimodalities should compensate for each other. Secondly, it is well known that ER both unimodal and multimodal architectures are susceptible to noise in the input. This was also confirmed by Mansouri Benssassi [39].

Even if the accuracy does drop there is a possibility to upgrade the current model with additional information about the source of noise (i.e. which modality) and thus redirect the focus and power of the model to the correct sensor in order to overpower the noisy sensor (in a similar fashion to the way such compensatory activity operates within 'human models').

The experiments will be conducted in steps where firstly, appropriate architecture will be selected, understood and an attempt made to obtain accuracy in a similar level of the state-of-art of the model. In the second step, unimodal architectures for both FER and SER will be designed and trained to obtain baseline. Thirdly, an intentional noise augmentation in modalities will be performed on the test split. Lastly, results of augmented datasets will be compared with the baselines and then examined in detail in order that a meaningful conclusion may be drawn.

CHAPTER 3

Methodology and Experiment design

3.1 Introduction

This section is focused on the methodology and experimental setup of this project. The methodology covers information about the tools, the architectures, datasets, and code implementation, whereas the account of the experimental setup leads the reader through the conducted experiments step by step; this includes the implementation and characteristics of performed augmentations.

3.2 Tools

Below are listed main software tools used in this thesis and all are open source libraries for python.

- **OpenCv**:A computer vision framework containing an array of optimised algorithms for image processing and machine learning. It is a robust and well-established libary in the field of computer vision[7]
- **Torch**: Torch is a powerful machine learning framework which contains a large selection of additional libraries for a plethora of machine learning tasks including optimisation algorithms, neural networks and customisible dataset loading. [11]
- **Torchvision**: Python's computer vision library containing state of the art datasets and architectures as well as image transform algorithms. [40]
- **Librosa**: A python audio library for analysing audio and music. [41]
- **Numpy**: A numerical library for mathematical operations [**np**].
- **Sklearn**: A machine learning library containing a multitude of machine learning , model selection and preprocessing algorithms. [47]
- **Yellowbrick**: A python library for machine learning visualisations, based on scikit-learn and matplotlib. [5]
- **Skimage**: A python library containing image processing algorithms [61]
- **Matplotlib**: A library for visualisations [27]

- **PIL**: Python Imaging Library, for reading/writing/editing image files. [34]

3.3 Architectures

Three DL architectures, i.e multisensory, unisensory audio, and unisensory visual, were used to access three baselines. The purpose of both unimodal models is to enable the performance comparison of the multimodal under various experimental conditions with uncertainty. The outline of the experiments and how these models were used are explained later in this chapter. In this section, the focus will be on the in-depth description of the models and details about the implementation.

1. **Multimodal AVER**

   Initially, two models (as detailed in Chapter 2, i.e. - VGGish (audio), Covariance + Temporal pooling (visual)) for feature extraction, were explored – one for each modality. Then attempted to reuse already existing and publicly available code ([17], [1]) and merge them in an end-to-end manner [59] in order to generate a fused output which is at the same time an input for a classifier Support Vector Machines (SVM). However, due to various reasons this intention resulted, rather, in an in-depth exploration of the unimodal architectures for feature extraction and being familiar with the libraries. One of the reasons was, for example, old versions of libraries versions (e.g. Tensor-Flow 1 vs current 2.2). The installation of an older version of TensorFlow was not found to be the solution and so the code was updated to a combination of both. However, this resulted in an issue when the Cream-D dataset library's class was intended to be used. Another example, lays in the difficulties of reusing already existing code due to some of the online resources lacking sufficient comments within the code or/and clear running instructions in README (e.g. Covariance and Temporal pooling code). Although a significant degree of progress was achieved, addressing these issues was time consuming and the level of complexity of the intended end-to-end architecture was underestimated.
   Therefore, a simpler Architecture based CNN was utilised for this project, adopted from Mansouri Benssassi's Doctoral dissertation [39] discussed in the Chapter 2. In her study, this architecture was used as one of the baseline models for a newly proposed multisensory architecture which was inspired by biological integration. The architecture was originally developed based on Snoek, Worring, and Smeulders study [57].

   - ***Implementation***
     Figure 3.1 highlights the structure of Mansouri Benssassi's multimodal architecture while figure 3.2 illustrates the same structure with further details. For the visual module it takes pre-processed features - collages from video frames - as an input, it has two convolution layers each with Rectified Linear Unit (ReLU) activation

function followed by a max pooling layer for further feature extraction. Further, after reshaping, two fully connected layers (fc1, 2) of linear data transformation are included again each being followed with ReLU.
In the case of the audio module, this takes .npy files which stores the necessary information including shape and dtype in order to reconstruct arrays containing audio MFCCs features correctly. Similarly as for the visual part, the audio one has also two convolution layers but only the first is followed by ReLU. Before passing further to concatenation for the late feature fusion, also this modality is reshaped.
After fusion and Batch normalisation of the two modalities for enhancing the speed of learning and the stability of the network, a 3 layer fully connected network - multi layer perceptron (MLP) is applied with ReLU. Finally, the output represents 8 values for all batches – although only the maximal value is considered as the decision of the network.

Undertaken changes:
The reported results in Mansouri Benssassi's dissertation [39] on the test split cited an accuracy of 81%. When original code was run the performance was as follows: train set 96%, 64% validation set, and 84% on test set. These results show a high overfitting of the model and for that reason an early stopping was included in the training process. This was done by checks of loss on validation set every 100 epochs. once the loss started to increase instead of decrease the training was stopped. However, early stopping occurred immediately after the first loop of one hundred epochs and the overfitting problems was still present. An exploration of parameters setting was undertaken, however, with no beneficial results. The resulting values of Momentum was changed from 0.9 to 0.5, original 24 batches, SGD optimiser but the number of epochs was at the end kept on 500 because of the lack of the early stopping effect on the test performance. After testing, due to the overfitting which can be a result of the complexity of the features (e.g. image is built from a number of pixels - each is a feature on its own) with respect to relatively small selected dataset (RAVDESS), the model is not able to learn the underlying interrelations connected to the output classes. To increase the number of samples, the validation set was included in the training to observe expected improvements.

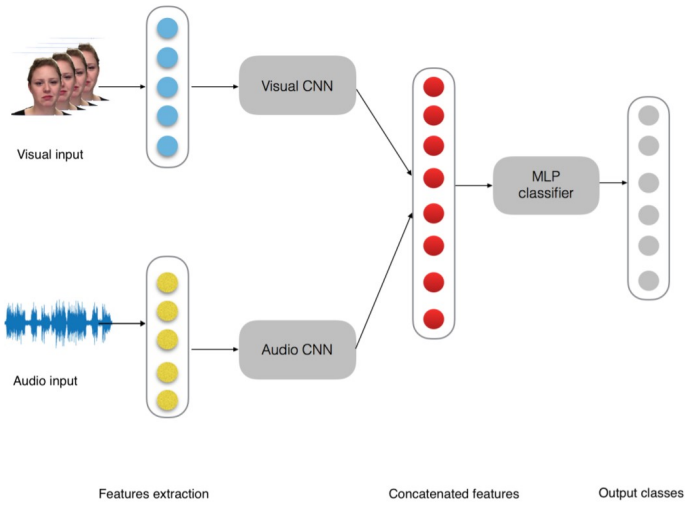

Figure 3.1: Multimodal - AVER model architecture [39].

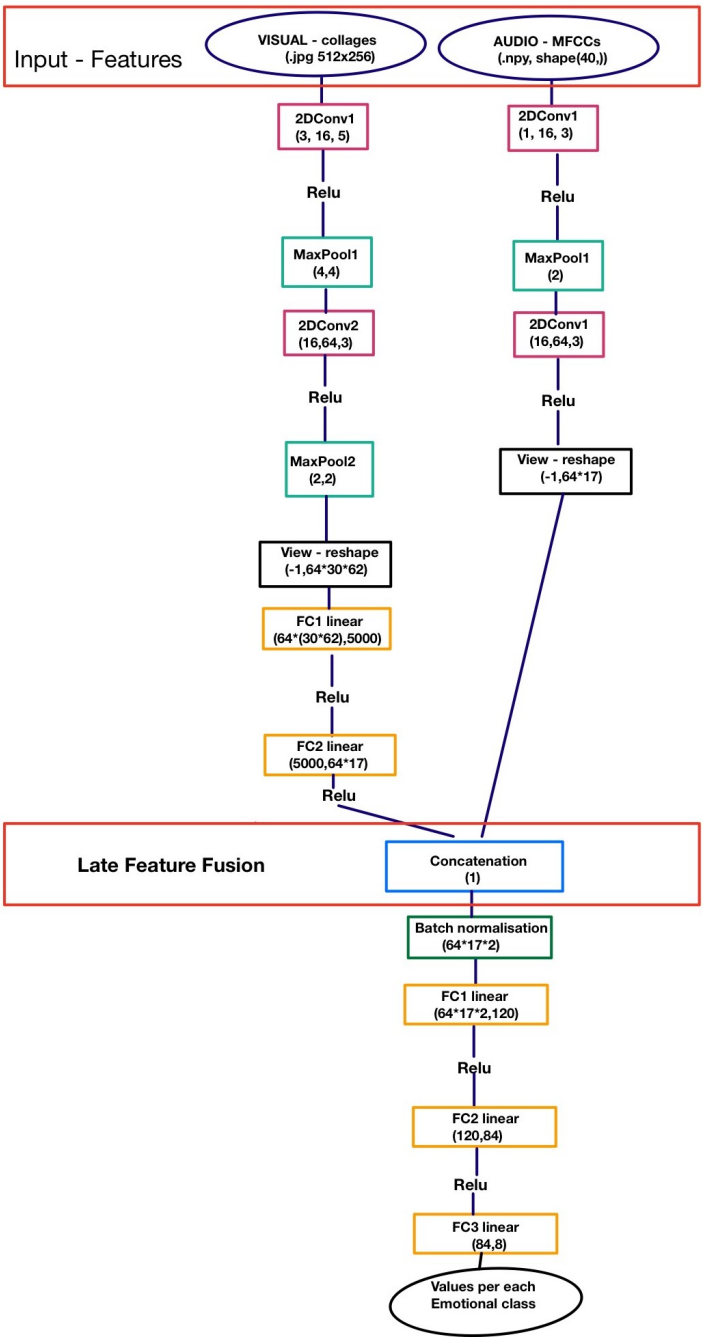

Figure 3.2: Multimodal - AVER model architecture with details.

2. **Unimodal FER**

The audio modality model was only extracted from the multimodal architecture in order to maintain consistency for the comparison of the

performance (see figure 3.3). After experimenting the selected values were for Momentum 0.9, learning rate 0.001 on 800 epochs.

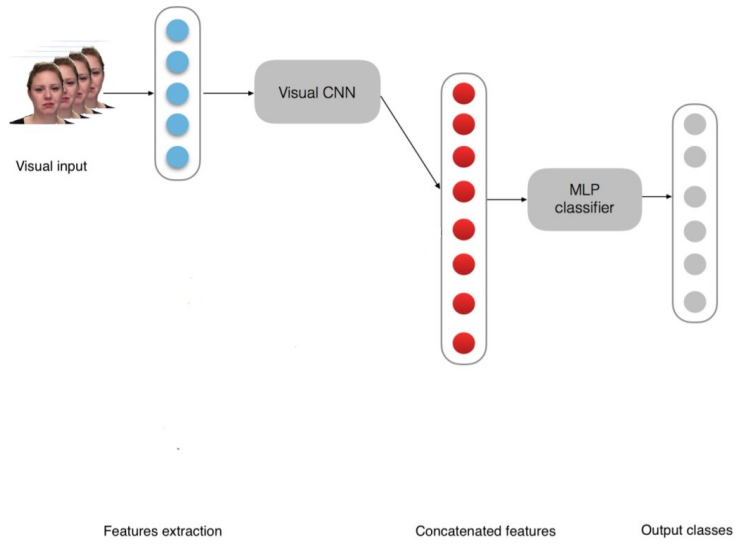

Figure 3.3: Unimodal visual FER model architecture.

3. Unimodal SER

Just in the case of the audio model the visual model was also extracted from the original multimodal architecture with momentum of 0.9 and learning rate 0.005 on 500 epochs (see figure 3.4).

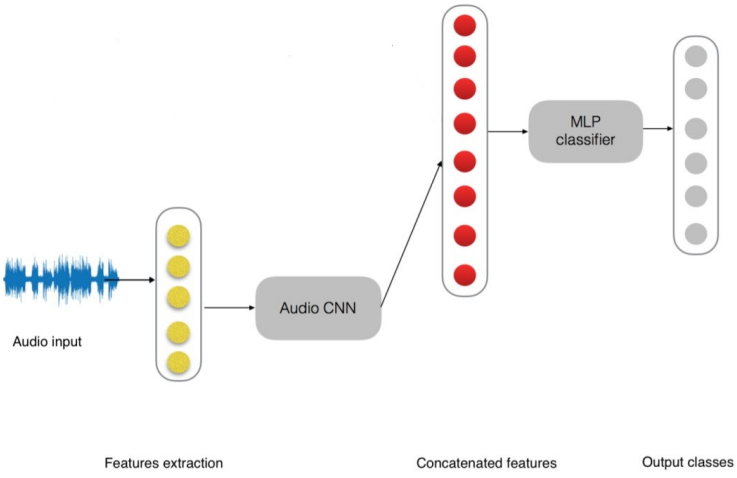

Figure 3.4: Unimodal audio SER model architecture.

3.4 Datasets

Two datasets were selected and explored with the aim that one would serve as a reserve if an anomaly behaviour would be detected in either of them. Both datasets contain audio, visual and audio-visual files of emotional content and both are accessible online free of charge for non-profitable use under a specific license. Moreover, the datasets are created from recordings of professional actors, with each sample with expressed emotion being validated by raters. This was done to decrease the emotional ambiguity and thus create more reliable datasets. Therefore, RAVDESS and CREMA-D [9], [10] the datasets are well suited to this study because trained models on a clean dataset would result in better observable effects of the noise augmentation on the performance.

RAVDESS

The database contains speech and singing from 24 actors (12 male, 12 female) with North American accents (see examples in figure 3.5). The content is built from two lexically-matched sentences which were expressed in different emotions in the recordings. Though, only speech files (1440 clips) with the eight emotional classes, namely, calm, happy, sad, angry, fearful, surprise, neutral and disgust were used for the study purposes. The dataset is balanced except neutral emotion class (see details in the figure 3.6). However, after features extraction takes place, the proportions are changed based on the length of the video (suggesting that emotions are expressed in different speed of talking) (see figure 3.7).

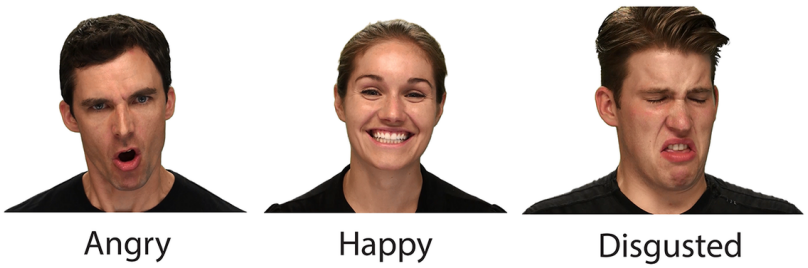

Figure 3.5: Examples from the dataset RAVDESS [38]

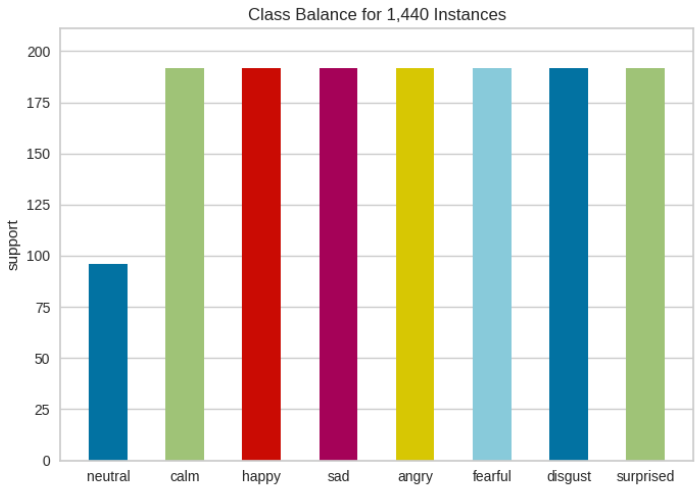

Figure 3.6: Class proportions in the dataset before FE.

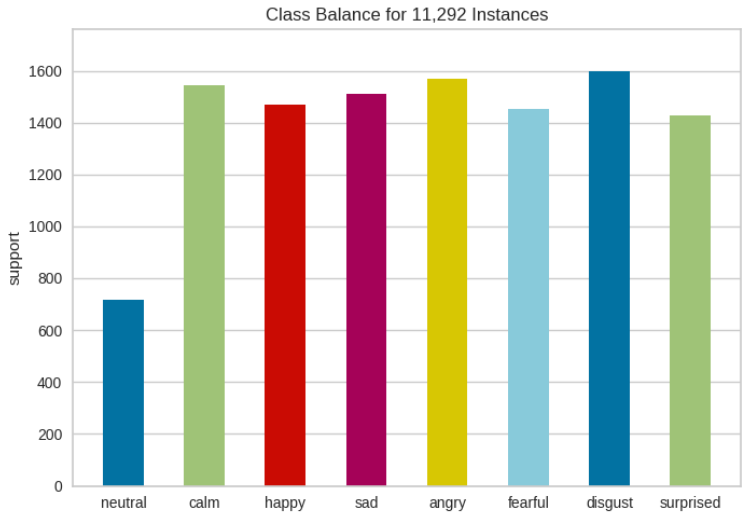

Figure 3.7: Class proportions in the dataset after FE.

CREMA-D

The database contains 7442 video clips from 91 actors (48 male, 43 female) from a variety of backgrounds (see examples in figure 3.8). In comparison to RAVDESS this dataset covers twelve sentences which also express various emotions but only with six different classes (Anger, Disgust, Fear, Happy,

Neutral, and Sad). Moreover, they included four emotion levels. Nonetheless, this classification is not relevant for this study.

Figure 3.8: Examples from the dataset CREMA-D [51]

3.5 Dataset pre-processing

- **Validation - Split**

  In the original code from Mansouri Benssassi [39] the code expects already prepared directories containing either just audio or video files. Hence, one extra step for dealing with the raw dataset was included. Then, the split to train, validation and test set is originally performed after feature extraction with each split having not only the possibility of containing videos from the same actor but also of contaning the features (visual collages, extended audio) of one video in each of the splits. Through the literature review this fact suggest a data leakage which was confirmed by testing a "leave-one-out" technique [51] where one actor was intentionally left out and the data were preprocessed the same way as all the splits and then tested on the trained model. In the literature survey various approaches are used such as overlaps of data from one subject in all splits [14], "leave-one-out" method [51], to splits based on subjects which prevent such overlaps that videos of one actor are in all splits [20], [58].

  It is sensible to split based on individuals because although there are characteristics of how emotion is expressed in all individuals, at the same time there are individual differences in how it is done [63]. In other words, the network should learn to generalise across the individuals in order to learn the connecting patterns rather than to learn the individuals' specifics.

  For a good practise this problem was addressed by splits (train, val, test) based on actors. This resulted in a new way of pre-processing and already existing code being altered or broken down and built based on the new needs.

- **Feature Extraction**

  FE was reused from the original code as it was for both modalities. Nevertheless, it was only broken down into smaller reusable functions for the data augmentation.

– **Video**

In the original code single frames were extracted from the raw video files by OpenCv library, resized and then from the resulting images collages containing two frames were created as the final features. Nevertheless, although the videos are created with a white background in the dataset RAVDESS it was found that the values of the pixeles are close to the white colour rather than actually white. This means that it adds extra irrelevant information to the network. Thus, after extracting the frames from the raw video, they were passed through a face detection (also OpenCv - Cascade Classifier), cropped to retain only the important facial information and reduce the irrelevant information. After that new resulting images were resized and merged into collages.

Figure 3.9: Visual features (collage) per one sample in this case 'neutral' class of the actor nr 8. [30].

– **Audio**

Audio features should represent extracted relevant components of the audio signal for defining specific emotions and reducing irrelevant information, thus decreasing feature complexity. Features are simply extracted from each loaded audio file by calling Librosa library method MFCCs. More specifically, (see figure 3.10), the hidden steps of the method include audio signal being split into overlapping frames on which Fast Fourier Transform is applied to obtain Mel-scale Spectograms, then logs of magnitude are calculated from the Fourier Transform. Finally, a last calculation of cosine transform is applied on the log power. This spectrum of the log represents the final output of 40 MFCCs over around 147 - 172 frames per sample (from seen examples through exploration) (see the visualised example from Mansouri Benssassi [39] in figure 3.11) [30], [44], [39]. The next step was to expand audio features to match the number of visual features which is necessary for the feature fusion. This was done per each class as follows: the total number of visual features in a class was divided by the total number of audio features in a class and with math.ceil function the smallest integral value was returned; this value represented the number of times the same audio needs to be duplicated to match the video features.

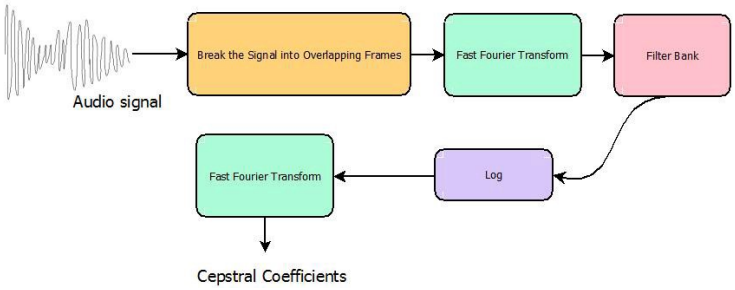

Figure 3.10: Steps to obtain from raw audio signal MFCCs [30].

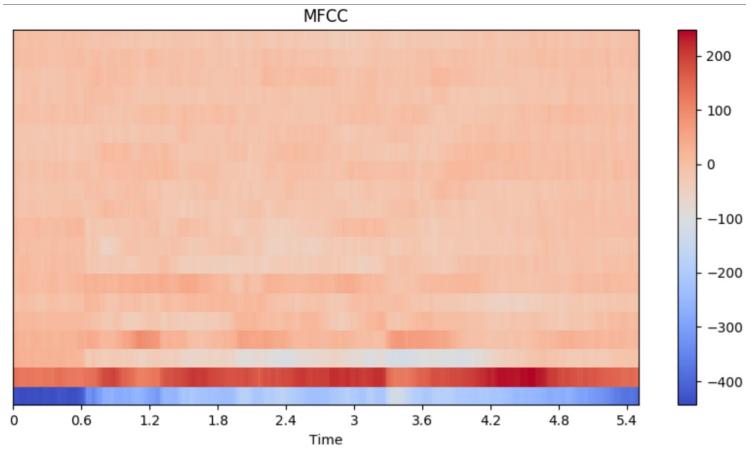

Figure 3.11: Example of MFCCs extracted from angry sample - used from [39]

3.6 Experimental setup

In general, noise augmentation is used for expanding the train set and achieve models which are able to generalise more effectively [24]. Nevertheless, the interest of this study lies in the comparison of the accuracy of the affected multisensory model with that of a unimodal model - especially to observe if there is a cooperation of the two modalities and if they can compensate for the loss of the other modality. Therefore, in order to explore aleitoric uncertainty effects caused by the noisy data on the multimodal architecture performance, a number of various noise augmentation on the test split were performed. In this section all details of the types of augmentations performed including visualised effects of each augmentation.

Note: images are loaded and saved through OpenCv library which uses BGR rather than RGB.

- **Visual Noise Augmentation**

  random_noise from *skimage* library was used to add the noise to the images in the first three types of augmentation and *torch* library was used for saving the new augmented images as tensors .

  – **Gaussian Noise**

    The variance of the random distribution was finalised at 0.9 (see figure 3.12). The higher the number the higher the level of noise is presented. For experimenting purposes a higher distortion of the images is desirable because then it is possible to observe any changes in the accuracy of the unimodal models.

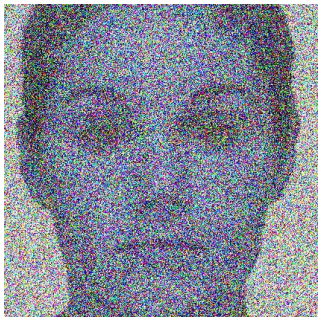

Figure 3.12: A resulted example of Gaussian noise augmentation

  – **Speckle Noise**

    Just as is the case of Gaussian noise, the variance of the random distribution parameter was set to 0.9 (see figure 3.13).

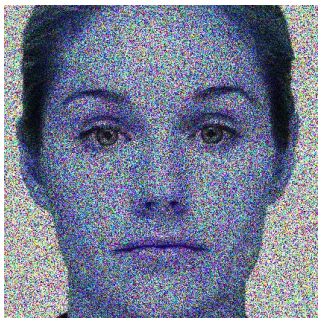

Figure 3.13: A resulted example of Speckle noise augmentation

  – **Salt & Pepper Noise**

    In the case of Salt and Pepper augmentation, the parameter was set to 0.99 because the noise was less significant in comparison to the two other augmentations (see figure 3.14).

Figure 3.14: A resulted example of Salt & Pepper noise augmentation

– **Non-cropped (original) images**

As previously discussed, the white background is considered as additional irrelevant information (noise), so for that reason, non cropped images were selected as the last type of a visual noise.

- **Audio Noise Augmentation**

librosa library was used to augment the original audio signal by stretching it or stretching the pitch in the signal and the *numpy* library was used for the remaining augmentation - White noise. All the newly resulting audio signals were saved and are here visualised with the spectrograms along with amplitude wave graphs

original sound graph:

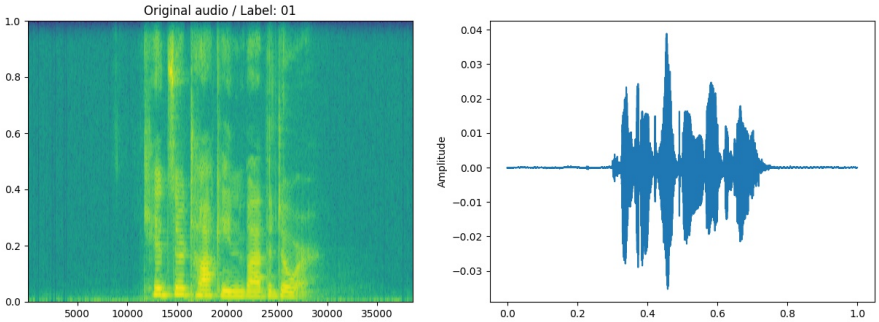

Figure 3.15: Representation of the orgininal sound before augmentation

– **White Noise**

A method random.rand from the *numpy* library was used to generate a random noise which was added to the original audio (see figure 3.15).

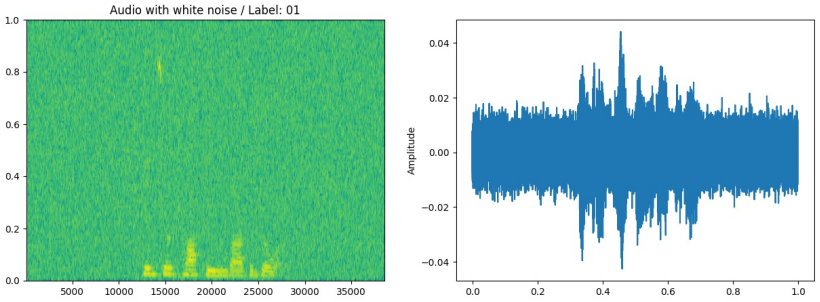

Figure 3.16: A resulted example of White noise augmentation

– **Stretch**

A method effects.stretch_time from the *librosa* library was used to stretch the original audio to be slower by 0.4 (see figure 3.16).

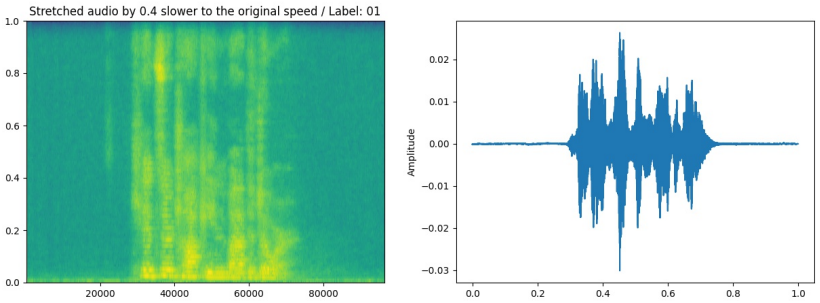

Figure 3.17: A resulted example of Stretch augmentation - 0.4 slower from the original speed

– **Stretch Pitch**

A method effects.pitch_shift from the *librosa* library was used with the number of 5 steps (see figure 3.17).

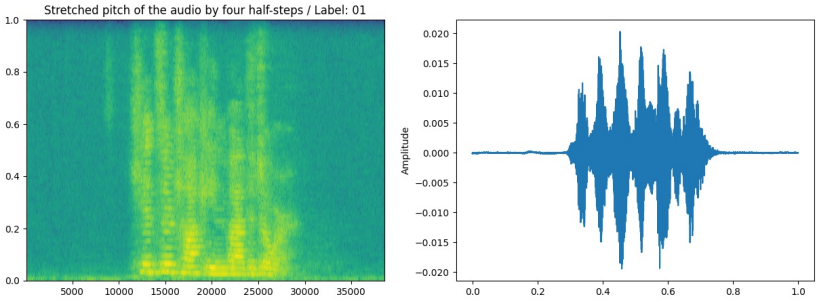

Figure 3.18: A resulted example of Stretch pitch augmentation being shifted with 5 fractional steps

3.7 Summary

This chapter as outlined detailed information regarding the research methodology. This includes the tools, the kinds of architectures and datasets used along with accompanying examples of visualisations and the rationale underpinning the motivation for why they were selected. Further, the details highlighted here relating to the experimental settings have covered the types of noise augmentation and their visualisations. Lastly, parts of code implementation have been provided with an emphasis on highlighting the changes made in comparison to the original code used from Mansouri Benssassi [39]

CHAPTER 4

The Fourth Chapter

4.1 Introduction

The focus of this project was on uncertainty in the input level and its effects on performance of a multimodal ER architecture. A number of experiments with different kinds of noise augmentation were conducted. Namely White noise, Stretch audio, and Stretch pitch in the audio test input and Gaussian, Salt & Pepper, and Speckle noise in the visual test input. Besides experiments focused on the effects of impaired input in one modality also combinations of augmentation in both modalities was explored. All experiments and results are accompanied with visualisations of different aspects of the experiments for deeper understanding. This chapter is divided into sections and subsections for clear overview of undertaken steps and experiments.

4.2 Baselines

This section summarises the performance of all three baselines (see table 4.1 with details of correct vs incorrect answers in figures 4.4 - 4.6 and per class in Appendix C): one multimodal AVER, two unimodal SER and FER. These results are important for further comparison with augmented noisy inputs from different modalities and their combinations. Despite the fact that the accuracy dropped rapidly (from originally in the paper [39] 81% to 58%) after preprocessing of the splits based on actors was performed, still, the pattern in the findings is in agreement with the literature [20]. Meaning, that multimodal architecture outperform both audio but also visual architecture.

In more detail, AVER was able to improve in recognition of the following emotions: calm, happy, sad, angry and surprised. However, in the case of fearful and disgust, the ability to correctly recognise the class declined. This means that when audio and visual modality was fused, besides overlapping information, they complement each other as mentioned in the literature review. On the other hand, it seems that there is a possibility that the multimodal model can inherit confusion from its modalities as it was in the case of issue in SER model with disgust class. This class was often miclassified as sad in SER but there was no significant problem in FER (for more details see figures 4.1 - 4.3). In another example, this kind of problem was not passed to the AVER, see the details about calm emotion which was problematic in FER but also SER, however, in AVER there was no that significant problem detected.

These overlaps of the classes should be detectable from t-sne plots in the Appendix D. This kind of visualisation demonstrates the class discriminability. In AVER in figure D.1 and FER in figure D.2 there is apparent overlap of the classes disgust and sad but this was not that clear in the case of SER which had this specific problem demonstrated in the confusion matrix. In the second mentioned example the t-sne visualisation corresponds to the confusion matrices.

Model type	Accuracy
AVER	58%
FER	45%
SER	36%

Table 4.1: Baselines of multimodal AVER and unimodal FER and SER architectures.

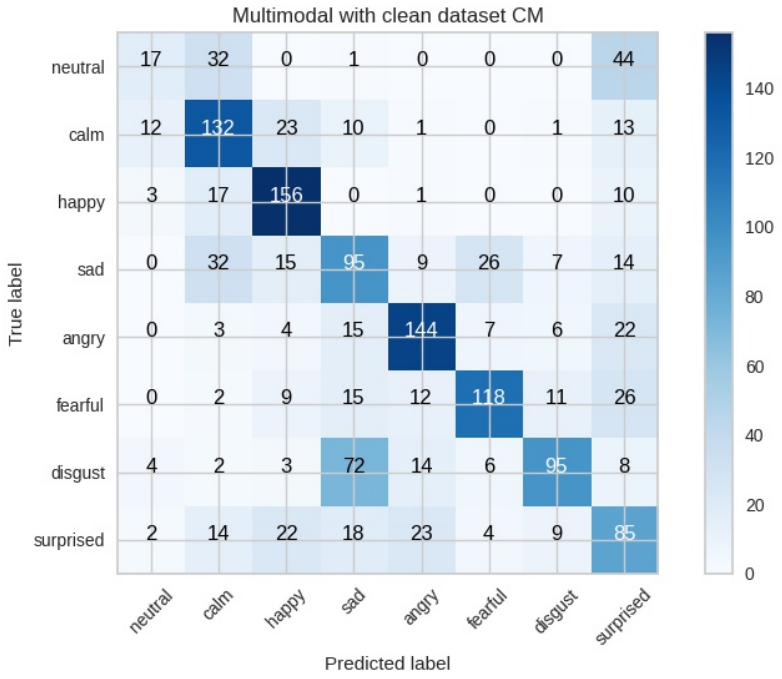

Figure 4.1: A Confusion Matrix for AVER baseline.

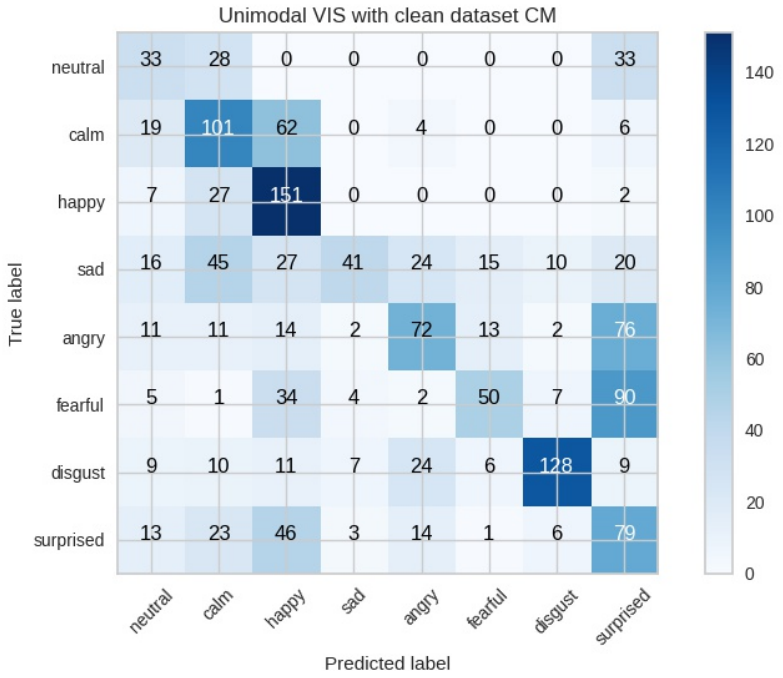

Figure 4.2: A Confusion Matrix for FER baseline.

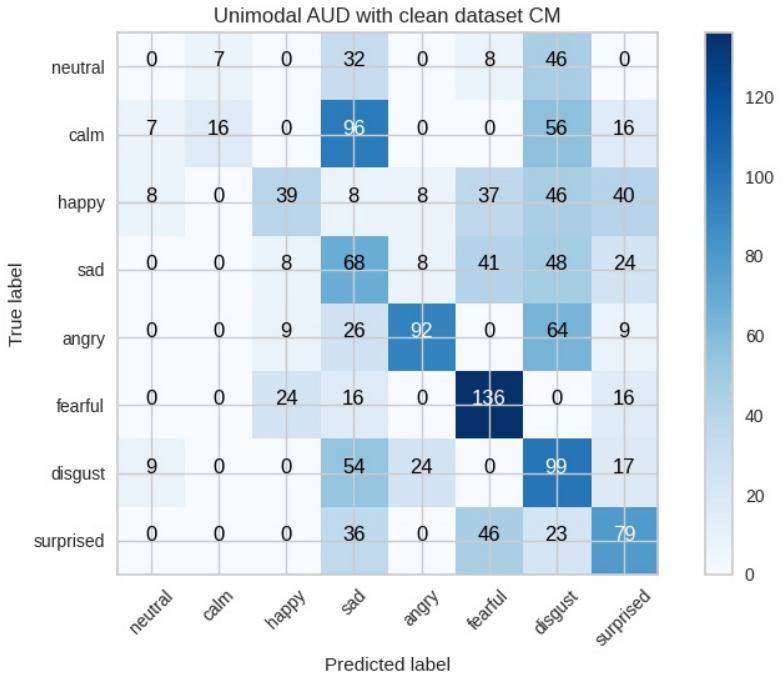

Figure 4.3: A Confusion Matrix for SER baseline.

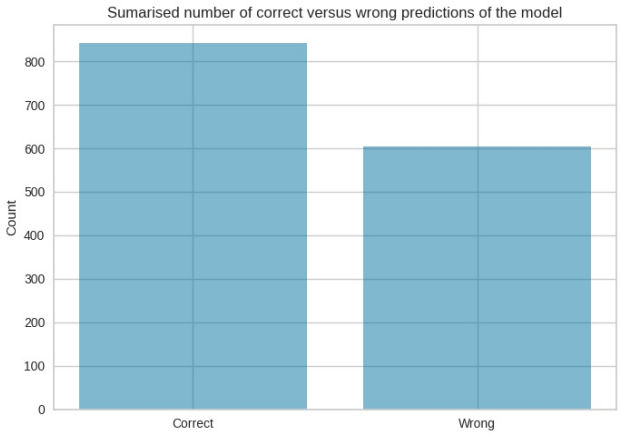

Figure 4.4: Total correct and incorrect predictions for AVER baseline.

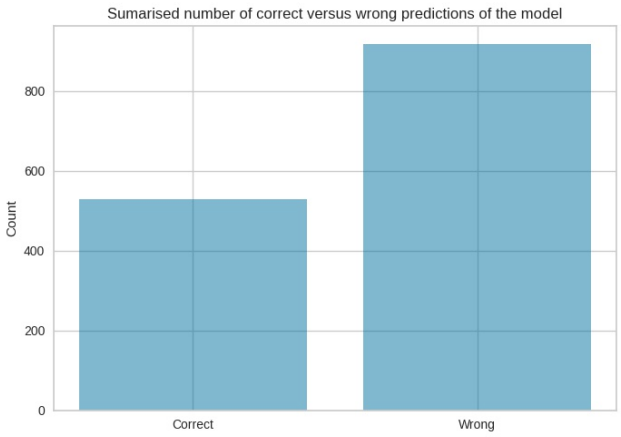

Figure 4.5: Total correct and incorrect predictions for SER baseline.

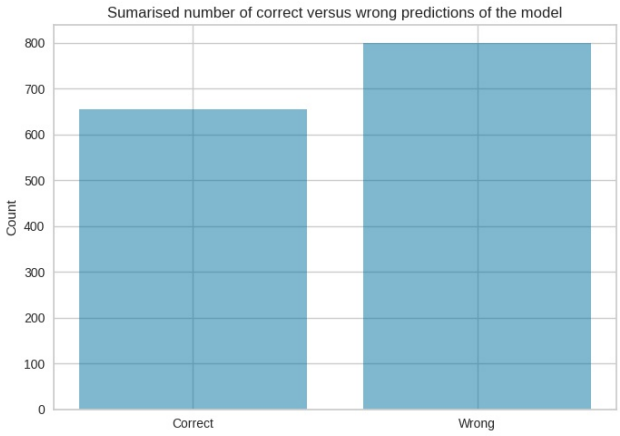

Figure 4.6: Total correct and incorrect predictions for FER baseline.

4.3 Visual Augmentations

This section overviews tables (4.2 and 4.3) with each performed augmentation and its scores is provided for both multimodal and unimodal architectures. However, in the Appendix A part A.1 specifically targeted tables are created for fast orientation in the achieved performance of each model. Then, confusion matrices are listed in Appendix B part B.1, bars of correct versus wrong answers per class in Appendix C part C.2, t-sne in Appendix D part D.2.

The most significant negative impact of the noise on AVER had usage of non-cropped frames. Raw frames contained a lot of extra noise and irrelevant information for the output. In this case clean audio modality in AVER have not decreased the negative impact, although the clean SER can achieve 36% on the accuracy in comparison to overall output of AVER with clean audio modality and augmented visual (non-cropped images) with only 19%. However, this is still more than achieved 14% with the same augmentation in FER. This suggests that some information was maintained from the clean modality and if extra information is passed to the model and importance of the modality would be modified then this loss of accuracy could be partly prevented.

The second most influential noise in the visual modality was caused by Gaussian noise with 31% in AVER. This is decrease of 27% in comparison to the clean dataset with 58%. However, in FER the drop of accuracy was not that severe from 45% to 35% which is 10% drop. It seems that influence of the noise also results not only in a confusion of the augmented modality but also of the clean one. Similarly, a decrease in accuracy occurred with the Speckle noise augmentation, but the negative impact was more apparent in AVER (from 58% to 41%) rather than in FER (from 45% to 42%).

On the other hand, the less influential noise augmentation in the visual modality was Salt & Pepper noise. This is due to the fact that the level of the noise was too low to have an impact on the models performance (possible to see in the example in Chapter 3 in the figure 3.14). Although, there was a insignificant decrease in AVER from 58% to 56% and no change in FER.

Model type	Augmentation type	Accuracy
FER	none	45%
FER	Visual - Salt and Pepper	45%
FER	Visual - Gaussian noise	35%
FER	Visual - Speckle	42%
FER	Visual - Non-cropped frames	14%

Table 4.2: Overview of all types of visual augmentation and its effects on FER

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Speckle	41%
AVER	Visual - Salt and Pepper	56%
AVER	Visual - Gaussian noise	31%
AVER	Visual - Non-cropped frames	19%

Table 4.3: Overview of all types of visual augmentation and its effects on AVER

4.4 Audio Augmentations

Similarly as in the visual augmentation section also here only the two main summary tables 4.5 and 4.5 are displayed. More details are located in appendixes such as tables designed for easier orientation in accuracy results in Appendix A part A.2, confusion matrices in Appendix B part B.2, bars of correct versus wrong answers per class in Appendix C part C.3, t-sne in Appendix D part D.3.

The most important finding in the audio augmentation is in the case of Stretch Pitch. Pitch appears to be notably linked to emotions because in SER performance dropped to 17%. Though in AVER the accuracy dropped, the decrease was only 10% (from 58%) in comparison to 19%. This could mean that visual modality has higher importance than the audio because the drop in the experiments in the visual modality in AVER were more or less similar to the results with the same augmentation in FER. Ambiguous results were obtained from White noise augmentation. In SER the results with augmented input actually increased accuracy of the model nonetheless, in the case of AVER the accuracy dropped by 8%. It is not clear why this is the case and these results may warrant further investigation of White noise augmentations and their levels.

Model type	Augmentation type	Accuracy
SER	none	36%
SER	Audio - White Noise	38%
SER	Audio - Stretch	34%
SER	Audio - Stretch Pitch	17%

Table 4.4: Overview of all types of audio augmentation and its effects on SER

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Audio - White Noise	50%
AVER	Audio - Stretch	58%
AVER	Audio - Stretch Pitch	48%

Table 4.5: Overview of all types of audio augmentation and its effects on AVER

4.5 Audio and Visual Augmentations combined

As in section 4.2 and 4.3, it also applies here that for further details and exhaustive visualisations the reader should explore appendixes. Namely, for preprepared structured tables see Appendix A part A.3, confusion matrices in Appendix B part B.3, bars of correct versus wrong answers per class in Appendix C part C.4, t-sne in Appendix D part D.4.

In some cases such as in the combined augmentation of audio as Stretch and visual as Salt & Pepper the accuracy remained close to the original. This is possible due to the fact that Salt & Pepper has no impact on FER and Stretch noise had minimal effect. A similar pattern was also found in the combination of visual as Salt & Pepper (no effect in FER) and Stretch pitch in audio (significant decrease of accuracy in SER but not so severe in AVER). However, foreseeable in previous experiments in previous sections the combination Gaussian noise (visual) and Stretch Pitch (audio) had one of the highest negative effects on the accuracy (resulted in 21%).

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Gaussian Noise and Audio - White Noise	31%
AVER	Visual - Gaussian Noise and Audio - Stretch	31%
AVER	Visual - Gaussian Noise and Audio - Stretch pitch	21%
AVER	Visual - Salt& Pepper and Audio - White Noise	47%
AVER	Visual - Salt& Pepper and Audio - stretch	56%
AVER	Visual - Salt& Pepper and Audio - Stretch Pitch	46%
AVER	Visual - Speckle and Audio - white Noise	42%
AVER	Visual - Speckle and Audio - Stretch	42%
AVER	Visual - Speckle and Audio - Stretch pitch	32%
AVER	Visual - non-cropped frames and Audio - White Noise	13%
AVER	Visual - non-cropped frames and Audio - Stretch	19%
AVER	Visual - non-cropped frames and Audio - Stretch pitch	15%

Table 4.6: Effects of Audio augmentation and visual augmentation - combined on AVER

4.6 Summary

This chapter provided an outline of the experiments conducted. Specifically the focus was first to highlight the baseline of all three architectures (AVER, SER and FER) and compare them. The multimodal model outperformed

both unimodal models. Then the focus was on augmentations of the test input data. Firstly, visual augmentations were covered such as Gaussian, Salt & Pepper, Speckle noise and non-cropped frames, followed by audio augmentations White, Stretch and Stretch pitch noise. Various visualisations accompanied each experiment including tables for a convenient overview of the accuracy of the models, as well as confusion matrices which provide detailed information about predicted versus true labels. Other provided visualisations are bars of correct versus incorrect predictions and lastly t-sne plots highlighting the distribution of classes of high-dimensional dataset and their overlaps in reduced dimensional space. In terms of the ability of the models to recognise the emotions, under noise augmentation some of the models were able to distinguish only a limited number of emotions and lost the ability in others.

CHAPTER 5

The Fifth Chapter

5.1 Limitations

This project covered a large number of experimental examples which resulted in problems concerning clarity in the structure. Furthermore, due to the time constraints not all examples in the appendices could be discussed and evaluated. Next, the architecture selected is a simple one and the results cannot be generalised to other architectures. This leads to another limitation that although most researched modalities are audio and video in ER, there are also others. Also, the differences in classes per each class in the augmented experiments should be explored in depth. Lastly, only certain kinds of noise were examined and this type of noise in the data does not necessarily represent the noise found in real life scenarios.

5.2 Future work

Derived from the limitations, the future research would benefit from examination of different architectures in ER, also various modalities used in ER (e.i. physiological or semantics in audio). Moreover, not only diverse kinds of noise but also various levels of the same noise may be beneficial to examine and explore the differences in the effect on AVER (similarly as it is done in [39]). In addition, results in one audio augmentation urge for an exploration of the importance of different modalities. Finally, different uncertainties and their interaction how they may influence each other may be important for further progress because as it was suggested before different types of uncertainties are interlinked or inseparable.

5.3 Conclusion

It is impossible to deny the significance of uncertainty in ML and DL in general. Noise in the input data in visual modalities resulted, in some cases, in a rapid drop of accuracy of AVER. However, based on the literature multi-modal architectures have the potential in recovering performance losses from impaired modalities by using the information from other modalities with clean input. This was the case in the experiment with impaired audio modality where visual modality was able to compensate the loss of information from audio. It seems that visual modality has higher importance not only thanks to achieved overall

accuracy of the model but also demonstrated in the example. A suggested upgrade to the current models is to focus on including information about the issue and increase the importance of the clean modality to access the potential in the clean modality.

Appendices

APPENDIX A

The First Appendix - TABLES experiments accuracy overview

A.1 First Section - tables representing the effects of different types of noise in Visual modality

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Gaussian noise	31%
FER	none	45%
FER	Visual - Gaussian noise	35%
SER	none	36%

Table A.1: Effects of Visual augmentation - Gaussian Noise.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Salt & Pepper	56%
FER	none	45%
FER	Visual - Salt & Pepper	45%
SER	none	36%

Table A.2: Effects of Visual augmentation - Salt & Pepper.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Speckle	41%
FER	none	45%
FER	Visual - Speckle	42%
SER	none	36%

Table A.3: Effects of Visual augmentation - Speckle Noise.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Non-cropped frames	19%
FER	none	45%
FER	Visual - Non-cropped frames	14%
SER	none	36%

Table A.4: Effects of Visual - Non-cropped frames.

A.2 Second Section - tables representing the effects of different types of noise in Audio modality

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Audio - White Noise	50%
FER	none	45%
SER	none	36%
SER	Audio - White Noise	38%

Table A.5: Effects of Audio augmentation - White Noise.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Audio - Stretch	58%
FER	none	45%
SER	none	36%
SER	Audio - Stretch	34%

Table A.6: Effects of Audio augmentation - Stretch.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Audio - Stretch Pitch	48%
FER	none	45%
SER	none	36%
SER	Audio - Stretch Pitch	17%

Table A.7: Effects of Audio augmentation - Stretch Pitch.

A.3 Third Section - tables representing the effects of different combinations of types of noise in both modalities

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Gaussian Noise and Audio - White Noise	31%
FER	none	45%
FER	Visual - Gaussian Noise	35%
SER	none	36%
SER	Audio - White Noise	38%

Table A.8: Effects of Audio augmentation - White Noise and Visual Augmentation - Gaussian Noise.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Gaussian Noise and Audio - Stretch	31%
FER	none	45%
FER	Visual - Gaussian Noise	35%
SER	none	36%
SER	Audio - Stretch	34%

Table A.9: Effects of Audio augmentation - Stretch and Visual Augmentation - Gaussian Noise.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Gaussian Noise and Audio - Stretch pitch	21%
FER	none	45%
FER	Visual - Gaussian Noise	35%
SER	none	36%
SER	Audio - Stretch pitch	17%

Table A.10: Effects of Audio augmentation - Stretch Pitch and Visual Augmentation - Gaussian Noise.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Salt& Pepper and Audio - White Noise	47%
FER	none	45%
FER	Visual - Salt& Pepper	45%
SER	none	36%
SER	Audio - White Noise	38%

Table A.11: Effects of Audio augmentation - White Noise and Visual Augmentation - Salt& Pepper.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Salt& Pepper and Audio - stretch	56%
FER	none	45%
FER	Visual - Salt& Pepper	45%
SER	none	36%
SER	Audio - Stretch	34%

Table A.12: Effects of Audio augmentation - Stretch and Visual Augmentation - Salt& Pepper.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Salt& Pepper and Audio - Stretch Pitch	46%
FER	none	45%
FER	Visual - Salt& Pepper	45%
SER	none	36%
SER	Audio - Stretch pitch	17%

Table A.13: Effects of Audio augmentation - Stretch Pitch and Visual Augmentation - Salt& Pepper.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Speckle and Audio - white Noise	42%
FER	none	45%
FER	Visual - Speckle	42%
SER	none	36%
SER	Audio - White Noise	38%

Table A.14: Effects of Audio augmentation - White Noise and Visual Augmentation - Speckle.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Speckle and Audio - Stretch	42%
FER	none	45%
FER	Visual - Speckle	42%
SER	none	36%
SER	Audio - Stretch	34%

Table A.15: Effects of Audio augmentation - Stretch and Visual Augmentation - Speckle.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - Speckle and Audio - Stretch pitch	32%
FER	none	45%
FER	Visual - Speckle	42%
SER	none	36%
SER	Audio - Stretch pitch	17%

Table A.16: Effects of Audio augmentation - Stretch Pitch and Visual Augmentation - Speckle.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - non-cropped frames and Audio - White Noise	13%
FER	none	45%
FER	Visual - non-cropped frames	14%
SER	none	36%
SER	Audio - White Noise	38%

Table A.17: Effects of Audio augmentation - White Noise and Visual - non-cropped frames.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - non-cropped frames and Audio - Stretch	19%
FER	none	45%
FER	Visual - non-cropped frames	14%
SER	none	36%
SER	Audio - Stretch	34%

Table A.18: Effects of Audio augmentation - Stretch and Visual Augmentation - non-cropped frames.

Model type	Augmentation type	Accuracy
AVER	none	58%
AVER	Visual - non-cropped frames and Audio - Stretch pitch	15%
FER	none	45%
FER	Visual - non-cropped frames	14%
SER	none	36%
SER	Audio - Stretch pitch	17%

Table A.19: Effects of Audio augmentation - Stretch Pitch and Visual Augmentation - non-cropped frames.

APPENDIX B

The Second Appendix - experiments Confusion Matrices overview

B.1 First Section - Visual augmentation

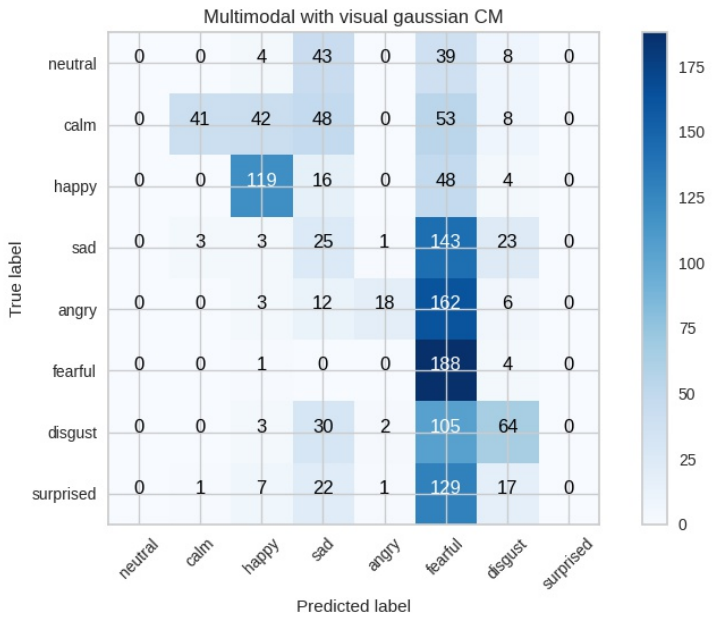

Figure B.1: A Confusion Matrix for AVER with visual augmentation - Gaussian Noise.

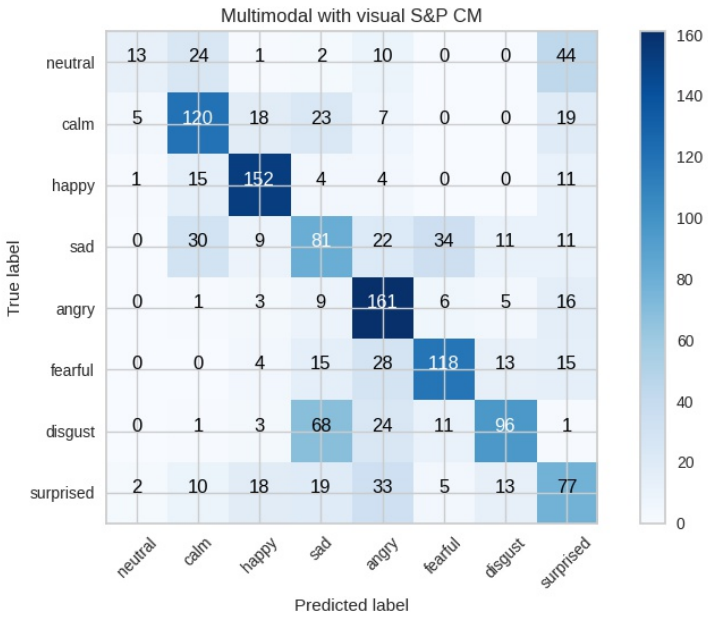

Figure B.2: A Confusion Matrix for AVER with visual augmentation - Salt & Pepper Noise.

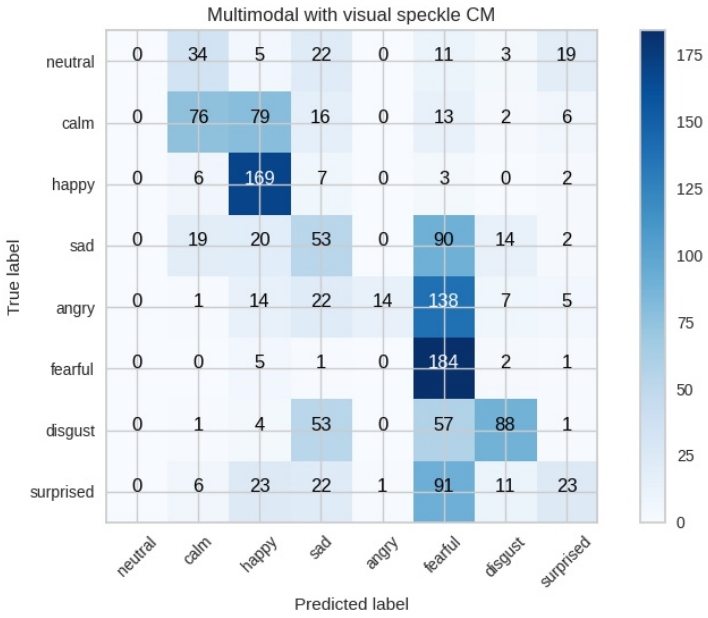

Figure B.3: A Confusion Matrix for AVER with visual augmentation - Speckle Noise.

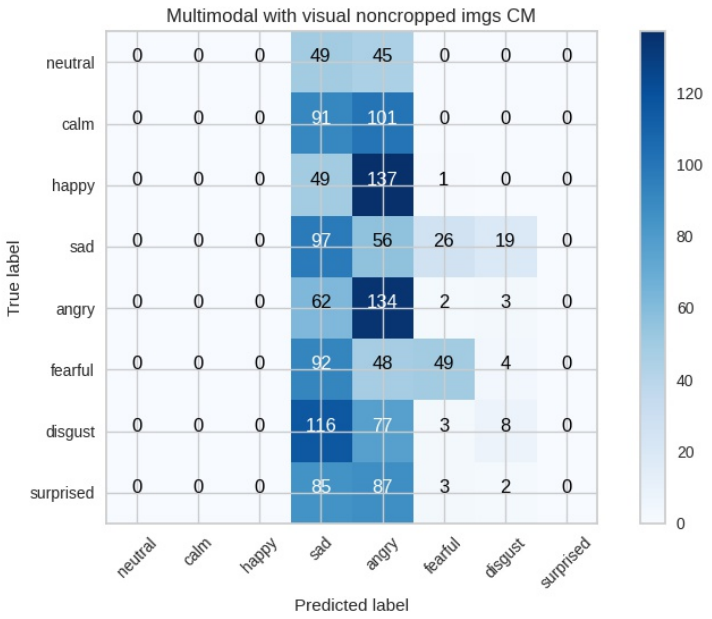

Figure B.4: A Confusion Matrix for AVER with visual augmentation - Non-cropped frames.

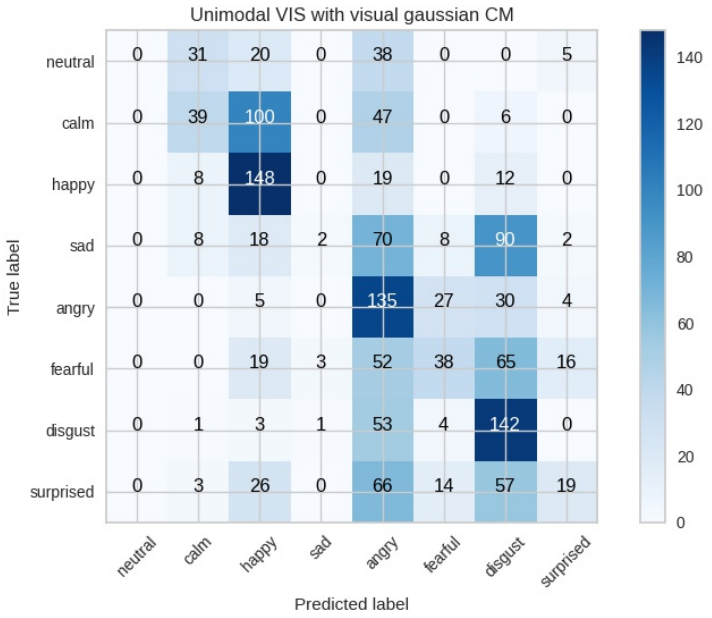

Figure B.5: A Confusion Matrix for FER with visual augmentation - Gaussian noise.

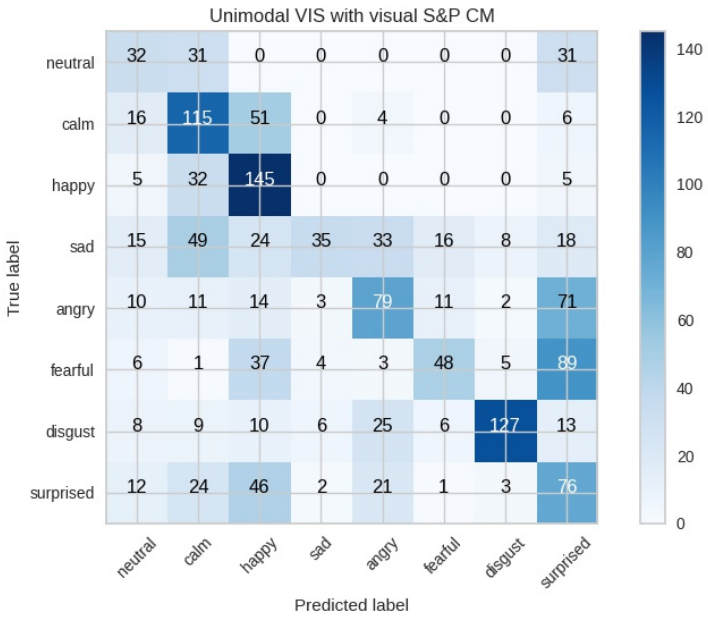

Figure B.6: A Confusion Matrix for FER with visual augmentation - Salt & Pepper Noise.

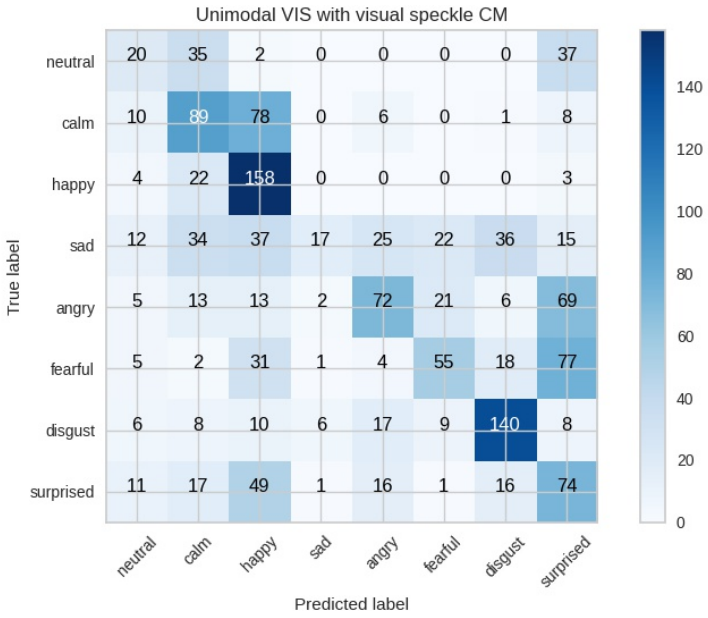

Figure B.7: A Confusion Matrix for FER with visual augmentation - Speckle Noise.

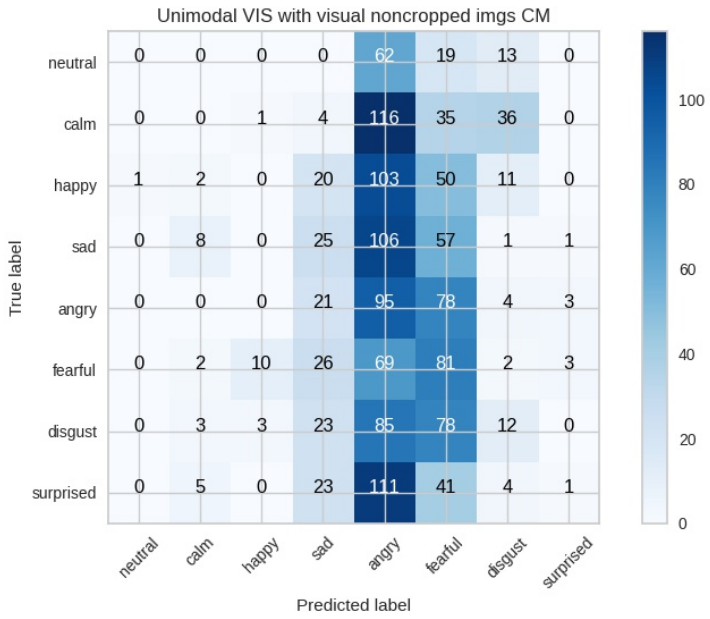

Figure B.8: A Confusion Matrix for FER with visual augmentation - Non-cropped frames.

B.2 Second Section - Audio augmentation

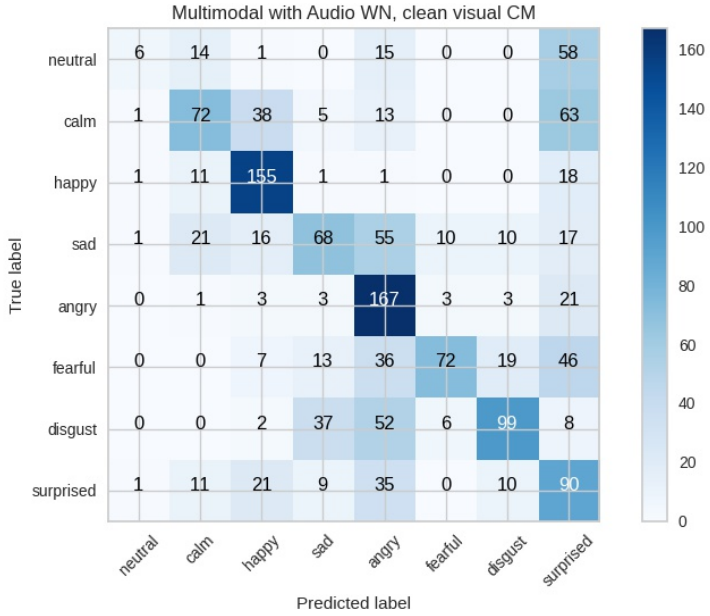

Figure B.9: A Confusion Matrix for AVER with audio augmentation - White Noise.

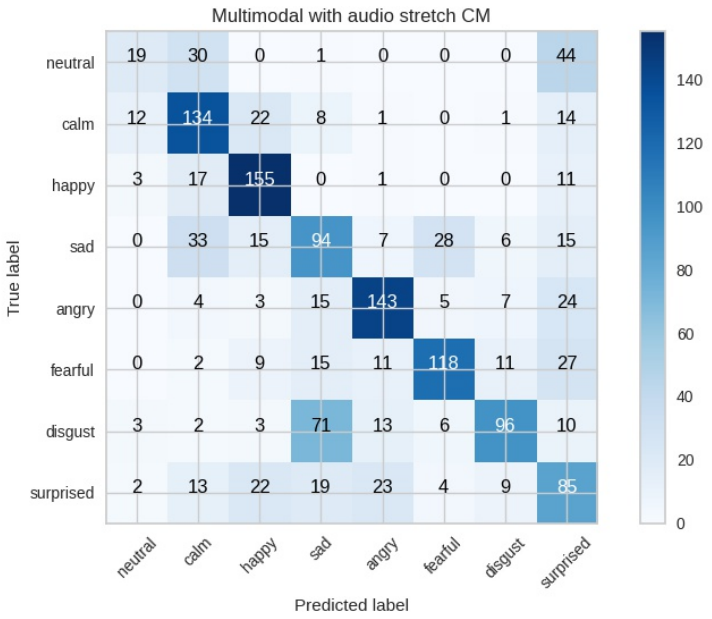

Figure B.10: A Confusion Matrix for AVER with audio augmentation - Stretch.

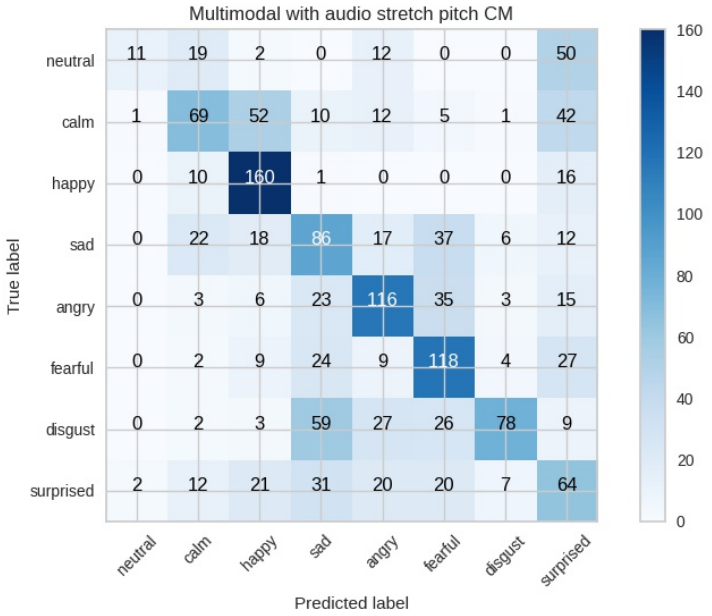

Figure B.11: A Confusion Matrix for AVER with audio augmentation - Stretch pitch.

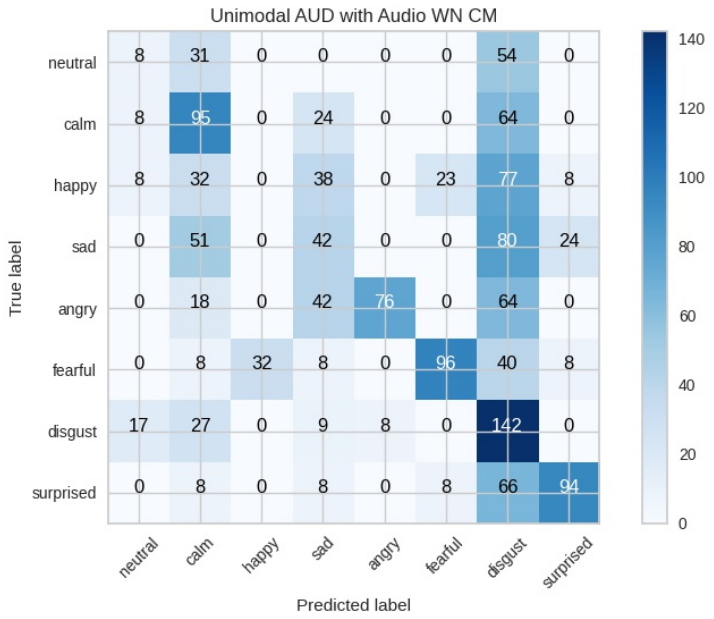

Figure B.12: A Confusion Matrix for SER with audio augmentation - White Noise.

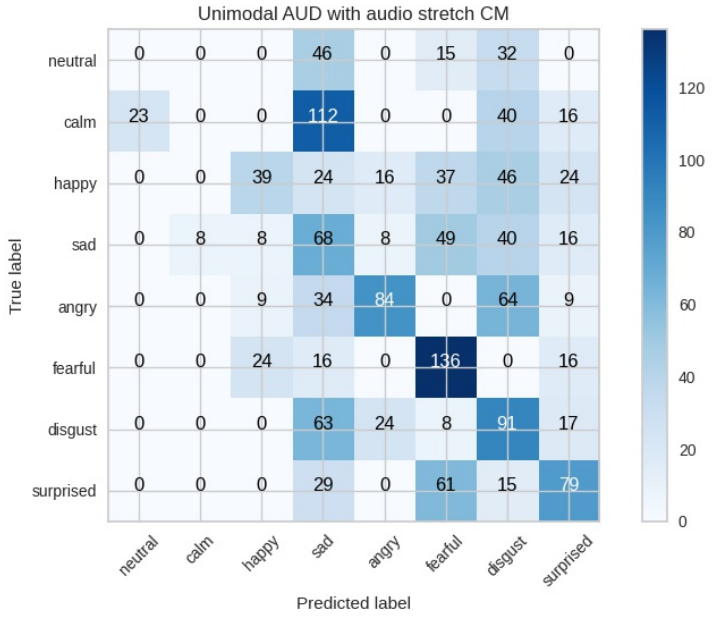

Figure B.13: A Confusion Matrix for SER with audio augmentation - Stretch.

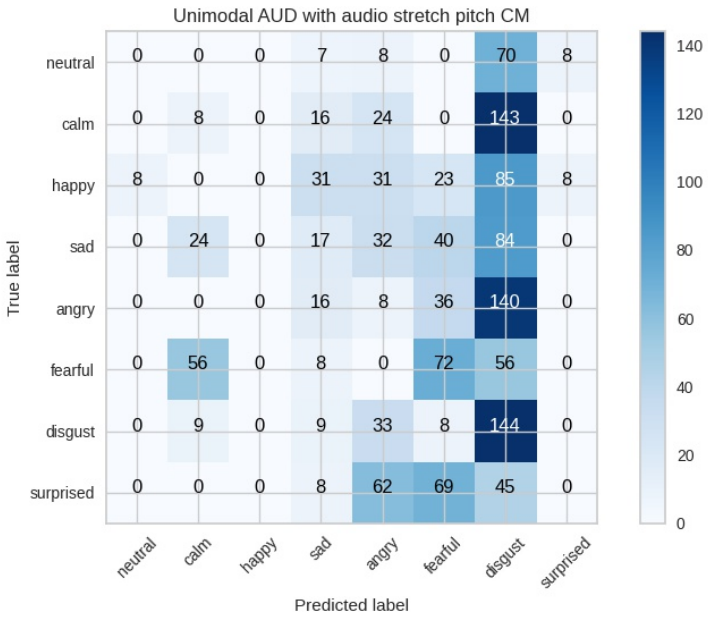

Figure B.14: A Confusion Matrix for SER with audio augmentation - Stretch pitch.

B.3 Third Section - Audio and Visual augmentation

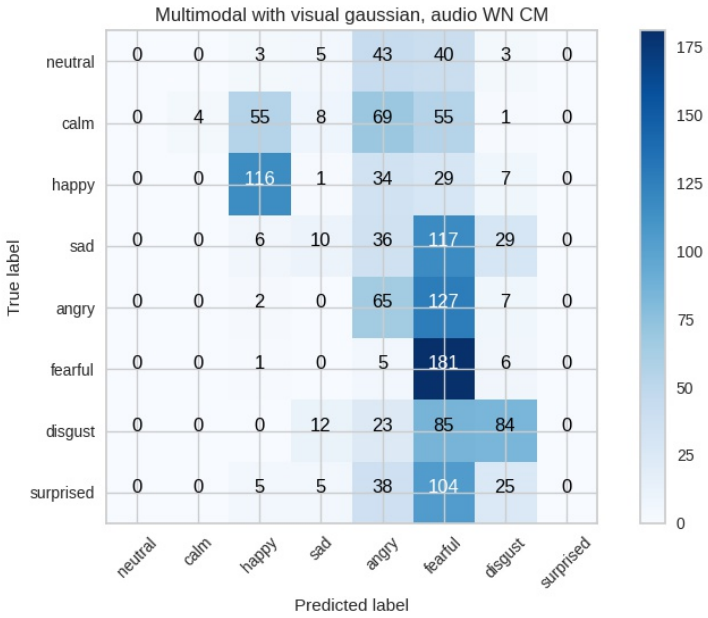

Figure B.15: A Confusion Matrix for AVER with augmentations visual - Gaussian Noise and audio - White Noise.

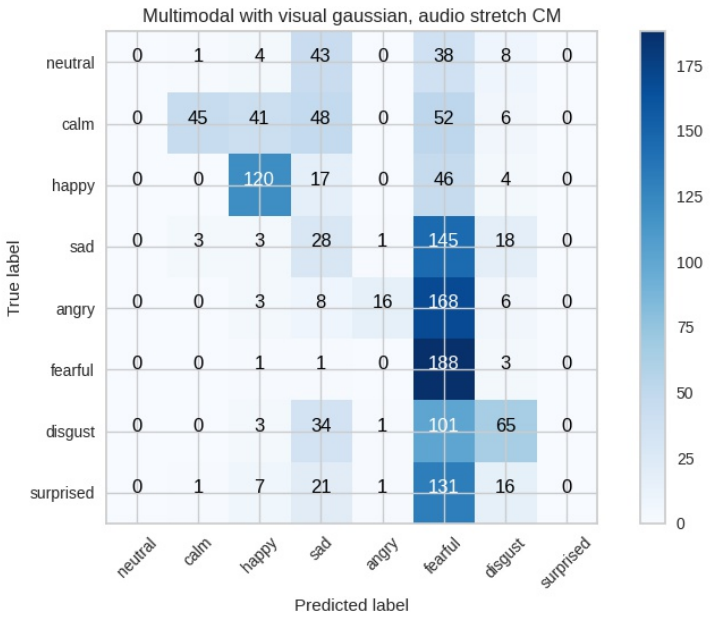

Figure B.16: A Confusion Matrix for AVER with augmentations visual - Gaussian Noise and audio - Stretch.

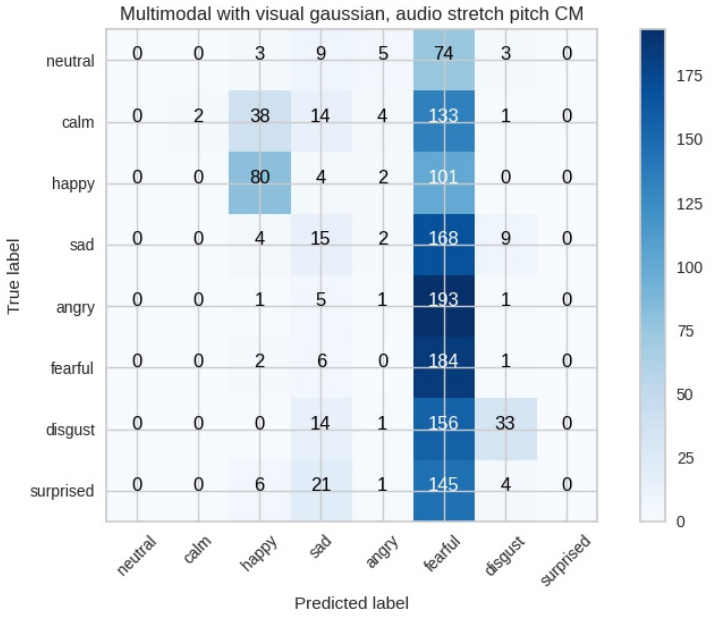

Figure B.17: A Confusion Matrix for AVER with augmentations visual - Gaussian Noise and audio - Stretch Pitch.

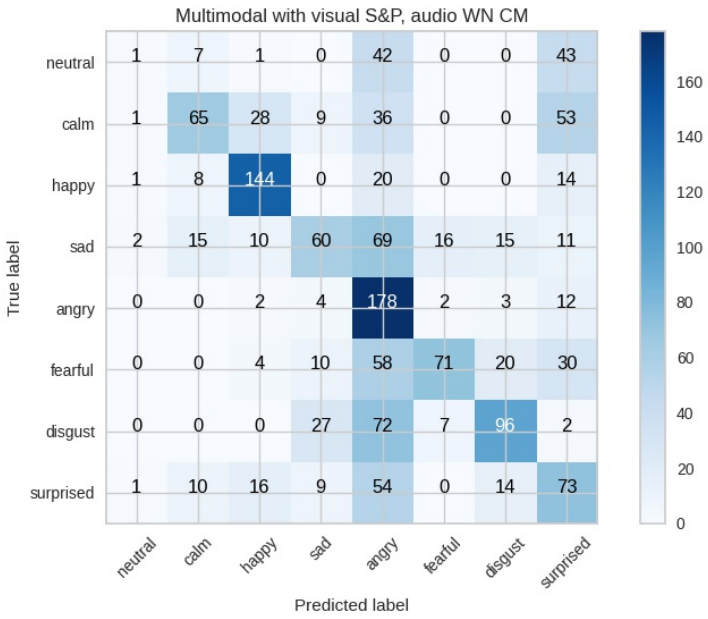

Figure B.18: A Confusion Matrix for AVER with augmentations visual - Salt & Pepper Noise and audio - White Noise.

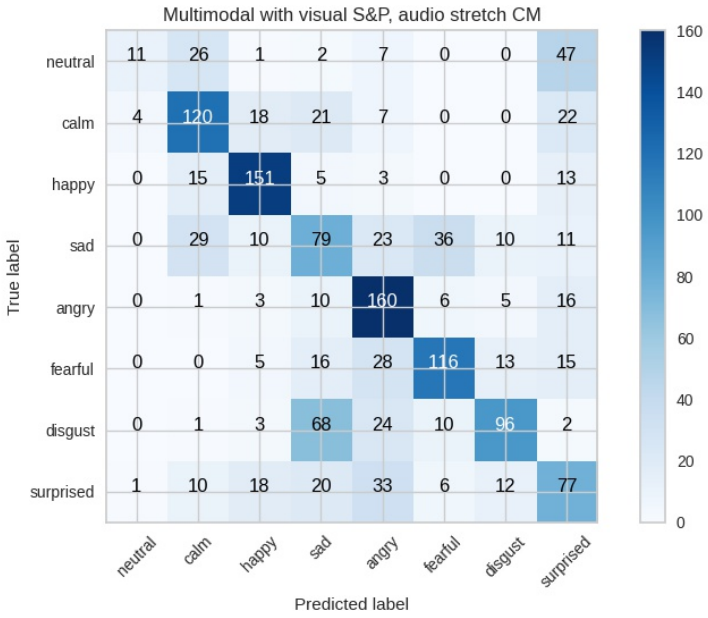

Figure B.19: A Confusion Matrix for AVER with augmentations visual - Salt & Pepper Noise and audio - Stretch.

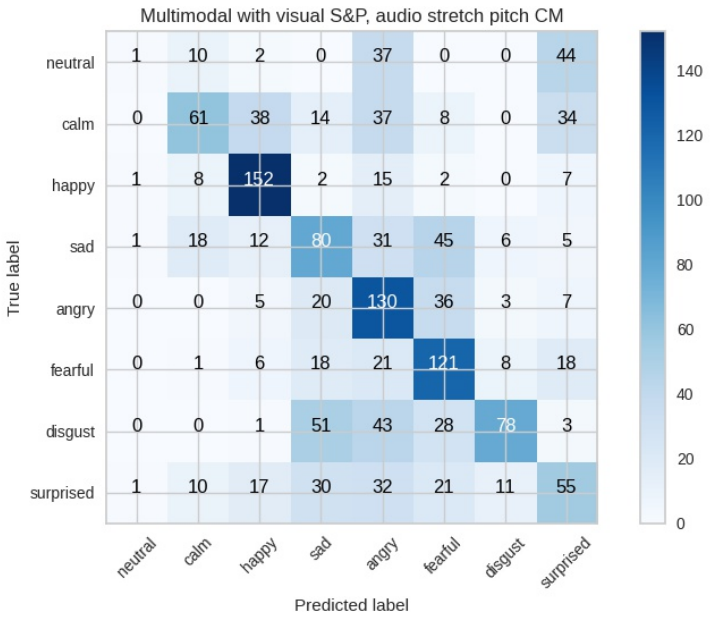

Figure B.20: A Confusion Matrix for AVER with augmentations visual - Salt & Pepper Noise and audio - Stretch Pitch.

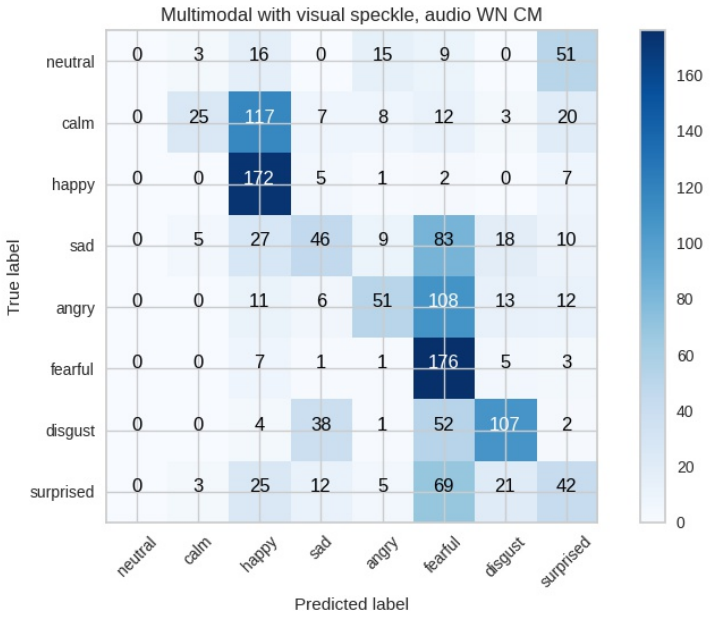

Figure B.21: A Confusion Matrix for AVER with augmentations visual - Speckle Noise and audio - White Noise.

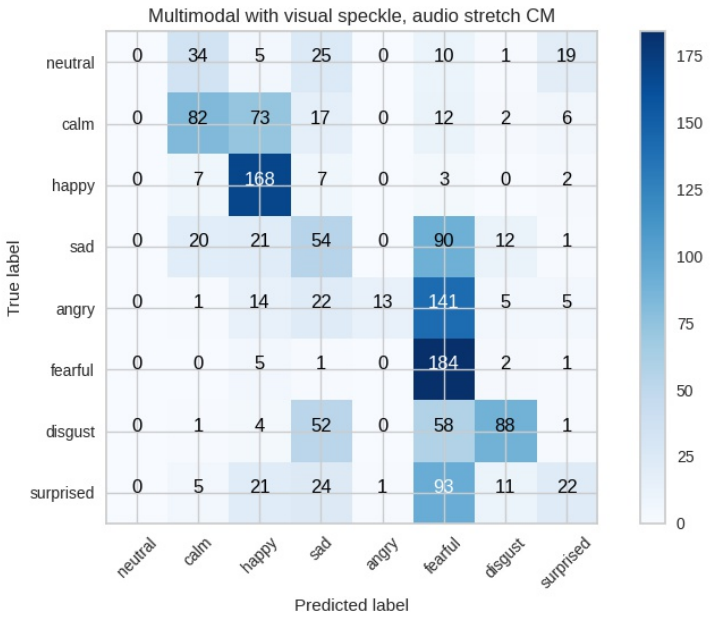

Figure B.22: A Confusion Matrix for AVER with augmentations visual - Speckle Noise and audio - Stretch.

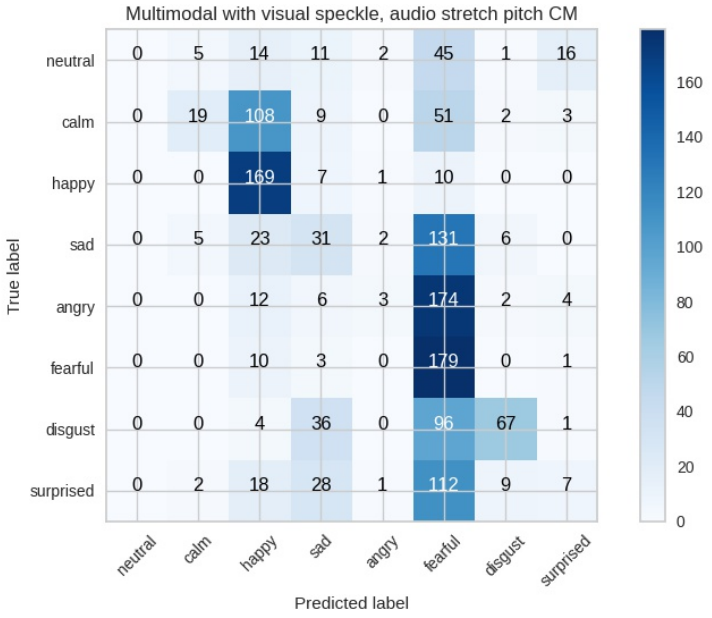

Figure B.23: A Confusion Matrix for AVER with augmentations visual - Speckle Noise and audio - Stretch Pitch.

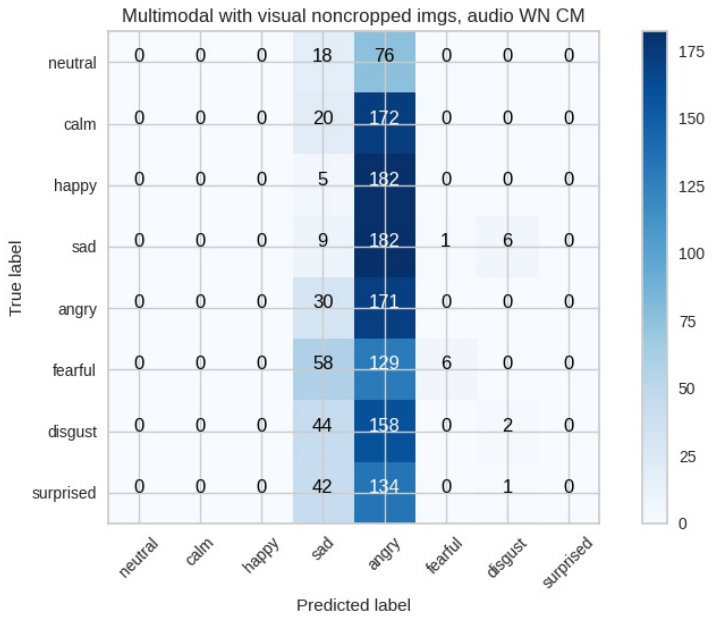

Figure B.24: A Confusion Matrix for AVER with augmentations visual - non-cropped frames and audio - White Noise.

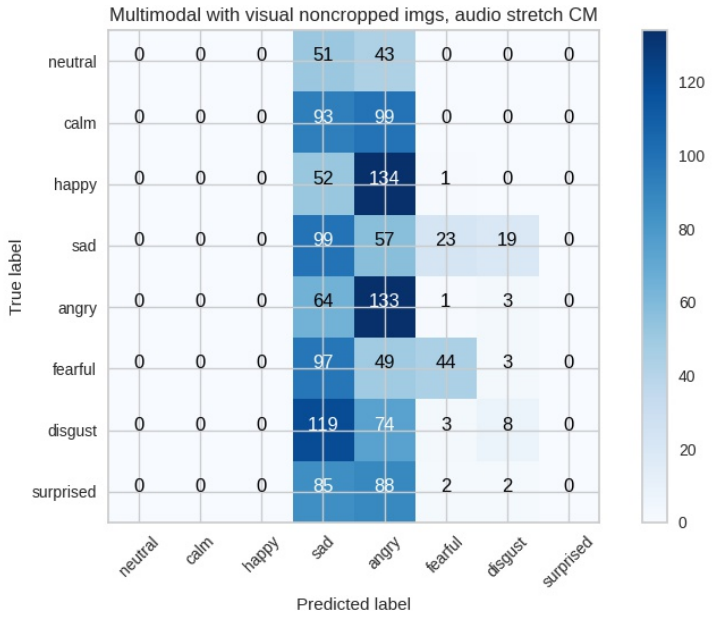

Figure B.25: A Confusion Matrix for AVER with augmentations visual - non-cropped frames and audio - Stretch.

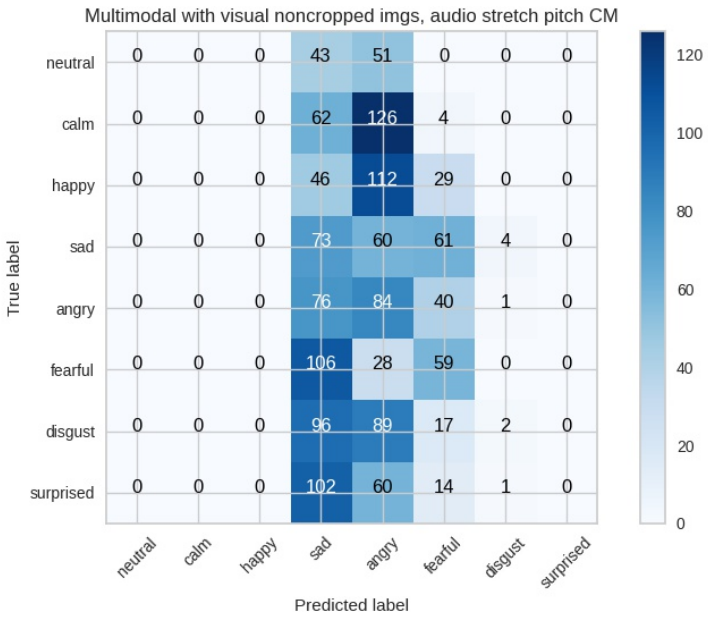

Figure B.26: A Confusion Matrix for AVER with augmentations visual - non-cropped frames and audio - Stretch Pitch.

APPENDIX C

The Third Appendix - experiments Bars correct vs incorrect answers

C.1 First Section - baselines

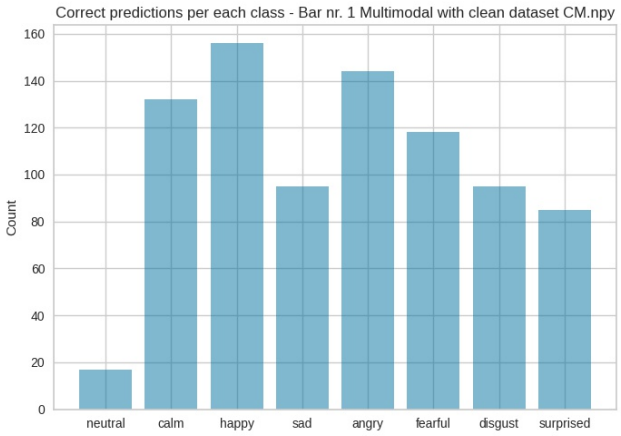

Figure C.1: An overview of correct predictions for AVER baseline.

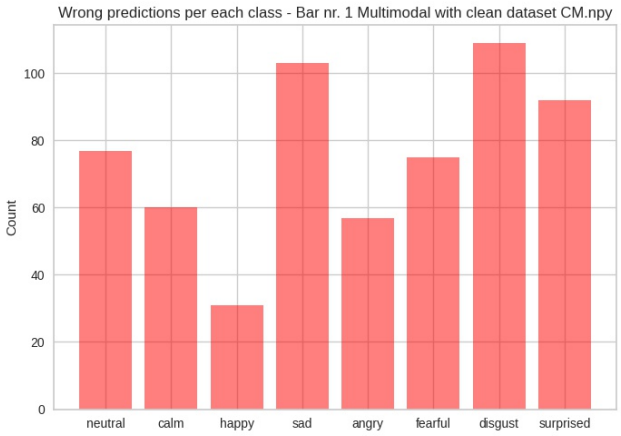

Figure C.2: An overview of incorrect predictions for AVER baseline.

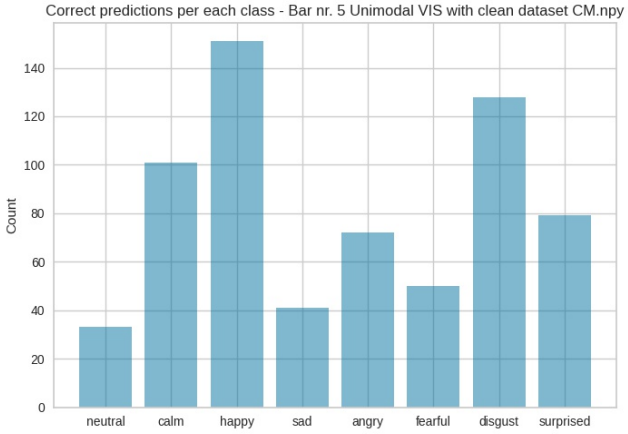

Figure C.3: An overview of correct predictions for FER baseline.

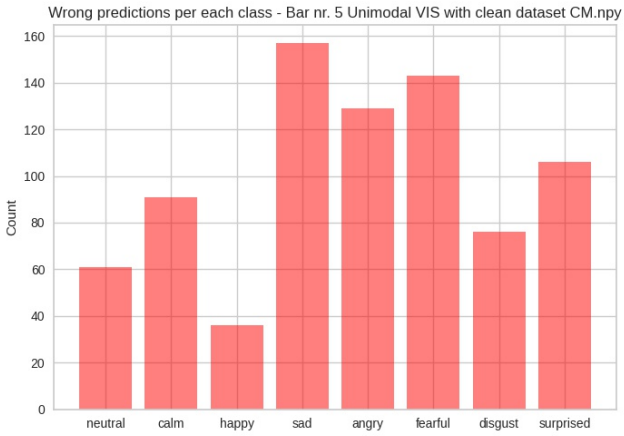

Figure C.4: An overview of incorrect predictions for FER baseline.

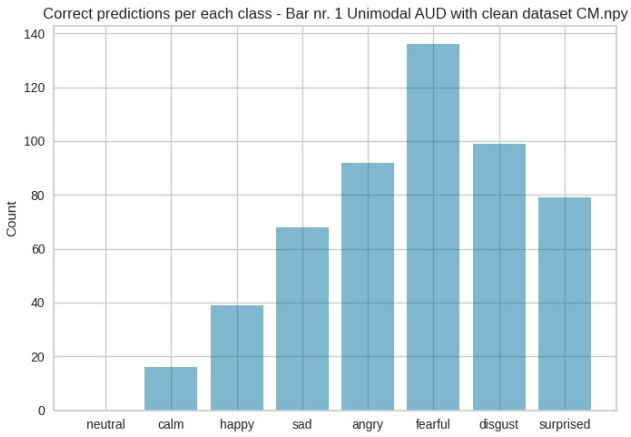

Figure C.5: An overview of correct predictions for SER baseline.

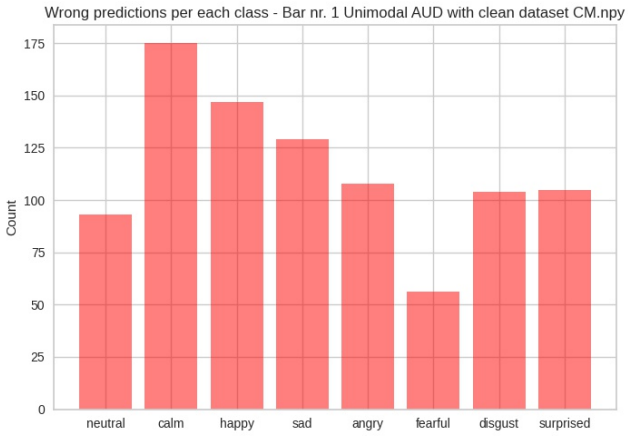

Figure C.6: An overview of incorrect predictions for SER baseline.

C.2 Second Section - Visual augmentation

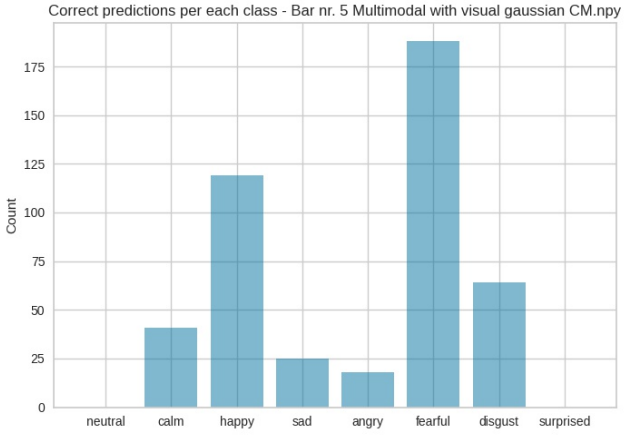

Figure C.7: An overview of correct predictions for AVER with visual augmentation - Gaussian Noise.

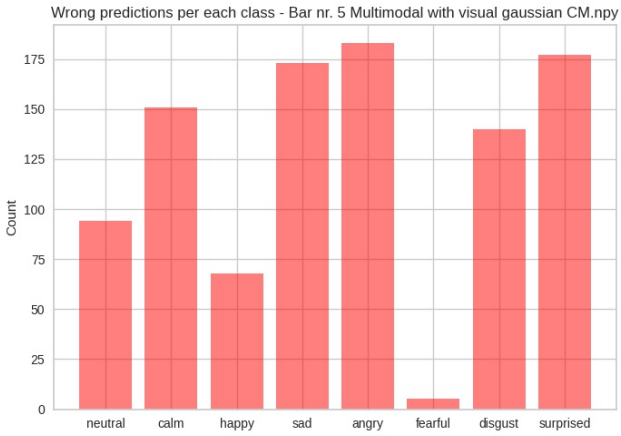

Figure C.8: An overview of incorrect predictions for AVER with visual augmentation - Gaussian Noise.

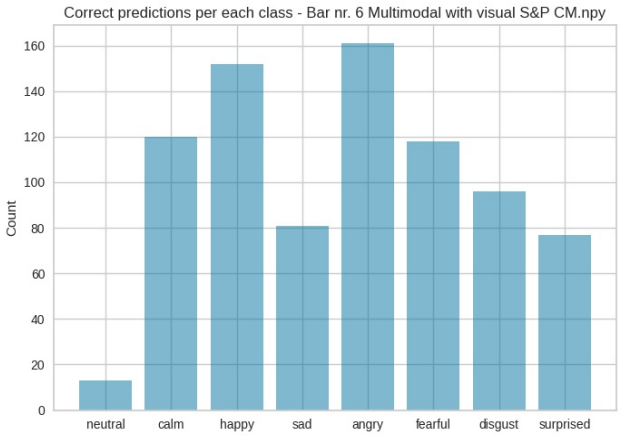

Figure C.9: An overview of correct predictions for AVER with visual augmentation - Salt & Pepper Noise.

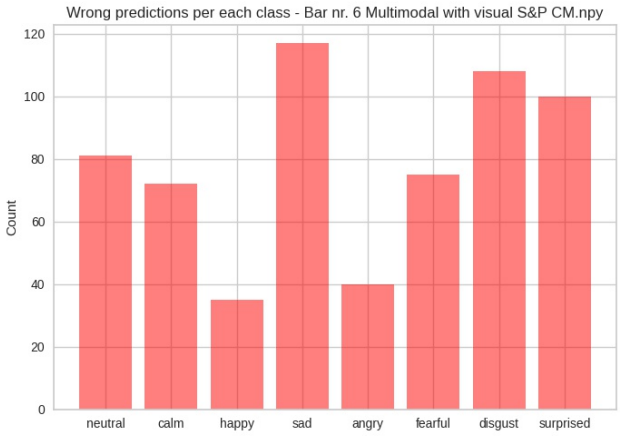

Figure C.10: An overview of incorrect predictions for AVER with visual augmentation - Salt & Pepper Noise.

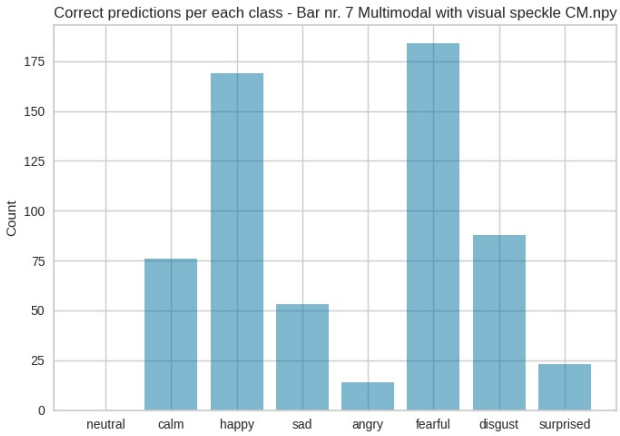

Figure C.11: An overview of correct predictions for AVER with visual augmentation - Speckle Noise.

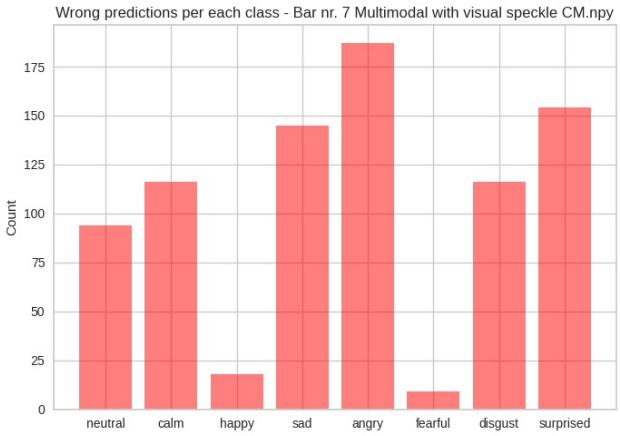

Figure C.12: An overview of incorrect predictions for AVER with visual augmentation - Speckle Noise.

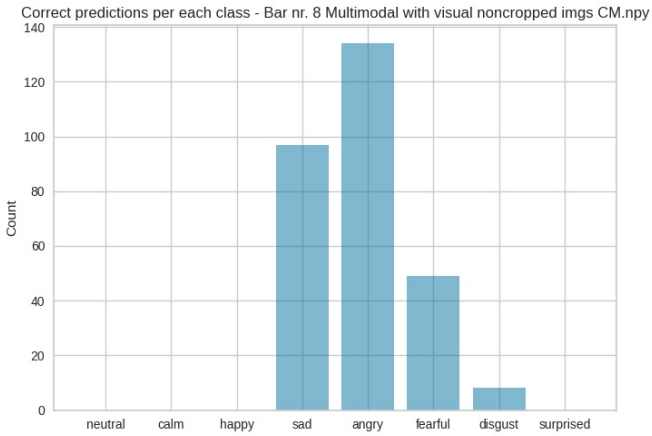

Figure C.13: An overview of correct predictions for AVER with visual augmentation - Non-cropped frames.

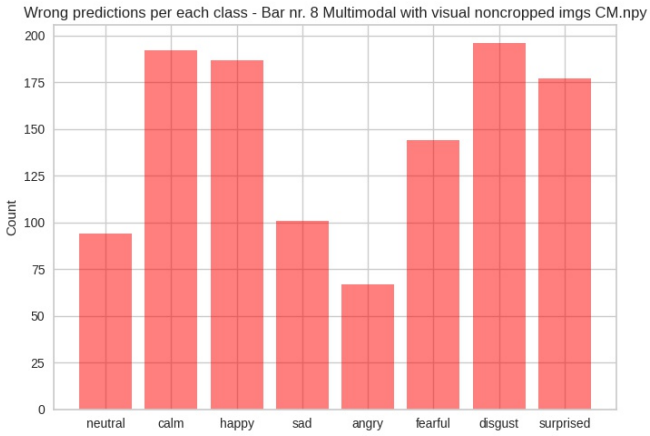

Figure C.14: An overview of incorrect predictions for AVER with visual augmentation - Non-cropped frames.

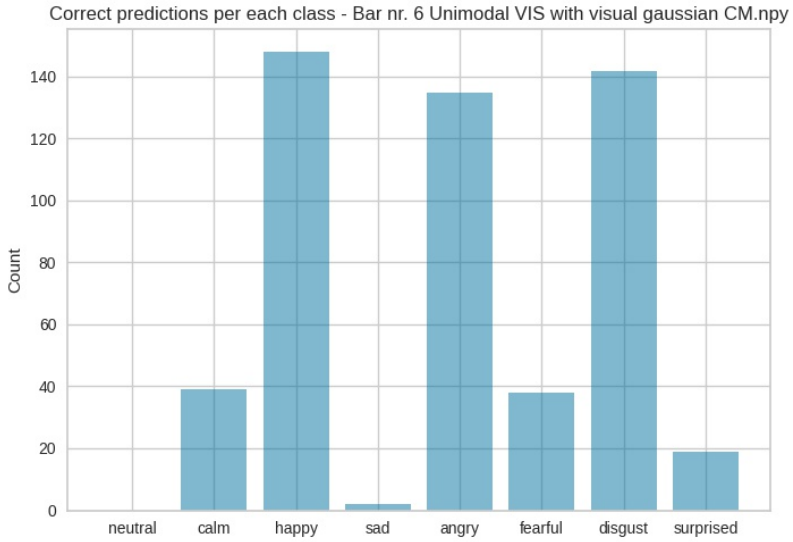

Figure C.15: An overview of correct predictions for FER with visual augmentation - Gaussian noise.

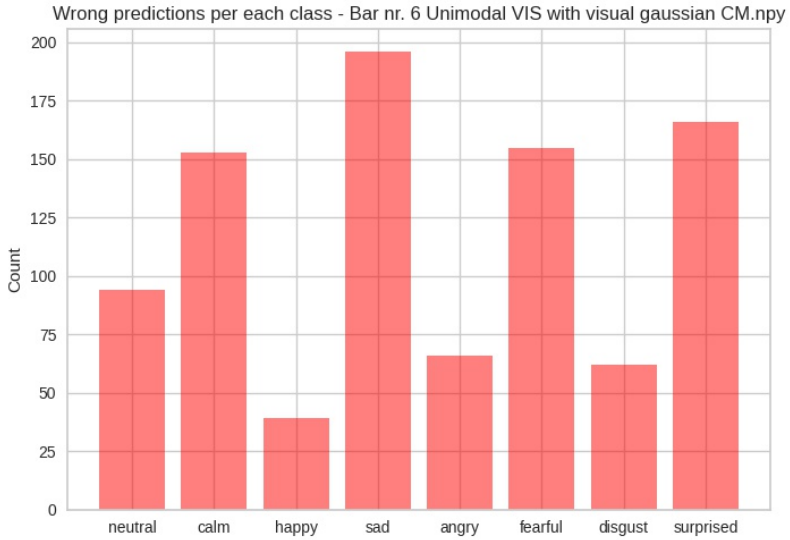

Figure C.16: An overview of incorrect predictions for FER with visual augmentation - Gaussian noise.

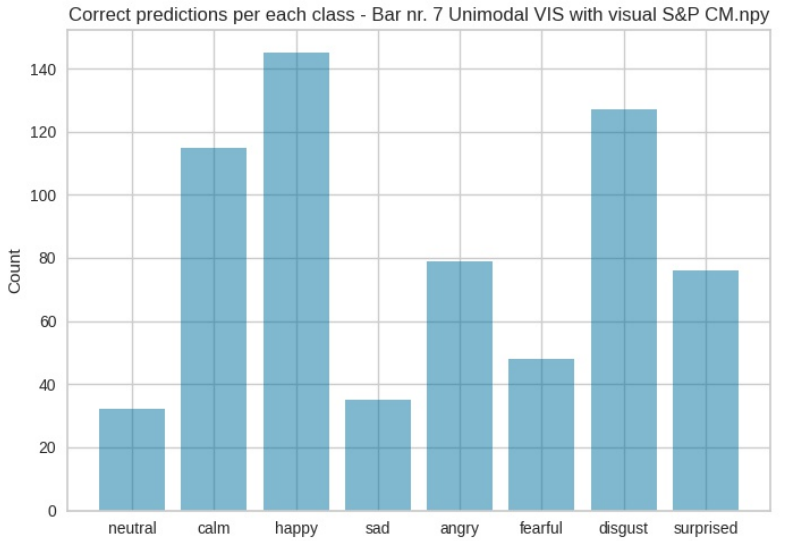

Figure C.17: An overview of correct predictions for FER with visual augmentation - Salt & Pepper Noise.

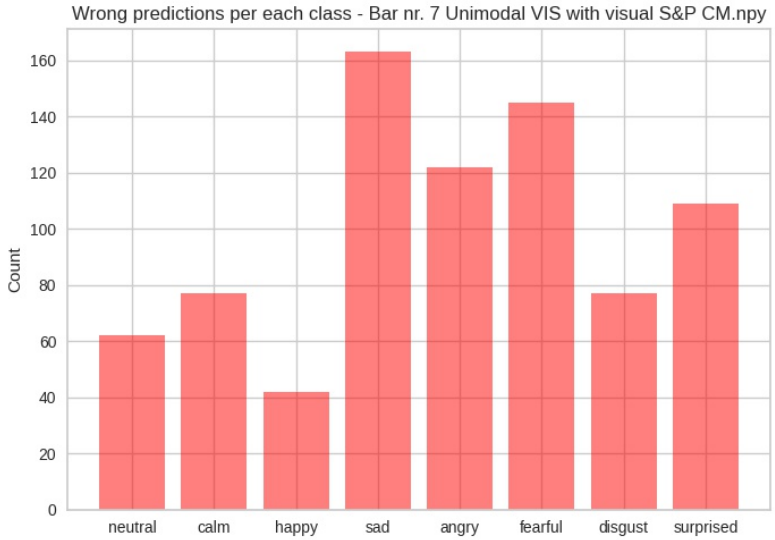

Figure C.18: An overview of incorrect predictions for FER with visual augmentation - Salt & Pepper Noise.

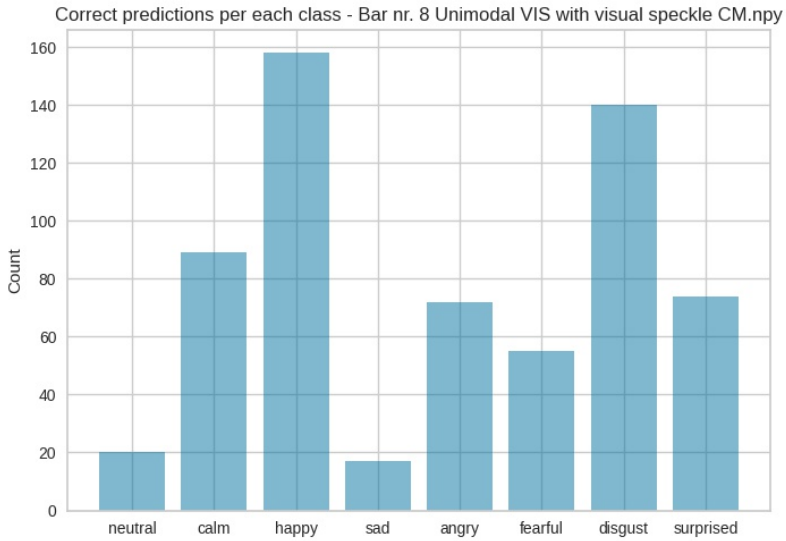

Figure C.19: An overview of correct predictions for FER with visual augmentation - Speckle Noise.

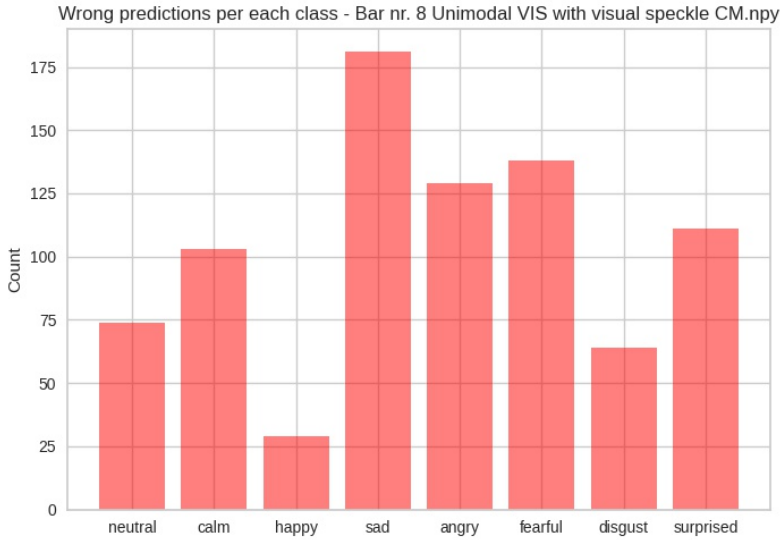

Figure C.20: An overview of incorrect predictions for FER with visual augmentation - Speckle Noise.

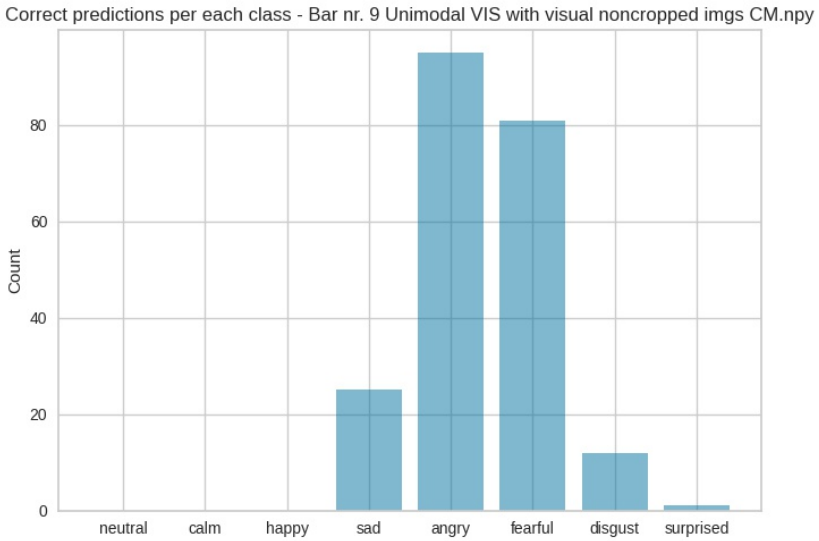

Figure C.21: An overview of correct predictions for FER with visual augmentation - Non-cropped frames.

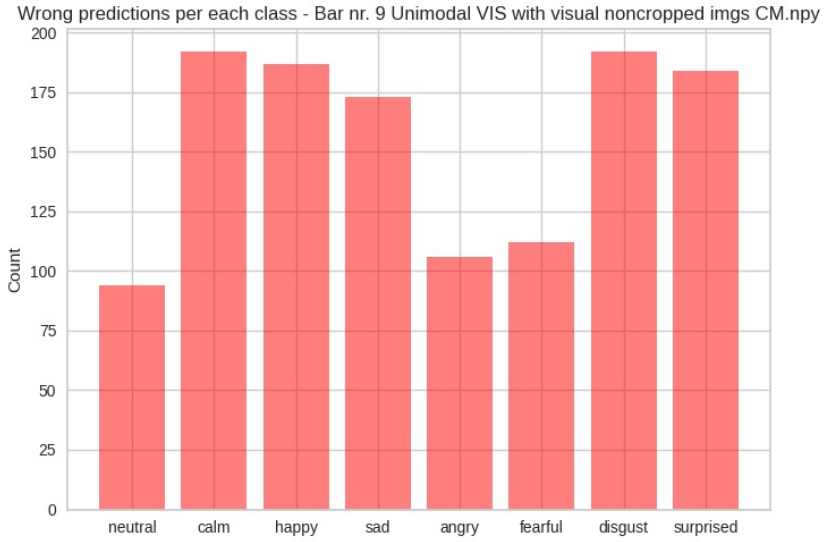

Figure C.22: An overview of correct predictions for FER with visual augmentation - Non-cropped frames.

C.3 Third Section - Audio augmentation

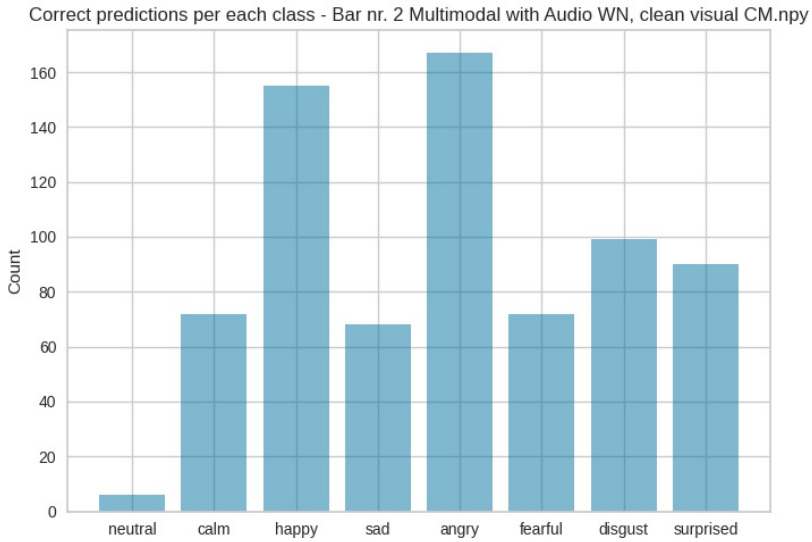

Figure C.23: An overview of correct predictions for AVER with audio augmentation - White Noise.

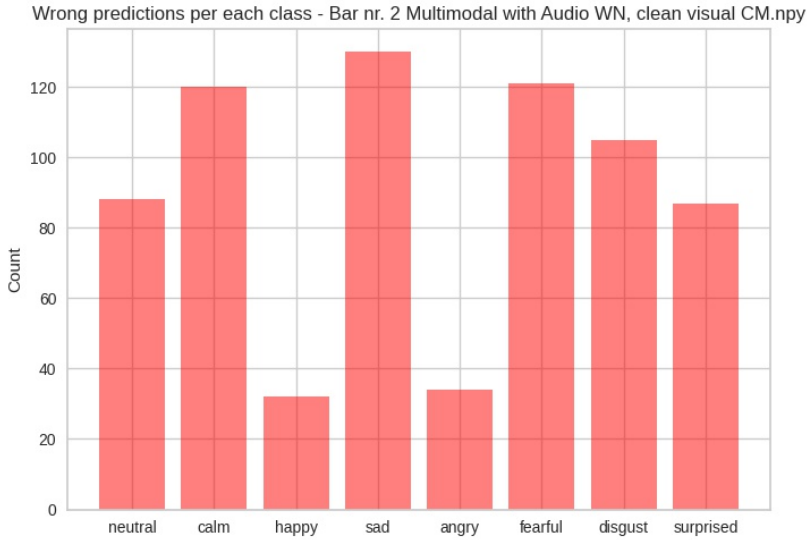

Figure C.24: An overview of incorrect predictions for AVER with audio augmentation - White Noise.

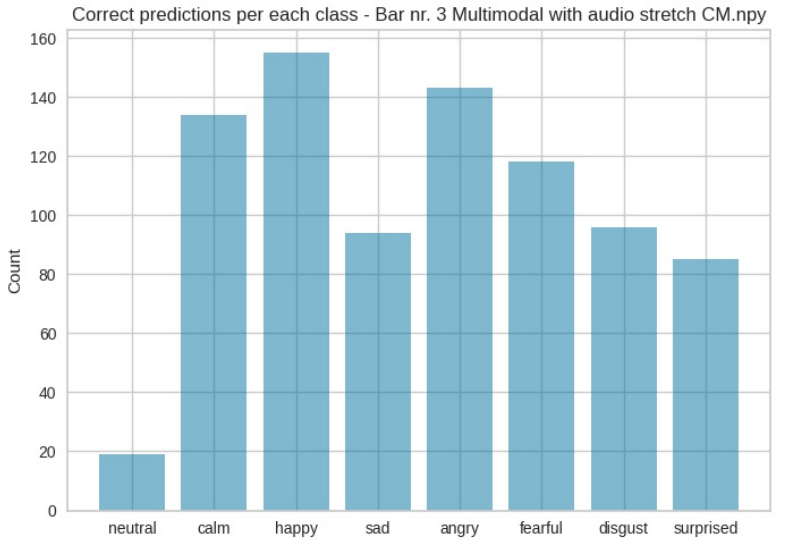

Figure C.25: An overview of correct predictions for AVER with audio augmentation - Stretch.

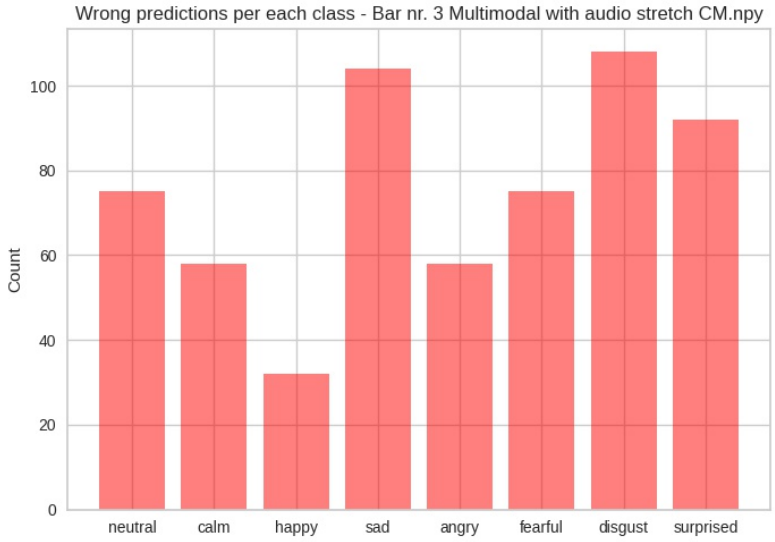

Figure C.26: An overview of incorrect predictions for AVER with audio augmentation - Stretch.

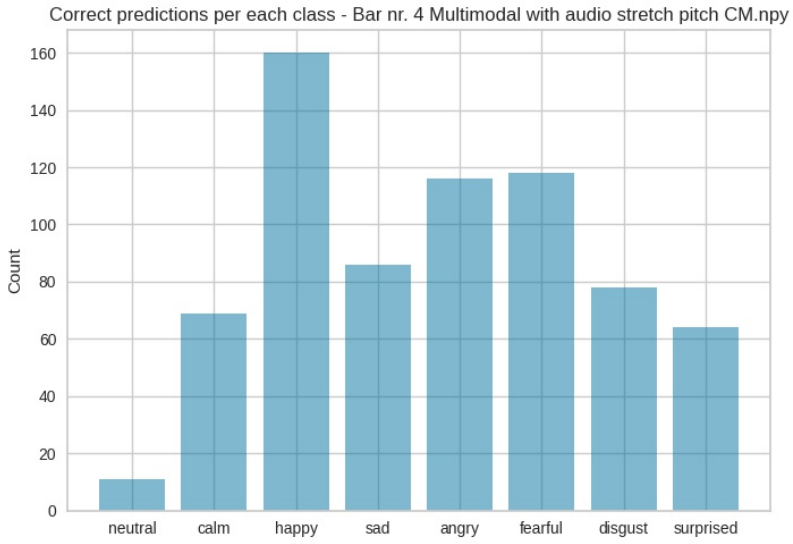

Figure C.27: An overview of correct predictions for AVER with audio augmentation - Stretch pitch.

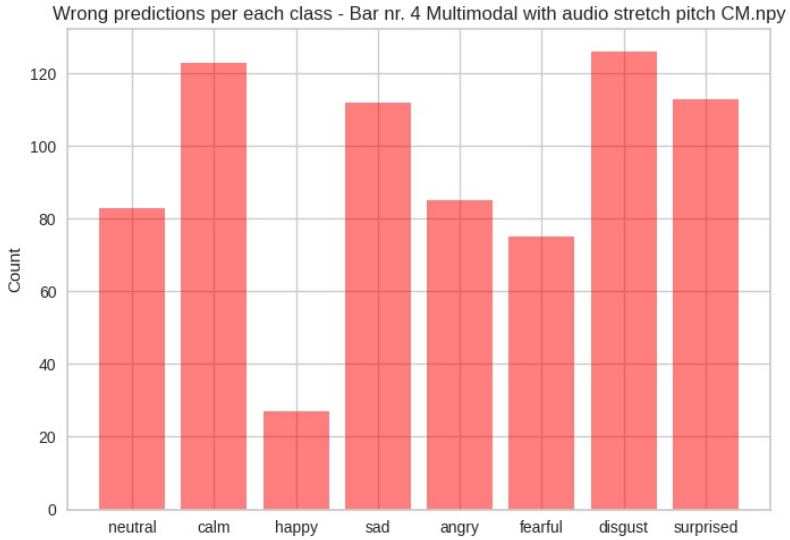

Figure C.28: An overview of incorrect predictions for AVER with audio augmentation - Stretch pitch.

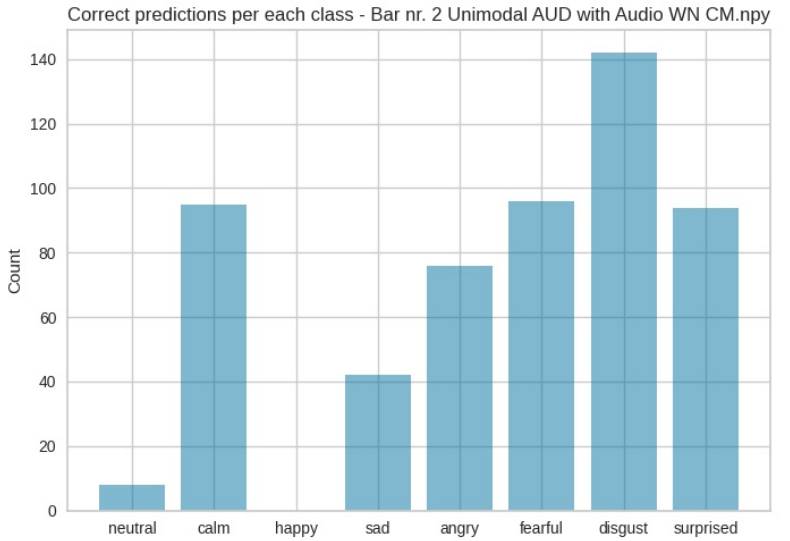

Figure C.29: An overview of correct predictions for SER with audio augmentation - White Noise.

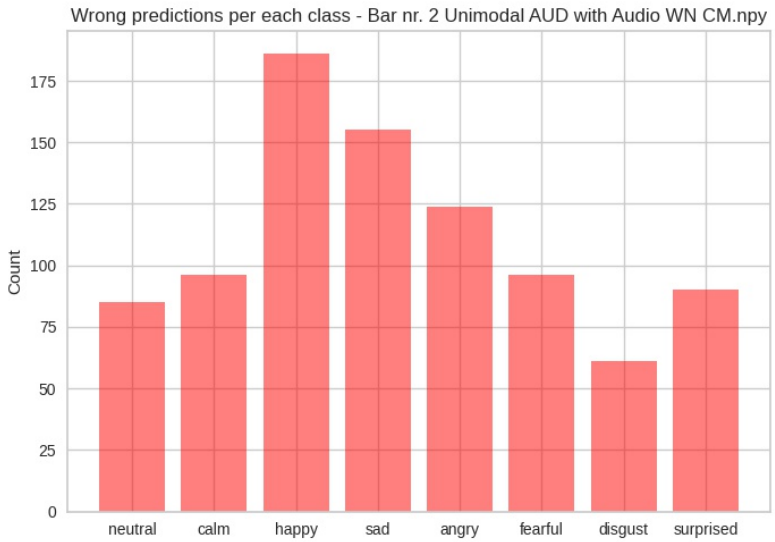

Figure C.30: An overview of incorrect predictions for SER with audio augmentation - White Noise.

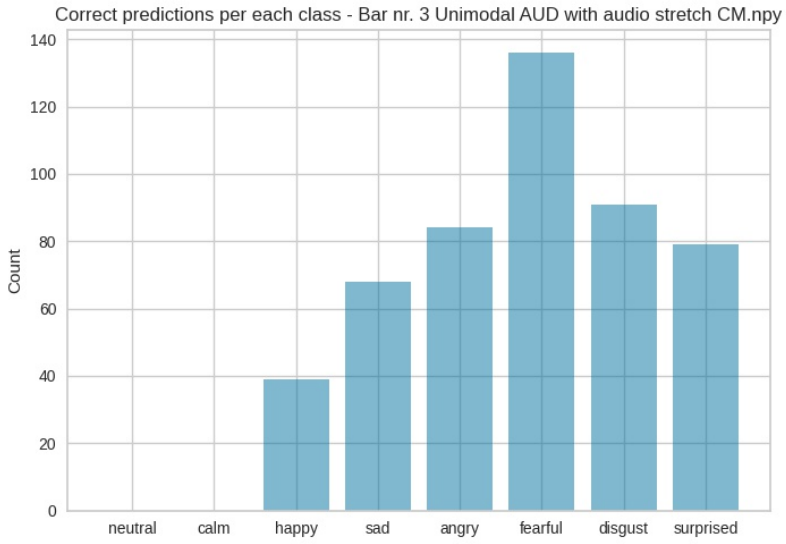

Figure C.31: An overview of correct predictions for SER with audio augmentation - Stretch.

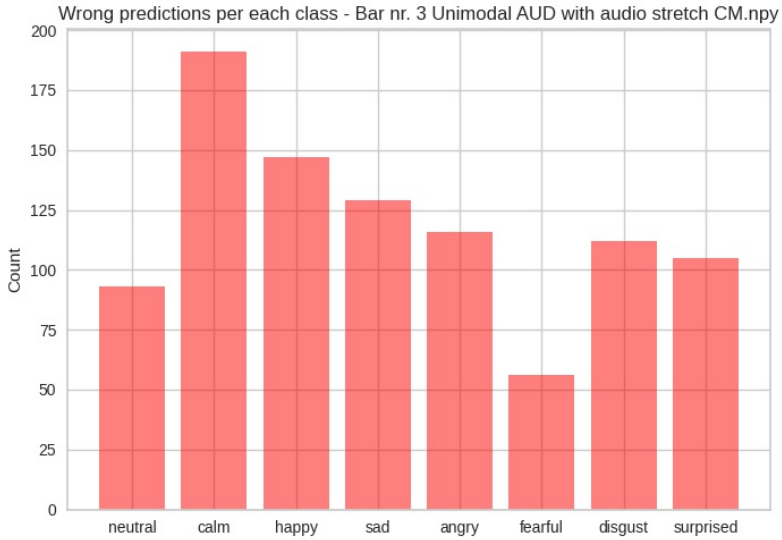

Figure C.32: An overview of correct predictions for SER with audio augmentation - Stretch.

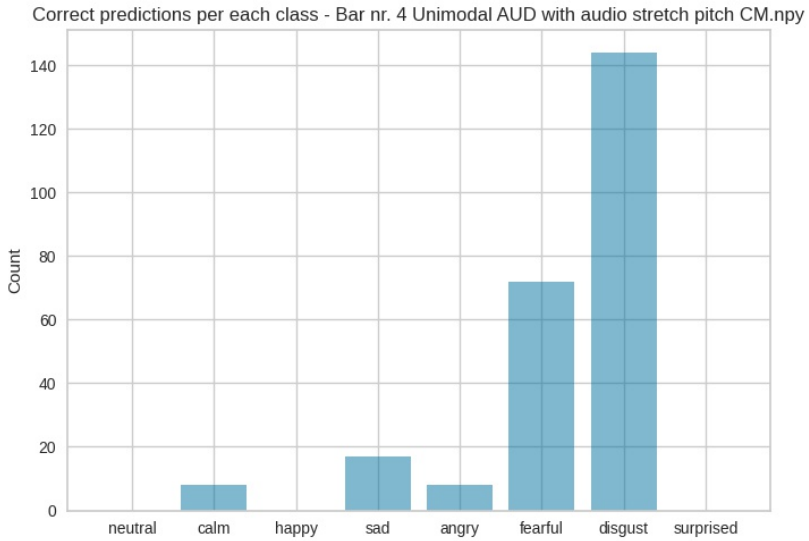

Figure C.33: An overview of incorrect predictions for SER with audio augmentation - Stretch pitch.

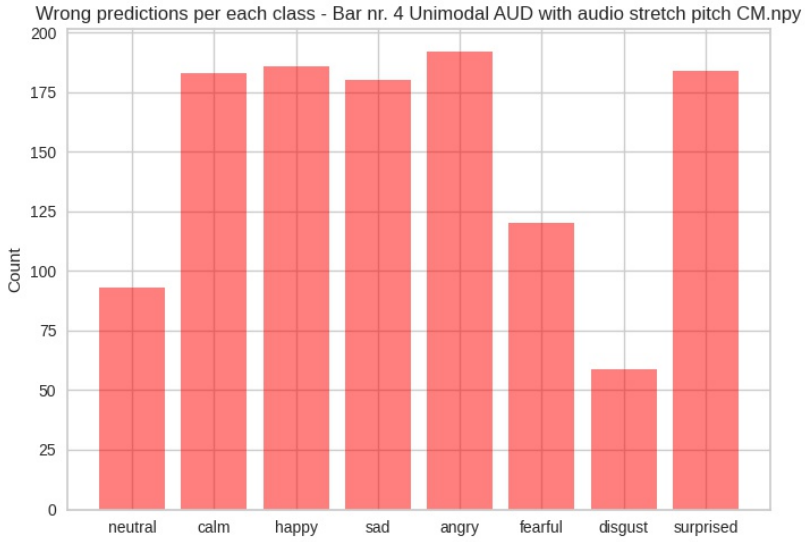

Figure C.34: An overview of incorrect predictions for SER with audio augmentation - Stretch pitch.

C.4 Fourth Section - Audio and Visual augmentation

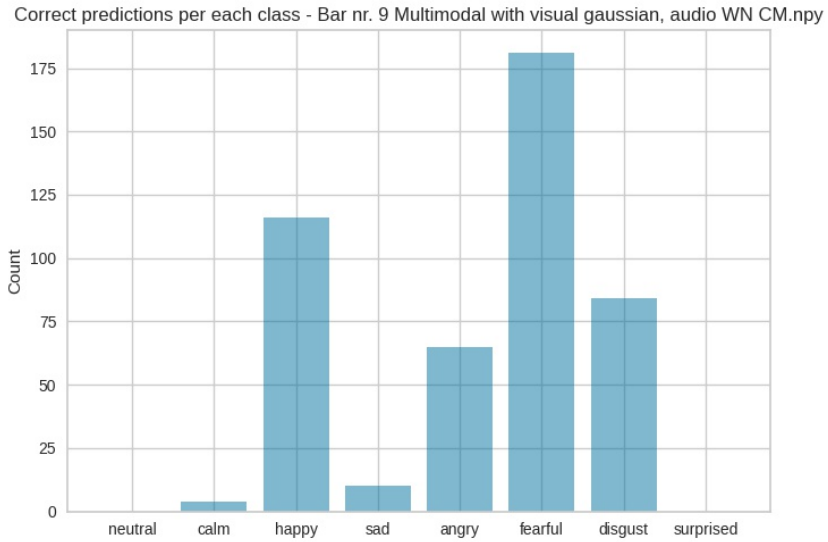

Figure C.35: An overview of correct predictions for AVER with augmentations visual - Gaussian Noise and audio - White Noise.

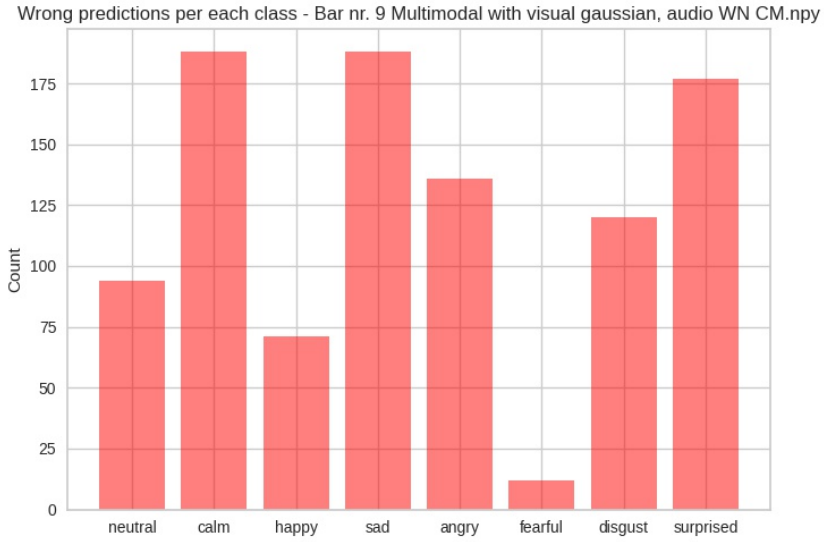

Figure C.36: An overview of incorrect predictions for AVER with augmentations visual - Gaussian Noise and audio - White Noise.

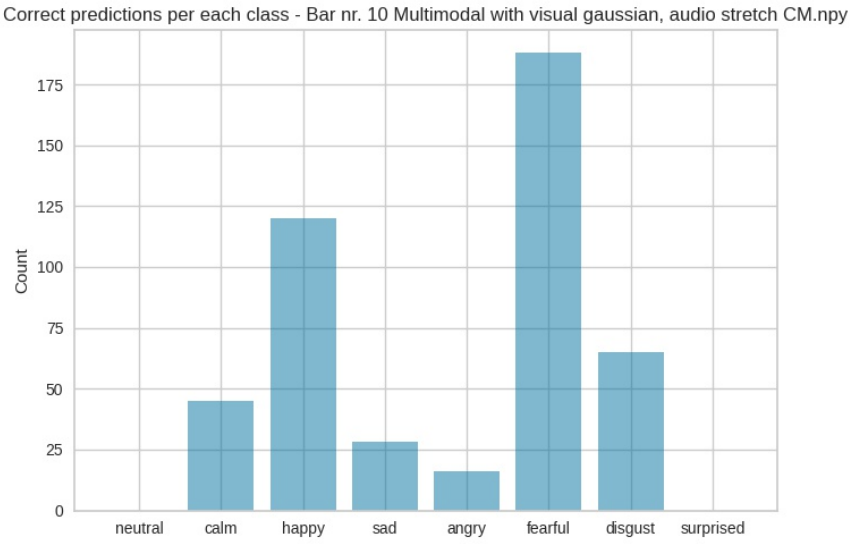

Figure C.37: An overview of correct predictions for AVER with augmentations visual - Gaussian Noise and audio - Stretch.

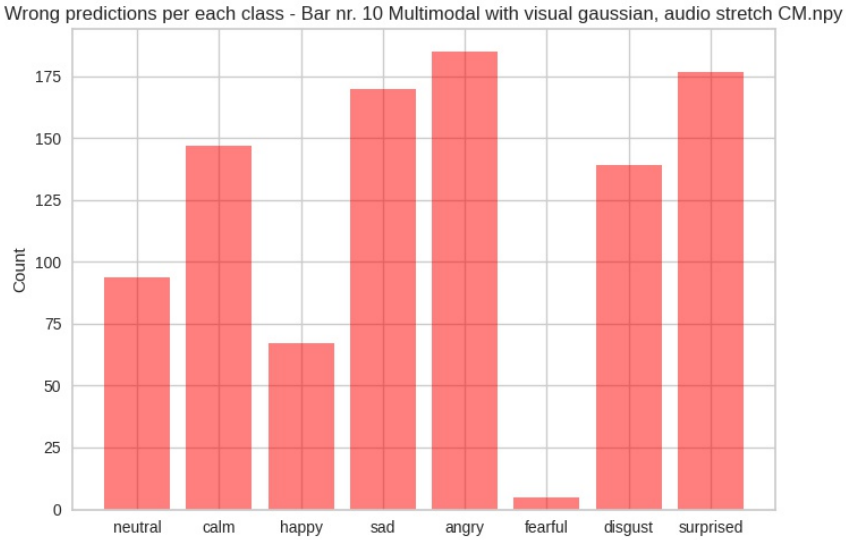

Figure C.38: An overview of incorrect predictions for AVER with augmentations visual - Gaussian Noise and audio - Stretch.

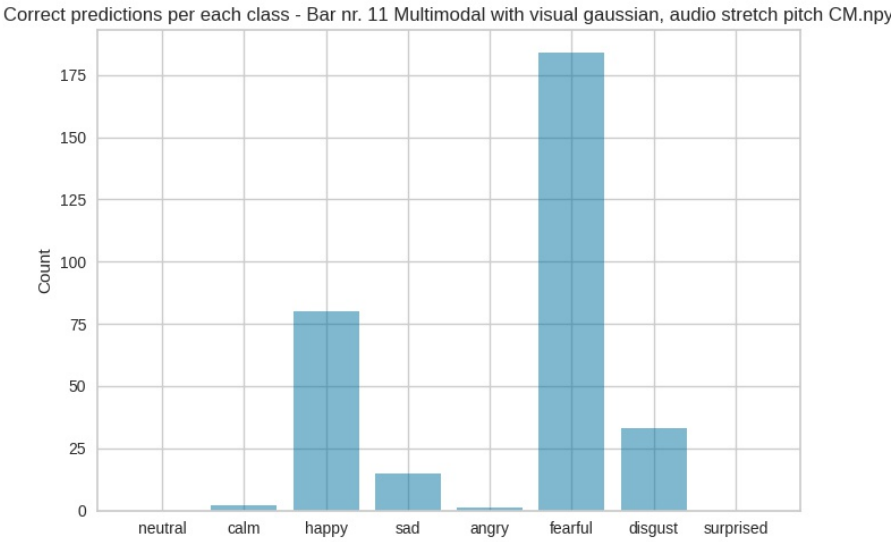

Figure C.39: An overview of correct predictions for AVER with augmentations visual - Gaussian Noise and audio - Stretch Pitch.

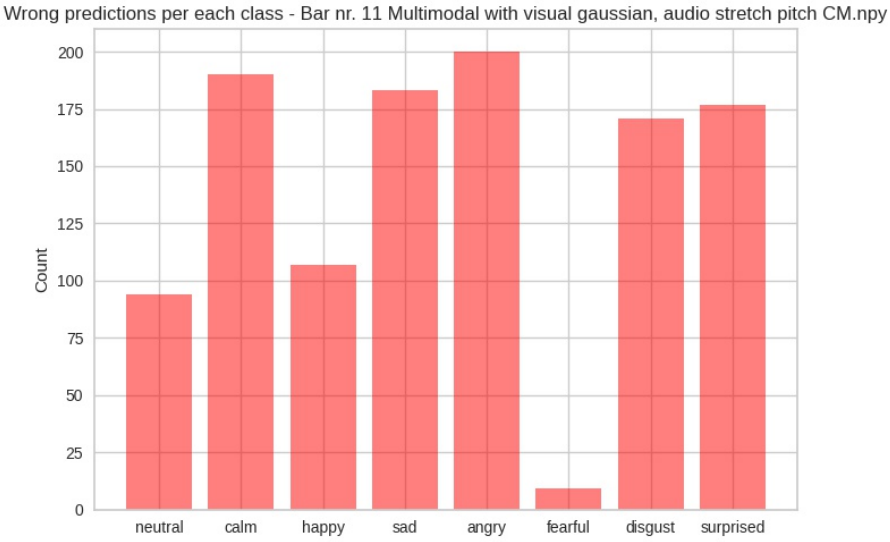

Figure C.40: An overview of incorrect predictions for AVER with augmentations visual - Gaussian Noise and audio - Stretch Pitch.

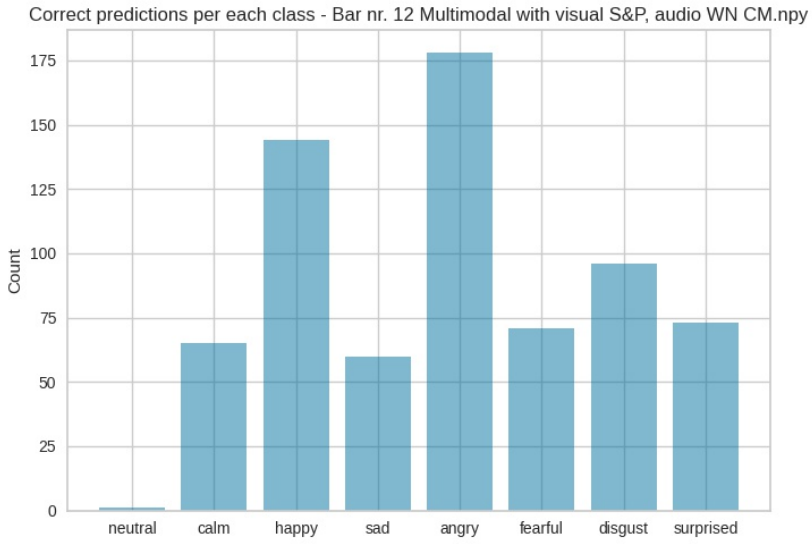

Figure C.41: An overview of correct predictions for AVER with augmentations visual - Salt & Pepper Noise and audio - White Noise.

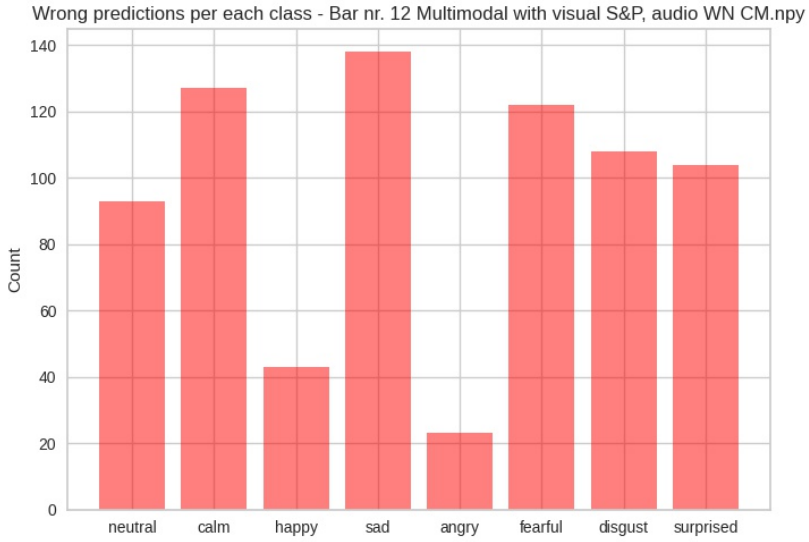

Figure C.42: An overview of incorrect predictions for AVER with augmentations visual - Salt & Pepper Noise and audio - White Noise.

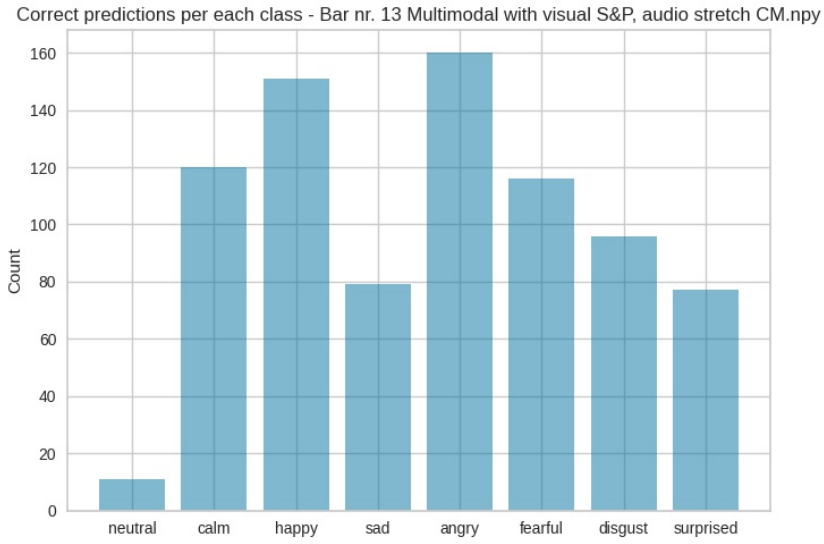

Figure C.43: An overview of correct predictions for AVER with augmentations visual - Salt & Pepper Noise and audio - Stretch.

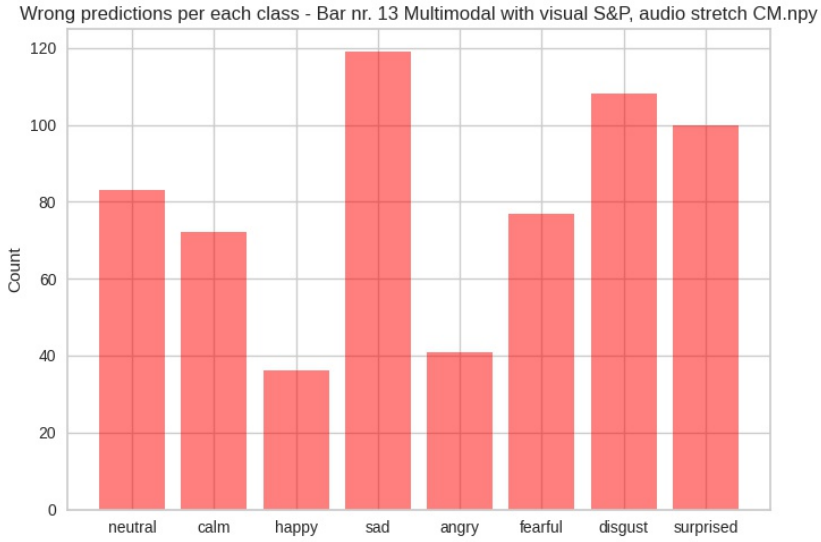

Figure C.44: An overview of incorrect predictions for AVER with augmentations visual - Salt & Pepper Noise and audio - Stretch.

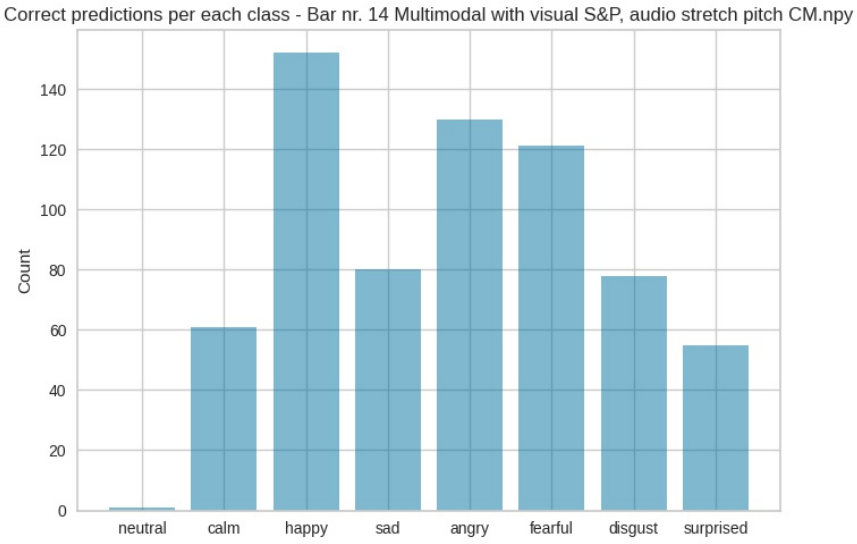

Figure C.45: An overview of correct predictions for AVER with augmentations visual - Salt & Pepper Noise and audio - Stretch Pitch.

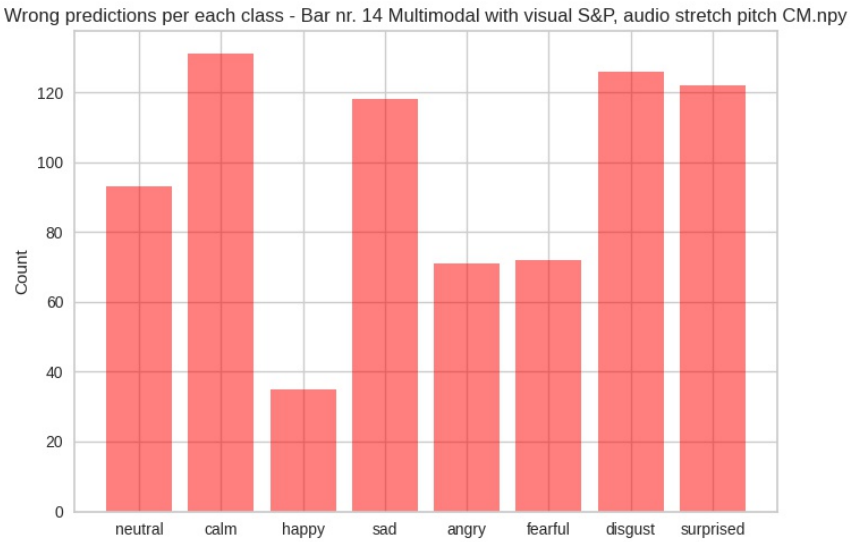

Figure C.46: An overview of incorrect predictions for AVER with augmentations visual - Salt & Pepper Noise and audio - Stretch Pitch.

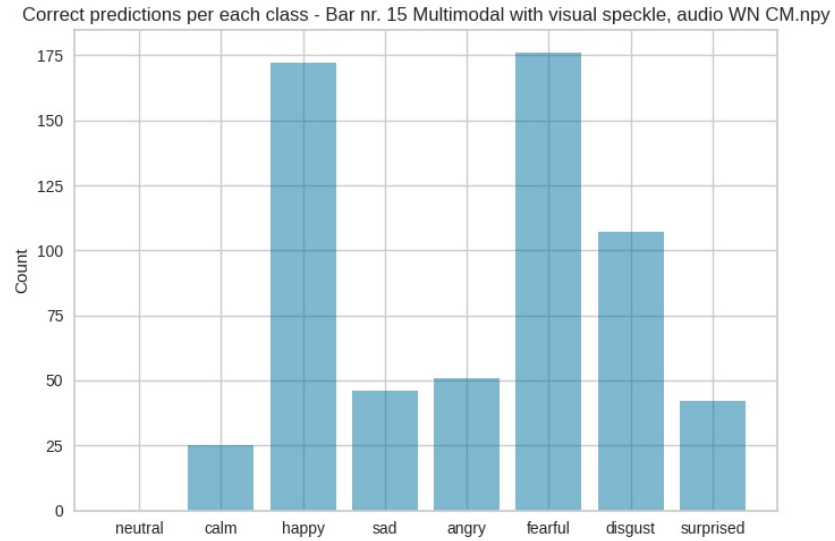

Figure C.47: An overview of correct predictions for AVER with augmentations visual - Speckle Noise and audio - White Noise.

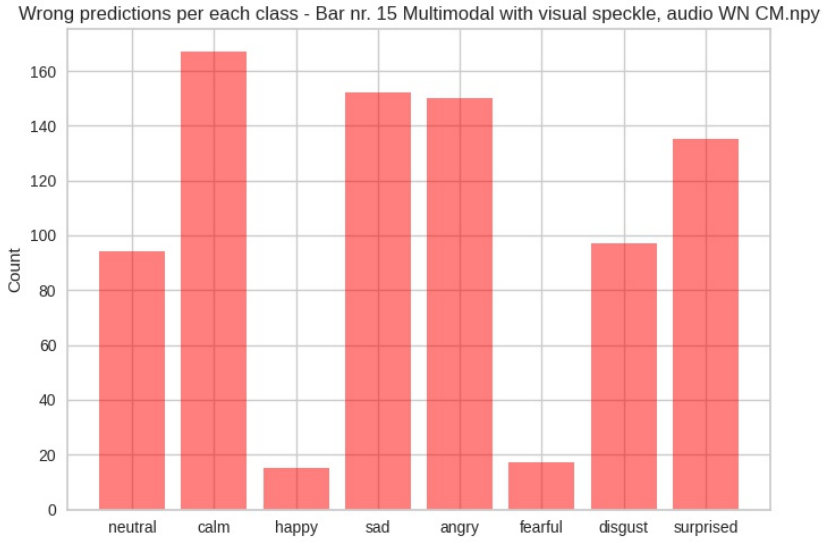

Figure C.48: An overview of incorrect predictions for AVER with augmentations visual - Speckle Noise and audio - White Noise.

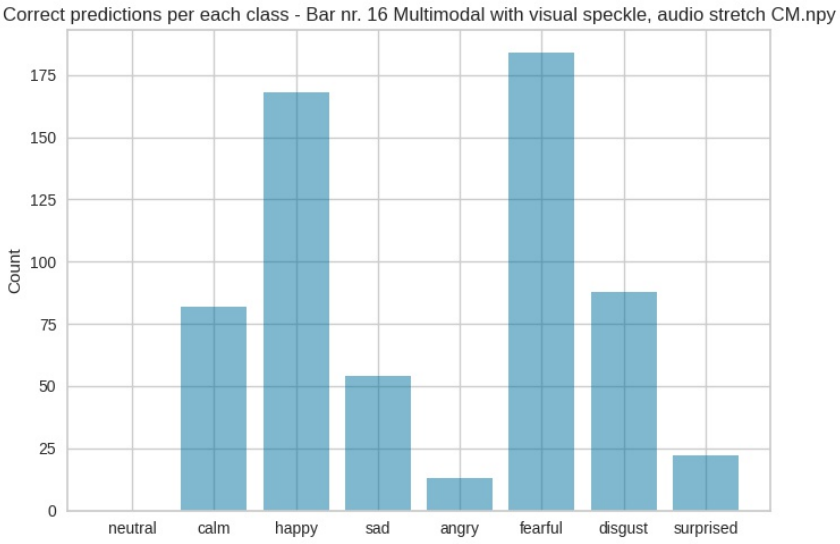

Figure C.49: An overview of correct predictions for AVER with augmentations visual - Speckle Noise and audio - Stretch.

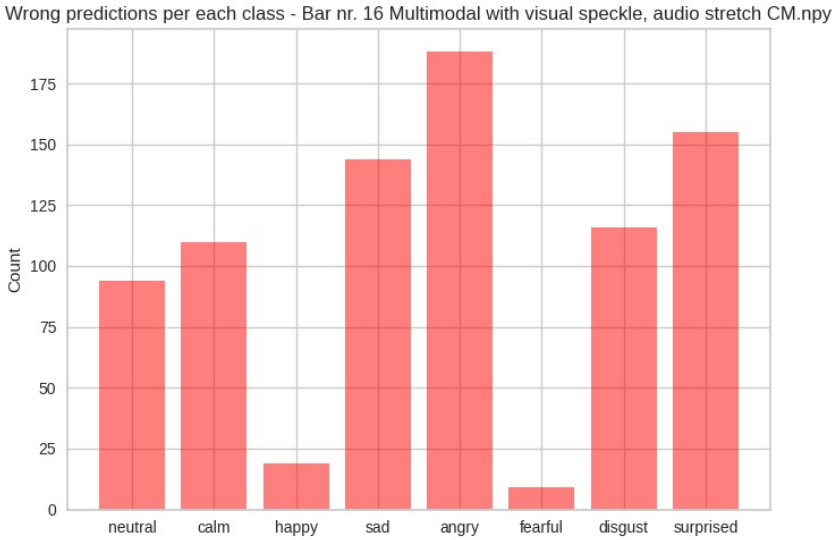

Figure C.50: An overview of incorrect predictions for AVER with augmentations visual - Speckle Noise and audio - Stretch.

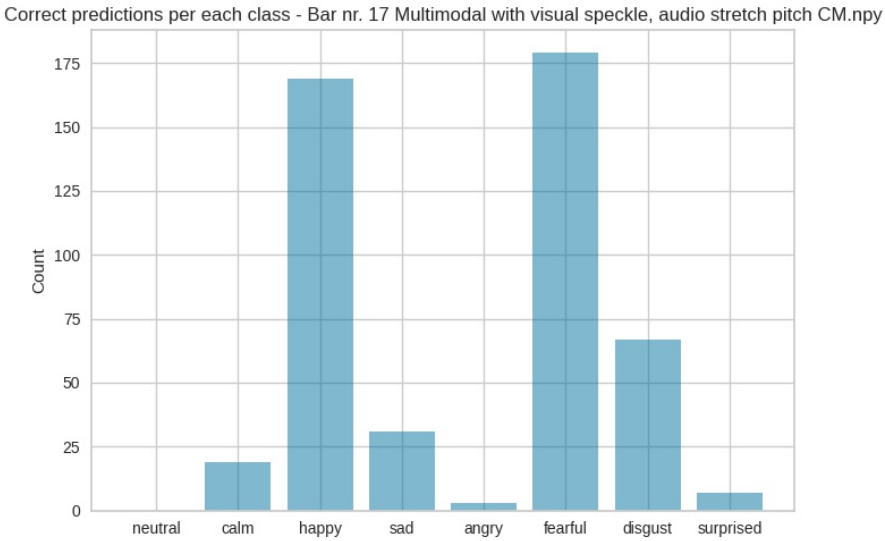

Figure C.51: An overview of correct predictions for AVER with augmentations visual - Speckle Noise and audio - Stretch Pitch.

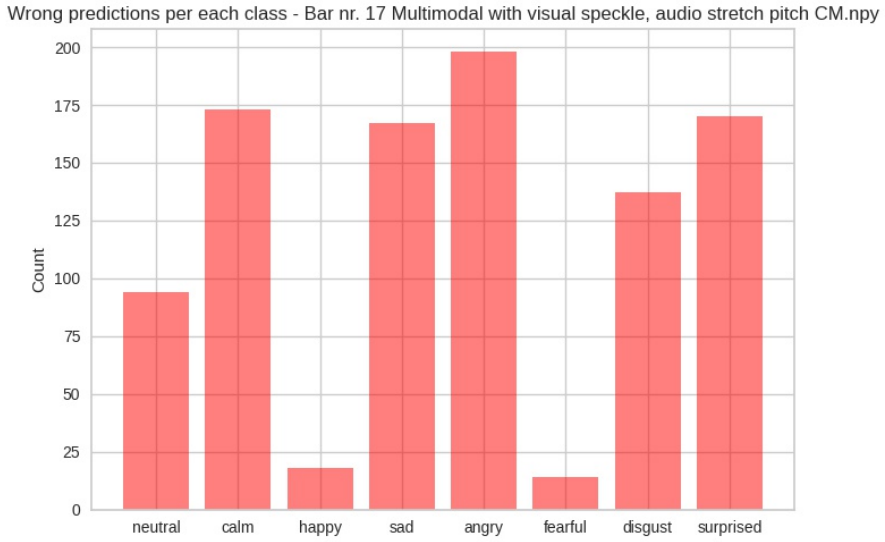

Figure C.52: An overview of incorrect predictions for AVER with augmentations visual - Speckle Noise and audio - Stretch Pitch.

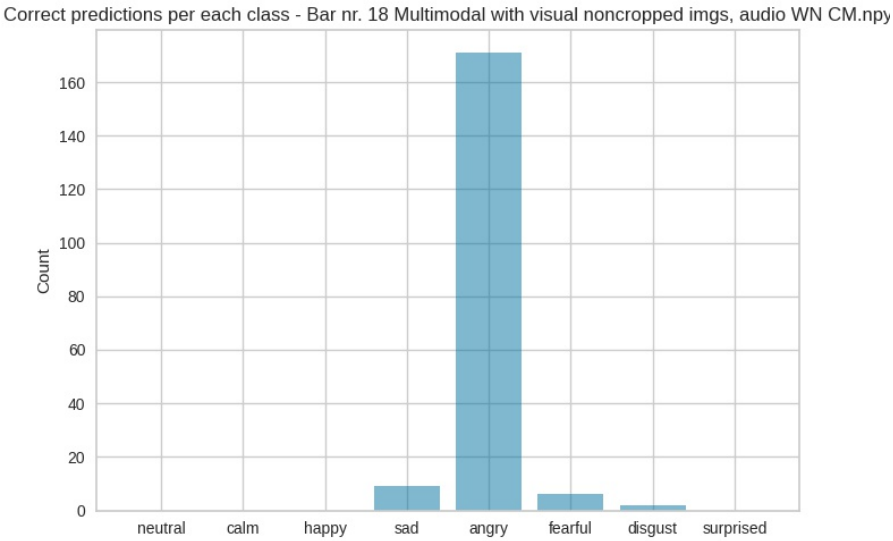

Figure C.53: An overview of correct predictions for AVER with augmentations visual - non-cropped frames and audio - White Noise.

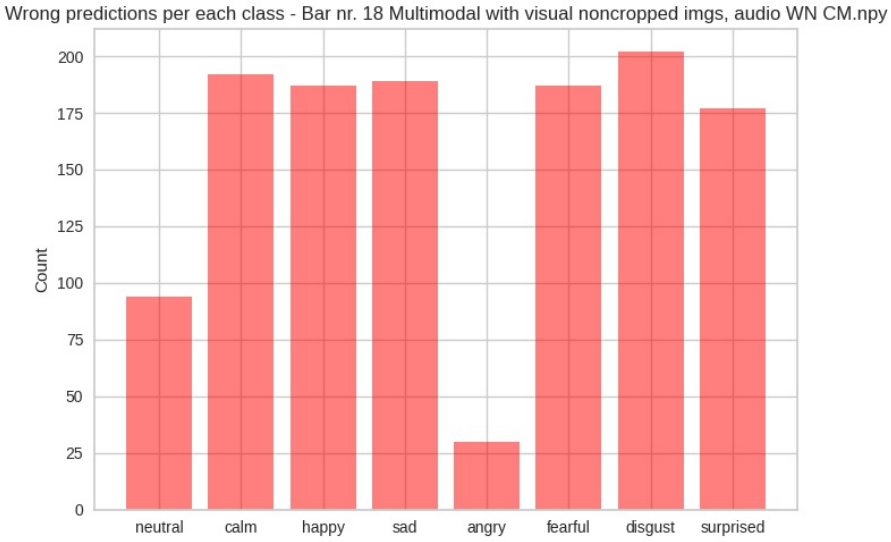

Figure C.54: An overview of incorrect predictions for AVER with augmentations visual - non-cropped frames and audio - White Noise.

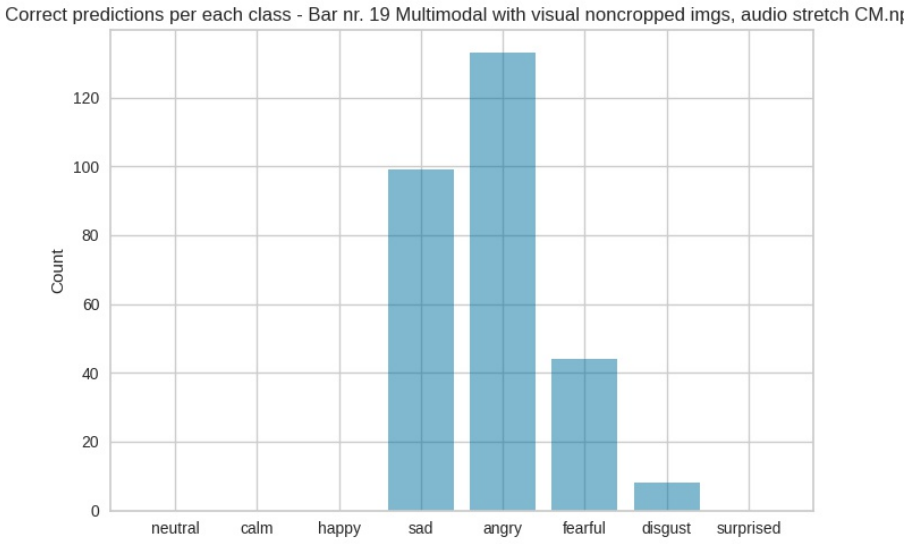

Figure C.55: An overview of correct predictions for AVER with augmentations visual - non-cropped frames and audio - Stretch.

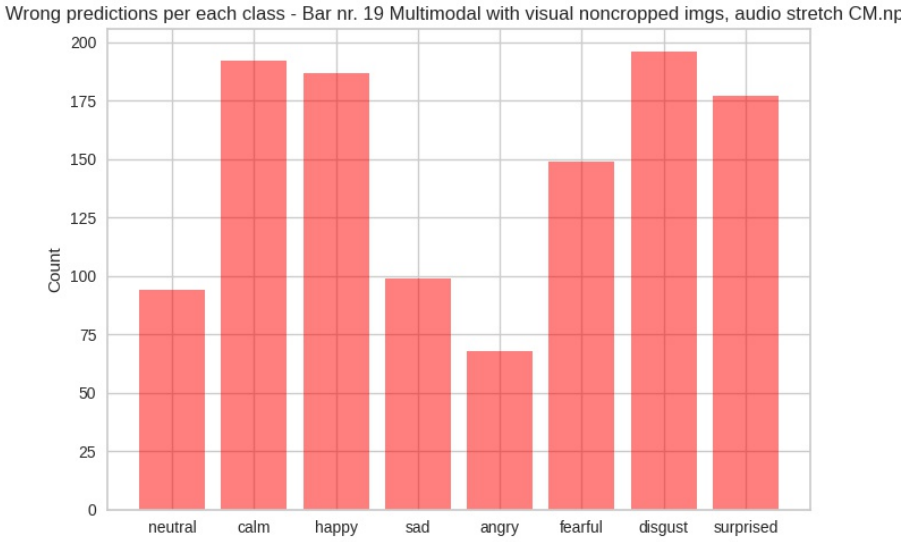

Figure C.56: An overview of incorrect predictions for AVER with augmentations visual - non-cropped frames and audio - Stretch.

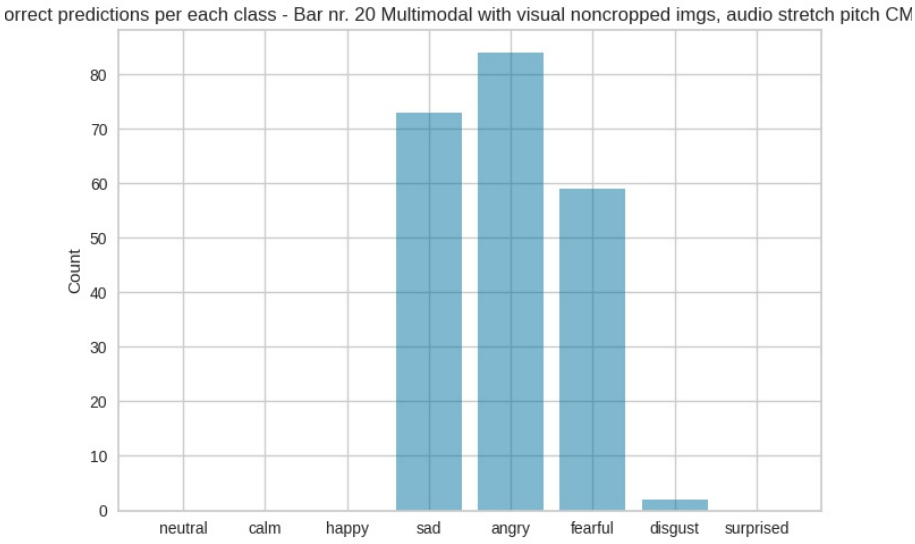

Figure C.57: An overview of correct predictions for AVER with augmentations visual - non-cropped frames and audio - Stretch Pitch.

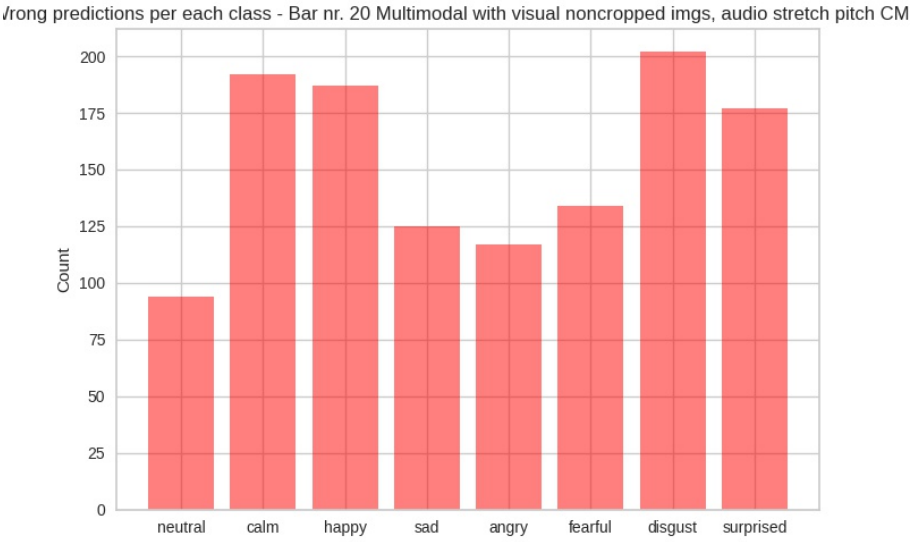

Figure C.58: An overview of incorrect predictions for AVER with augmentations visual - non-cropped frames and audio - Stretch Pitch.

APPENDIX D

The fourth Appendix - experiments t-sne figures overview

It was detected that emotion surprised failed to be displayed.

D.1 First Section - baselines

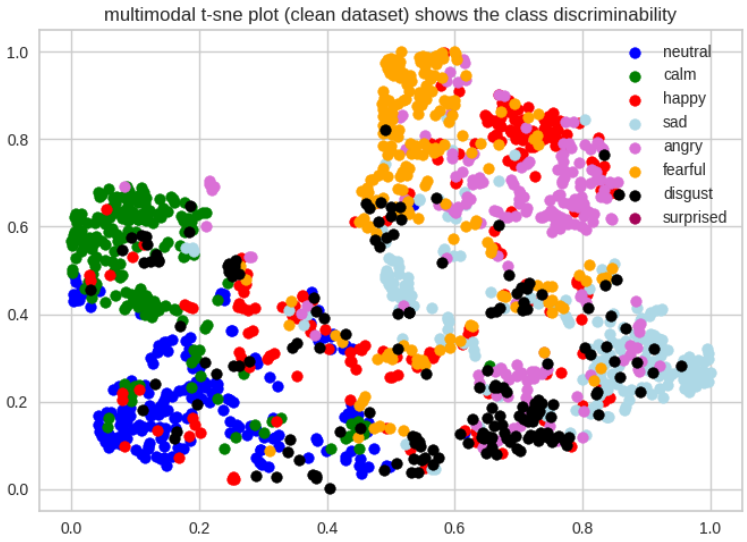

Figure D.1: T-sne for AVER Baseline.

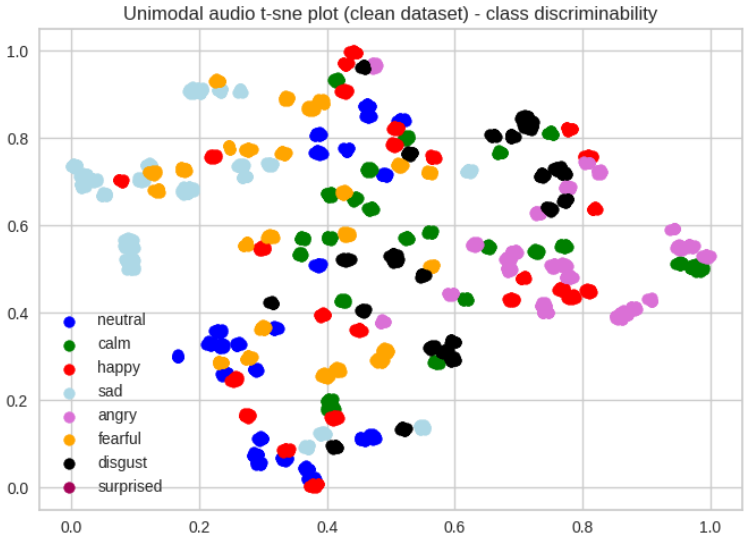

Figure D.2: T-sne for SER Baseline.

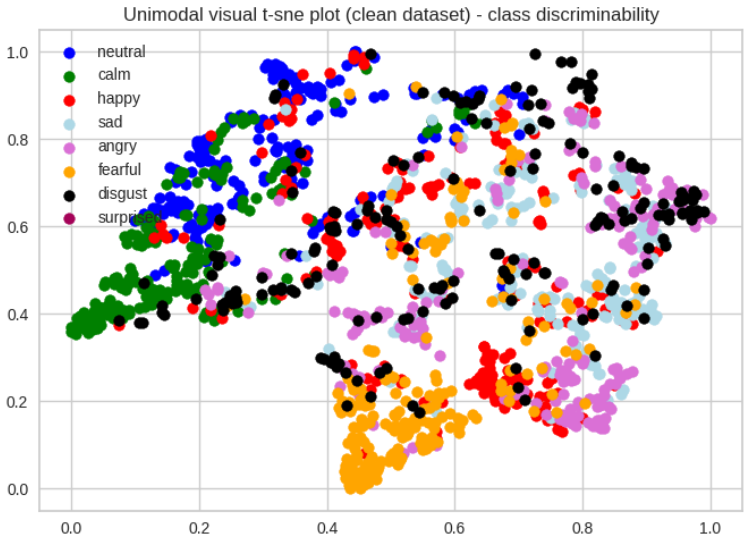

Figure D.3: T-sne for FER Baseline.

D.2 Second Section - Visual augmentation

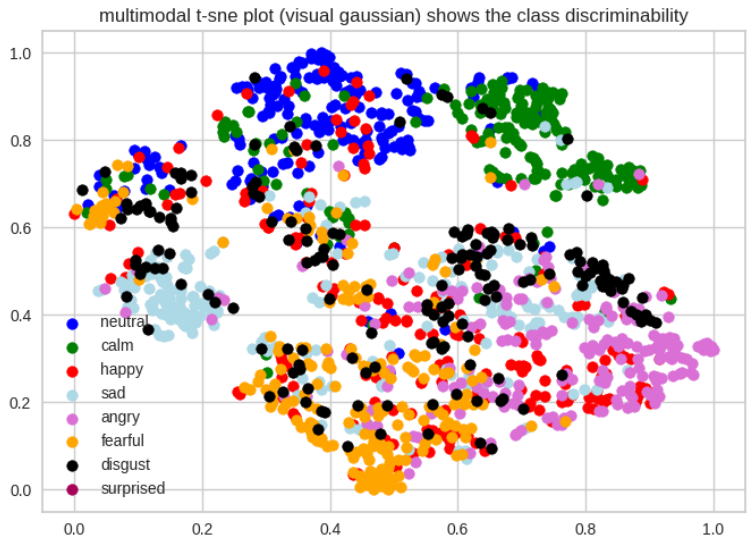

Figure D.4: T-sne for AVER with visual augmentation - Gaussian Noise.

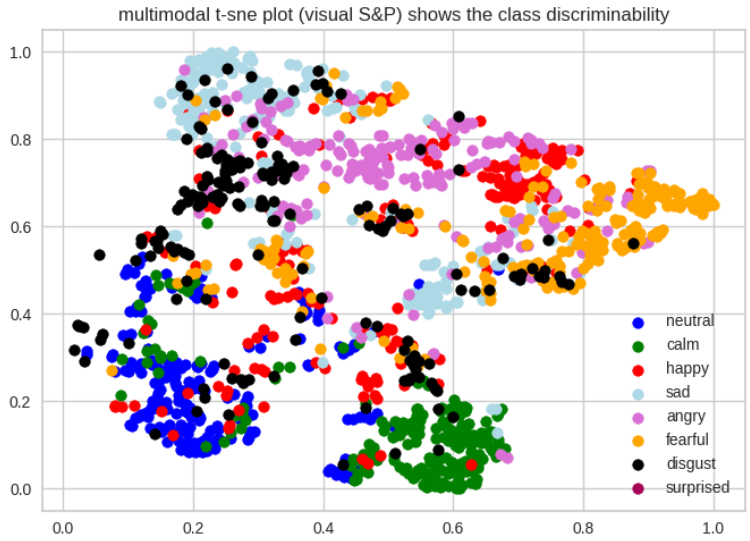

Figure D.5: T-sne for AVER with visual augmentation - Salt & Pepper Noise.

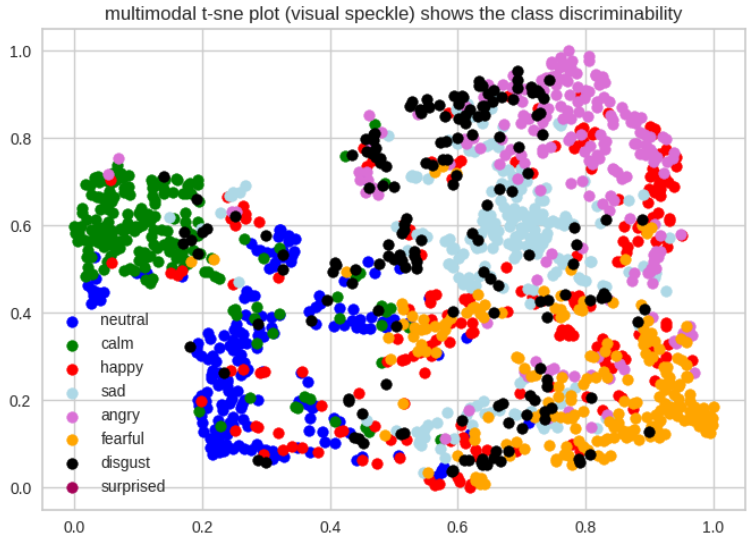

Figure D.6: T-sne for AVER with visual augmentation - Speckle Noise.

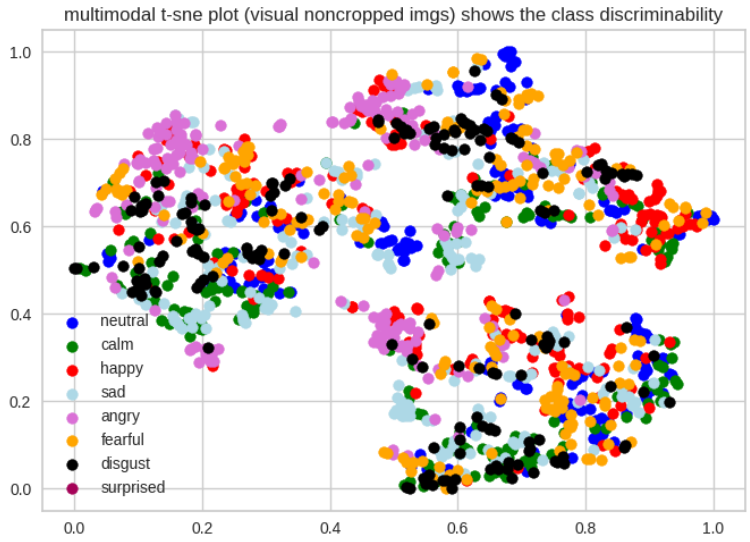

Figure D.7: T-sne for AVER with visual augmentation - Non-cropped frames.

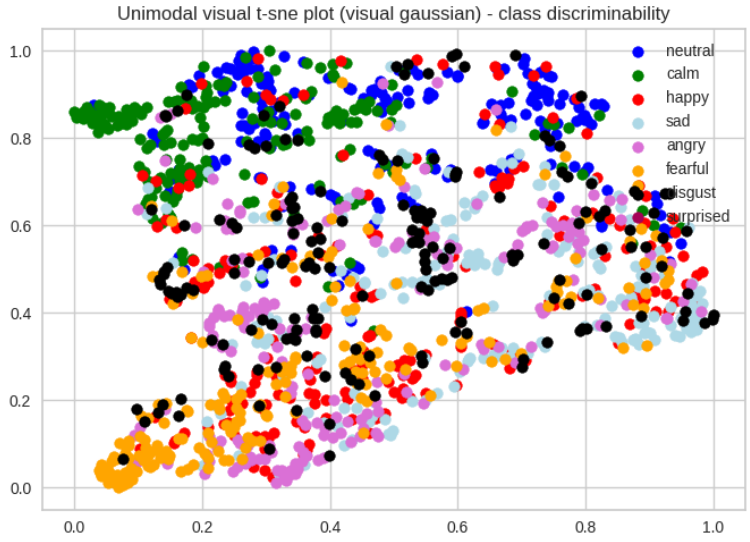

Figure D.8: T-sne for FER with visual augmentation - Gaussian noise.

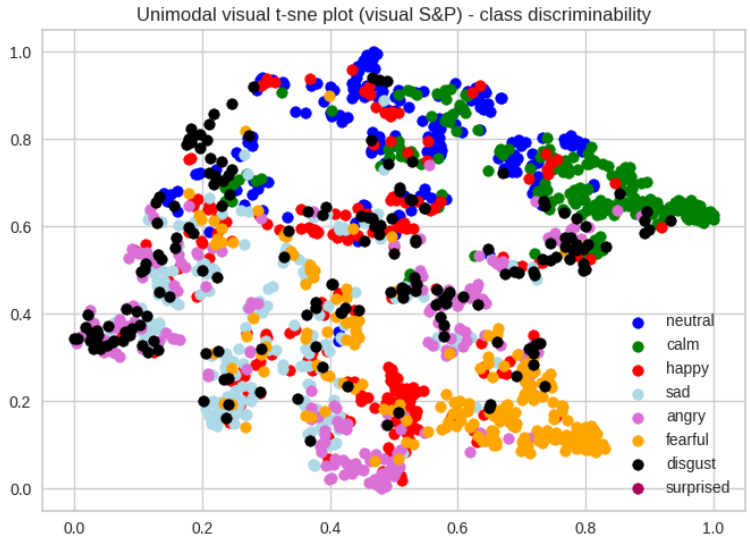

Figure D.9: T-sne for FER with visual augmentation - Salt & Pepper Noise.

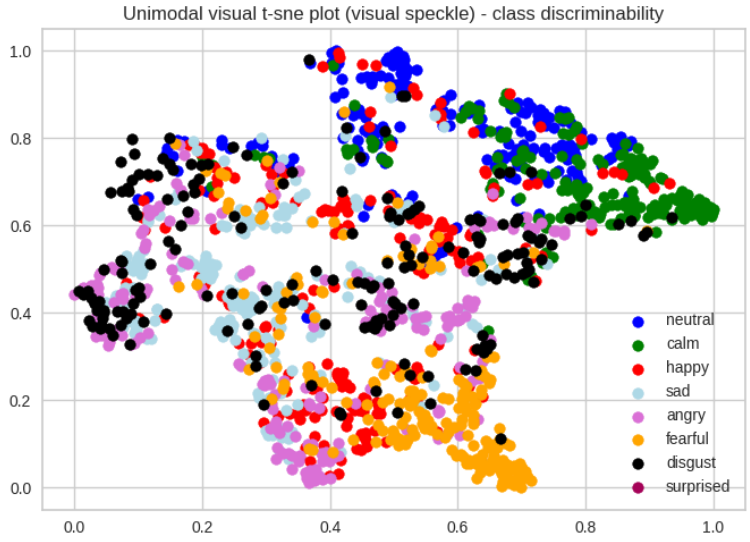

Figure D.10: T-sne for FER with visual augmentation - Speckle Noise.

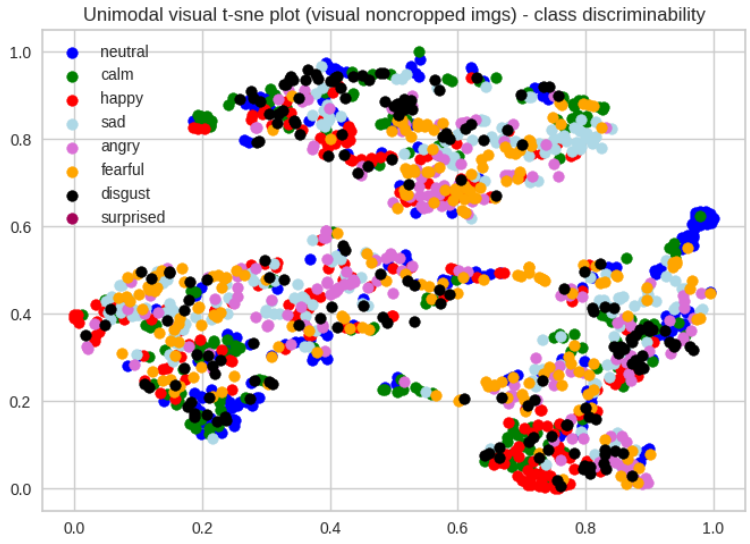

Figure D.11: T-sne for FER with visual augmentation - Non-cropped frames.

D.3 Third Section - Audio augmentation

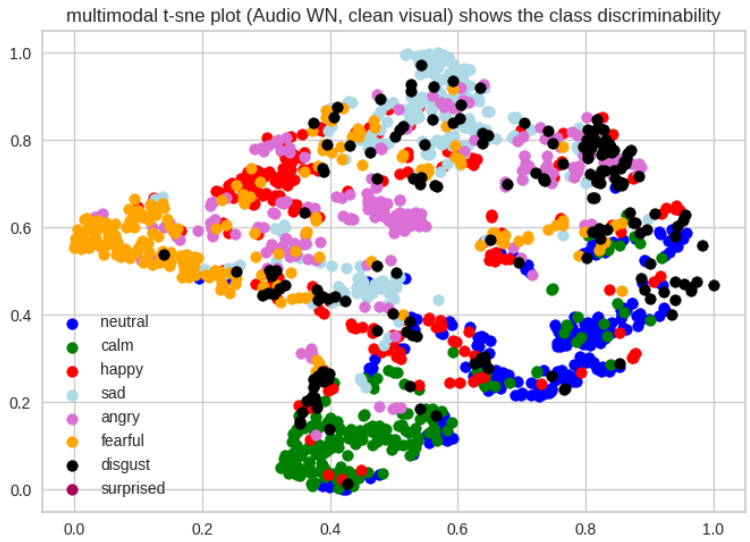

Figure D.12: T-sne for AVER with audio augmentation - White Noise.

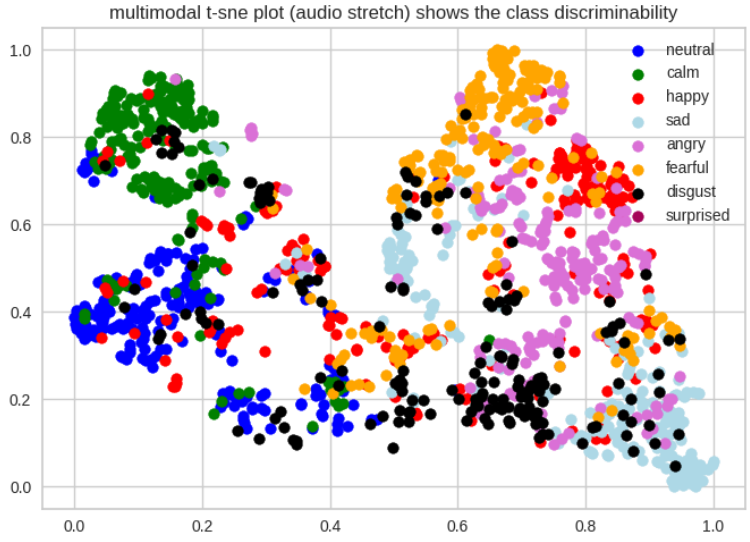

Figure D.13: T-sne for AVER with audio augmentation - Stretch.

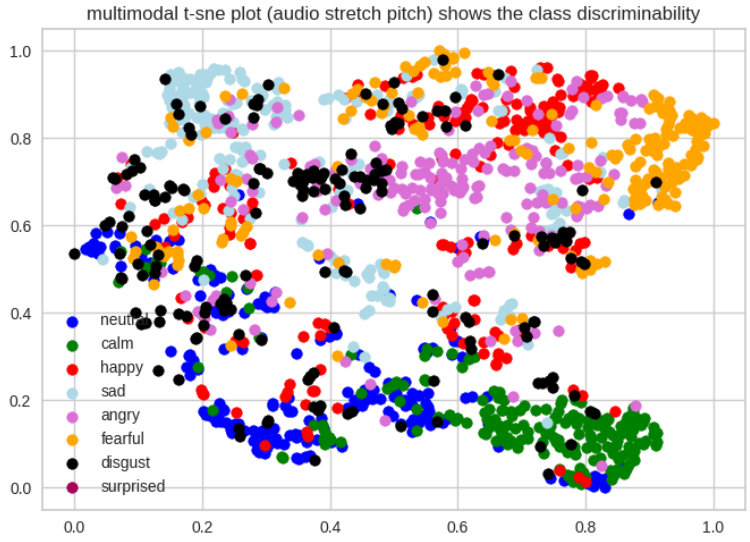

Figure D.14: T-sne for AVER with audio augmentation - Stretch pitch.

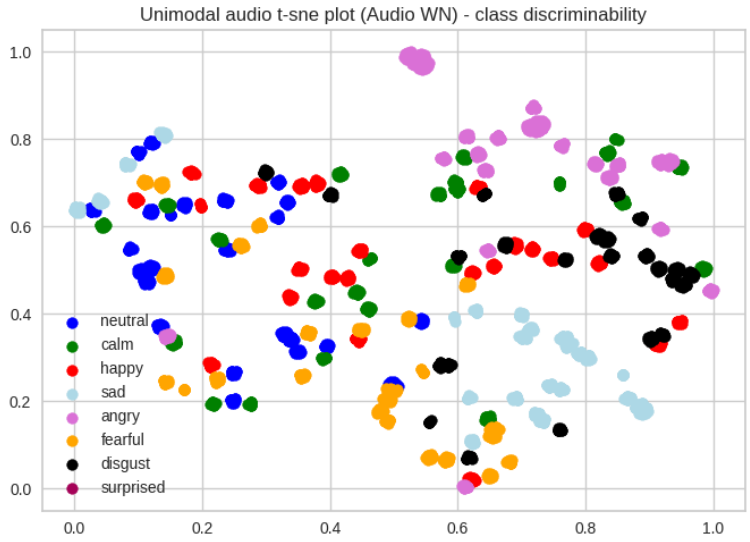

Figure D.15: T-sne for SER with audio augmentation - White Noise.

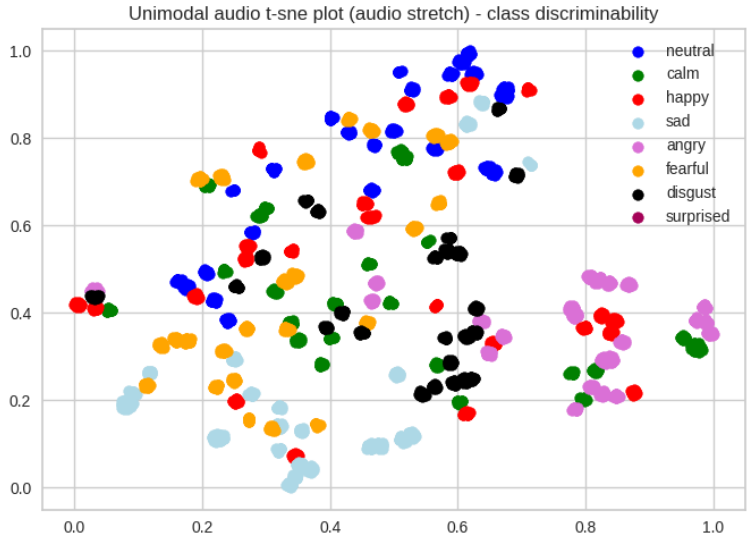

Figure D.16: T-sne for SER with audio augmentation - Stretch.

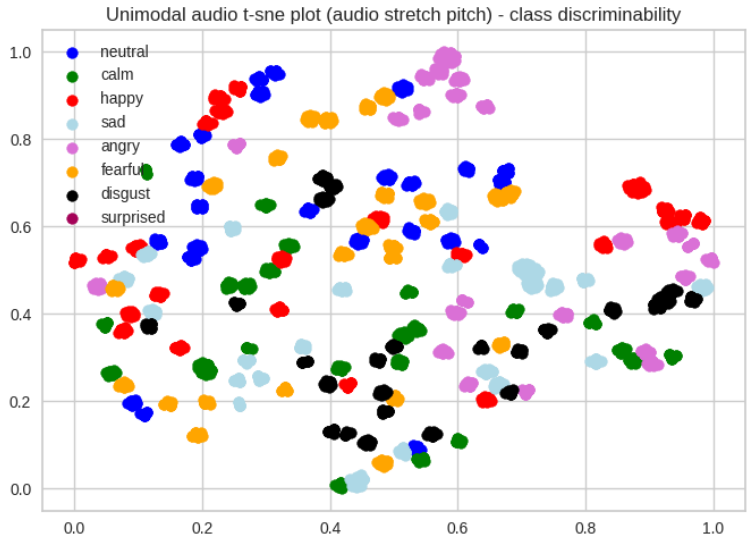

Figure D.17: T-sne for SER with audio augmentation - Stretch pitch.

D.4 Fourth Section - Audio and Visual augmentation

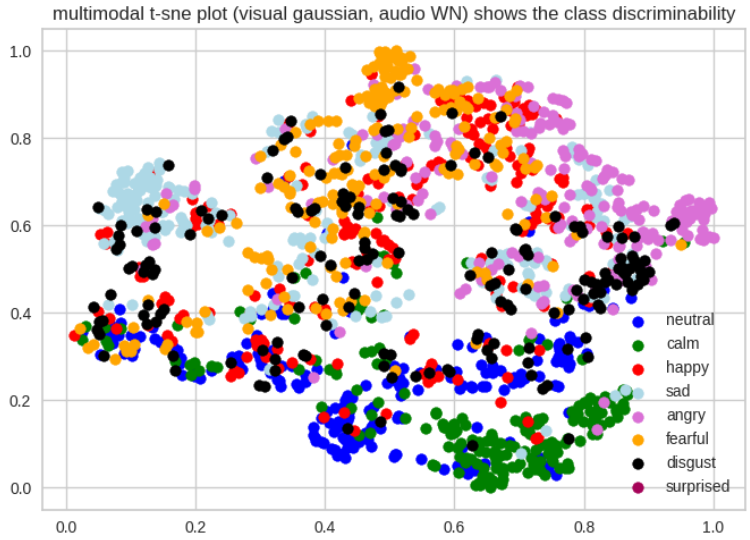

Figure D.18: T-sne for AVER with augmentations visual - Gaussian Noise and audio - White Noise.

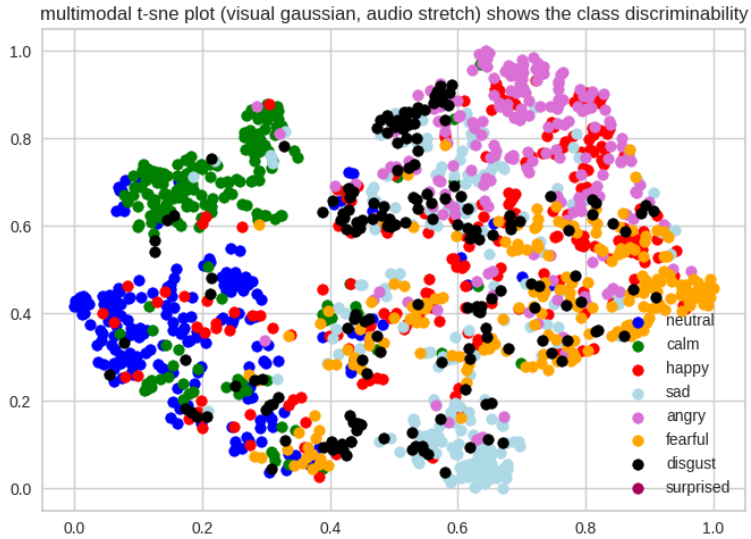

Figure D.19: T-sne for AVER with augmentations visual - Gaussian Noise and audio - Stretch.

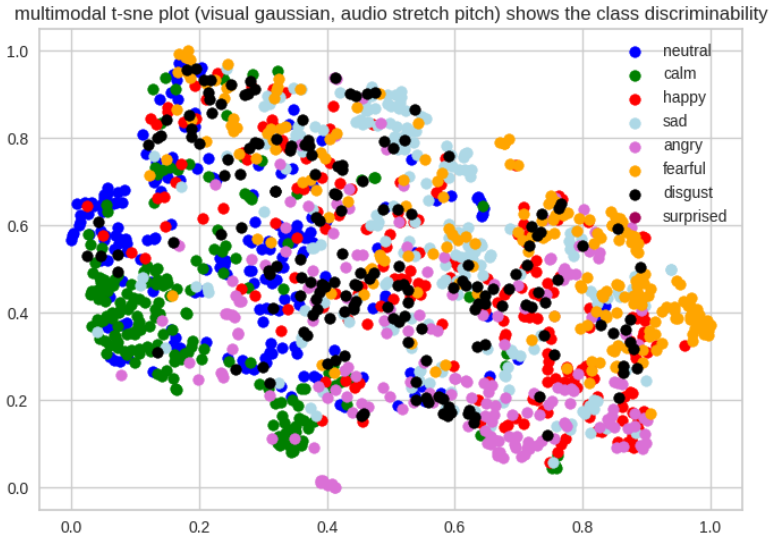

Figure D.20: T-sne for AVER with augmentations visual - Gaussian Noise and audio - Stretch Pitch.

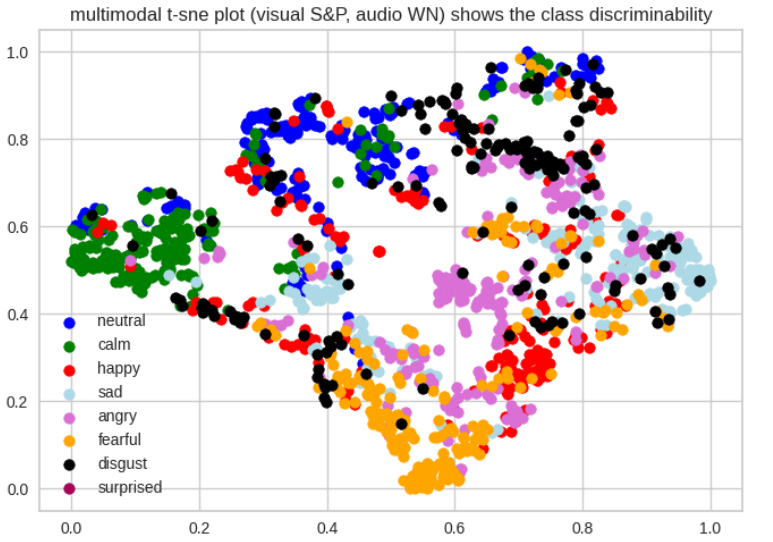

Figure D.21: T-sne for AVER with augmentations visual - Salt & Pepper Noise and audio - White Noise.

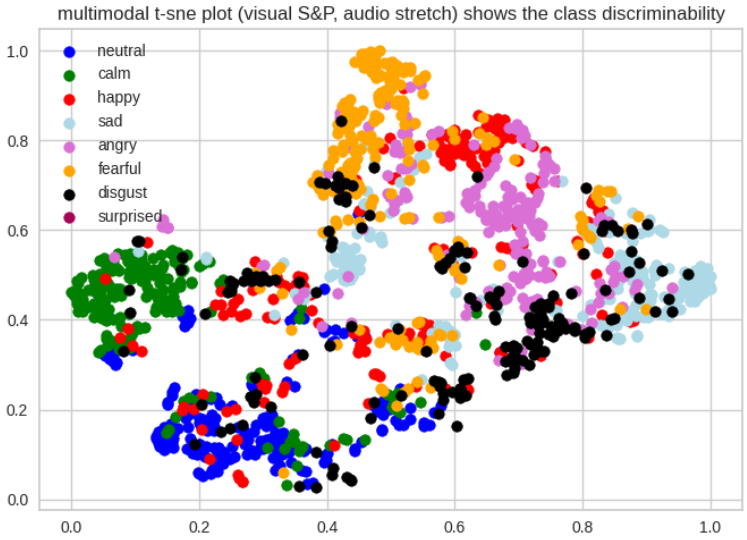

Figure D.22: T-sne for AVER with augmentations visual - Salt & Pepper Noise and audio - Stretch.

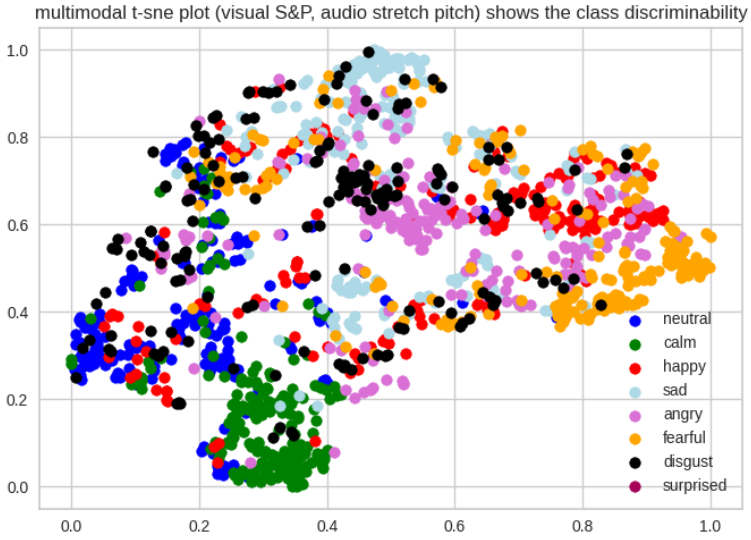

Figure D.23: T-sne for AVER with augmentations visual - Salt & Pepper Noise and audio - Stretch Pitch.

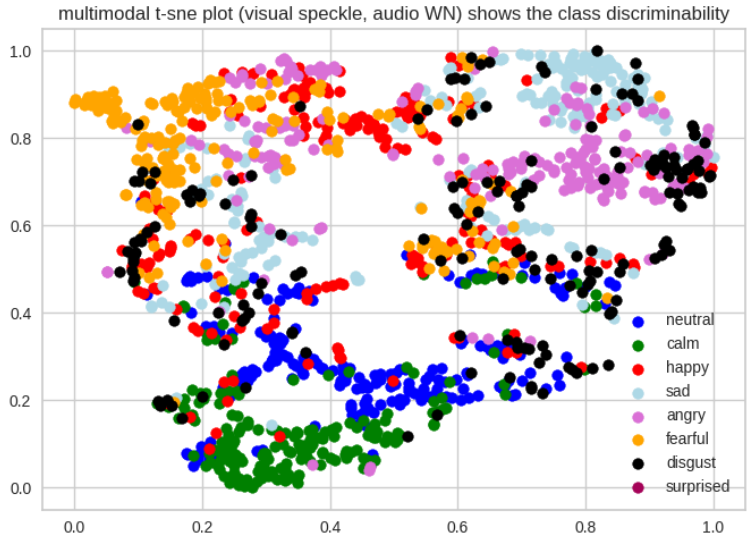

Figure D.24: T-sne for AVER with augmentations visual - Speckle Noise and audio - White Noise.

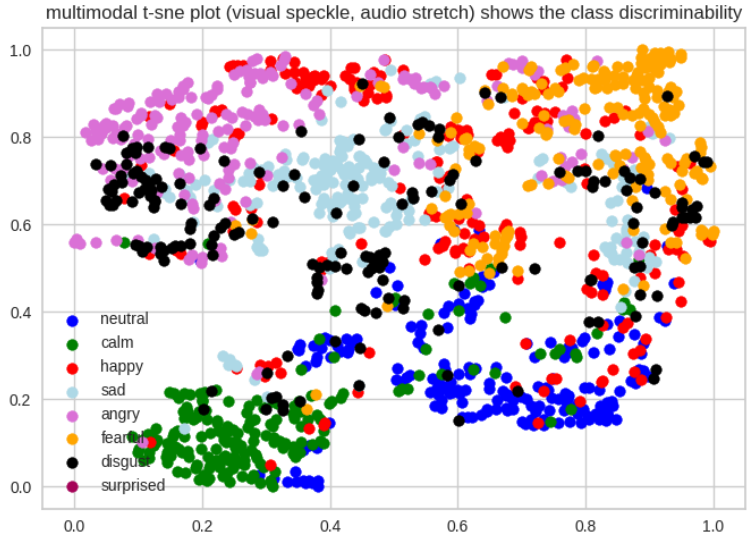

Figure D.25: T-sne for AVER with augmentations visual - Speckle Noise and audio - Stretch.

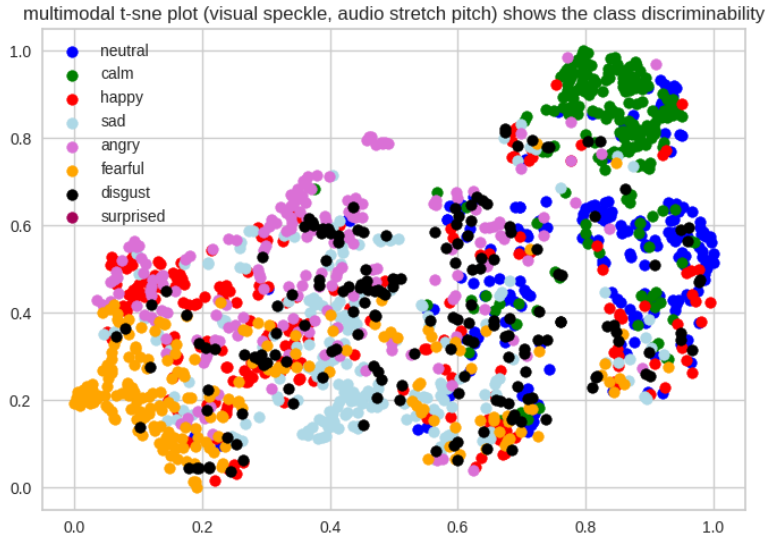

Figure D.26: T-sne for AVER with augmentations visual - Speckle Noise and audio - Stretch Pitch.

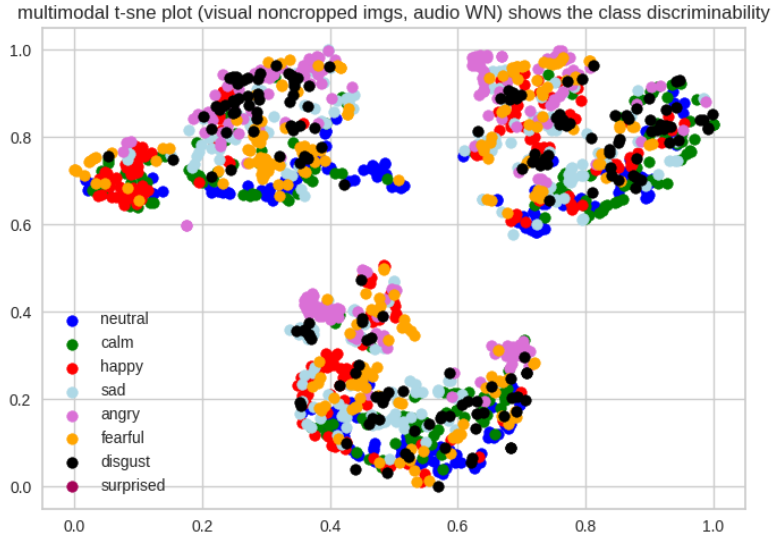

Figure D.27: T-sne for AVER with augmentations visual - non-cropped frames and audio - White Noise.

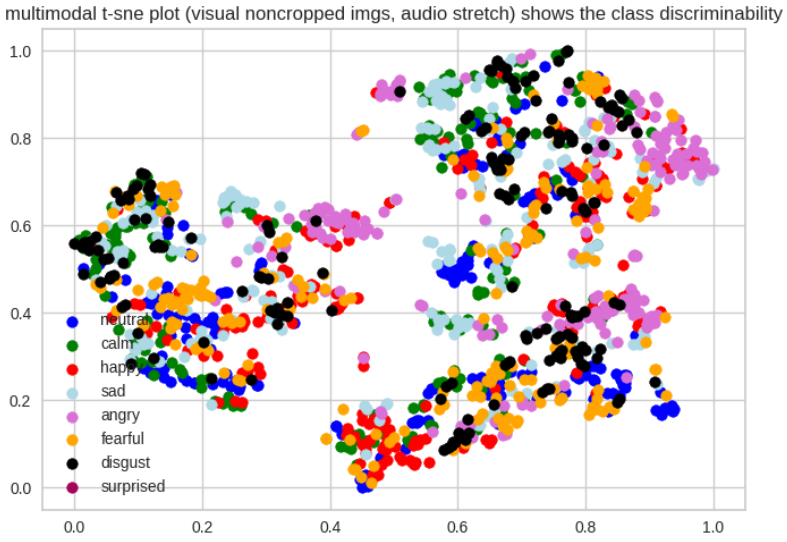

Figure D.28: T-sne for AVER with augmentations visual - non-cropped frames and audio - Stretch.

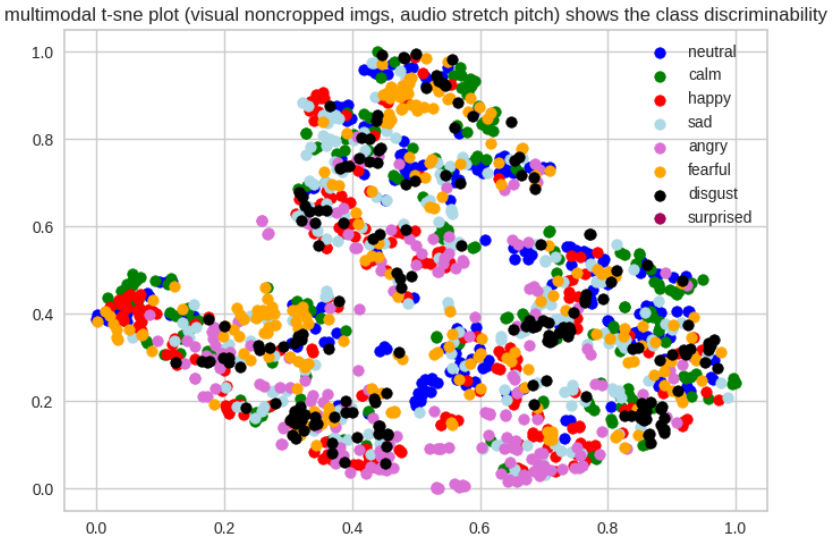

Figure D.29: T-sne for AVER with augmentations visual - non-cropped frames and audio - Stretch Pitch.

Bibliography

[1] Dinesh Acharya. *CovPoolFER*. https://github.com/d-acharya/CovPoolFER. Sept. 2018.

[2] Dinesh Acharya et al. 'Covariance pooling for facial expression recognition'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 367–374.

[3] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou and Ioannis Giannoukos. 'Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011'. In: *Artificial Intelligence Review* 43.2 (2015), pp. 155–177.

[4] Tadas Baltrušaitis, Chaitanya Ahuja and Louis-Philippe Morency. 'Multimodal machine learning: A survey and taxonomy'. In: *IEEE transactions on pattern analysis and machine intelligence* 41.2 (2018), pp. 423–443.

[5] Benjamin Bengfort et al. *Yellowbrick*. Version 0.9.1. 14th Nov. 2018. DOI: 10.5281/zenodo.1206264. URL: http://www.scikit-yb.org/en/latest/.

[6] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[7] G. Bradski. 'The OpenCV Library'. In: *Dr. Dobb's Journal of Software Tools* (2000).

[8] Joseph P Campbell, Douglas A Reynolds and Robert B Dunn. 'Fusing high-and low-level features for speaker recognition'. In: *Eighth European Conference on Speech Communication and Technology*. 2003.

[9] Houwei Cao et al. 'CREMA-D: Crowd-sourced emotional multimodal actors dataset'. In: *IEEE transactions on affective computing* 5.4 (2014), pp. 377–390.

[10] David Cooper Cheyney. *CREMA-D (Crowd-sourced Emotional Mutimodal Actors Dataset)*. https://github.com/CheyneyComputerScience/CREMA-D. June 2018.

[11] R. Collobert, K. Kavukcuoglu and C. Farabet. 'Torch7: A Matlab-like Environment for Machine Learning'. In: *BigLearn, NIPS Workshop*. 2011.

[12] Navneet Dalal and Bill Triggs. 'Histograms of oriented gradients for human detection'. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE. 2005, pp. 886–893.

[13] Mehmet Dedeoglu, Jessica Zhang and Runli Liang. 'Emotion Classification Based on Audiovisual Information Fusion Using Deep Learning'. In: *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2019, pp. 131–134.

[14] Mehmet Dedeoglu, Jessica Zhang and Runli Liang. 'Emotion Classification Based on Audiovisual Information Fusion Using Deep Learning'. In: *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2019, pp. 131–134.

[15] E.Hüllermeier and W. Waegeman. 'Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction'. In: *arXiv preprint arXiv:1910.09457* (2019).

[16] Paul Ekman. *Emotion in the human face . Cambridge Cambridgeshire*. 1982.

[17] Dan Ellis and Manoj Plakal. *VGGish*. https://github.com/tensorflow/models/tree/master/research/audioset/vggish. June 2019.

[18] Haytham M Fayek, Margaret Lech and Lawrence Cavedon. 'Evaluating deep learning architectures for Speech Emotion Recognition'. In: *Neural Networks* 92 (2017), pp. 60–68.

[19] Zoubin Ghahramani. 'Probabilistic machine learning and artificial intelligence'. In: *Nature* 521.7553 (2015), pp. 452–459.

[20] Esam Ghaleb, Mirela Popa and Stylianos Asteriadis. 'Multimodal and Temporal Perception of Audio-visual Cues for Emotion Recognition'. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2019, pp. 552–558.

[21] Asma Ghandeharioun et al. 'Uncertainty modeling in affective computing'. In: (2019).

[22] Daniel Goleman. *Emotional intelligence. New York, NY, England*. 1995.

[23] Frédéric Gougoux et al. 'A functional neuroimaging study of sound localization: visual cortex activity predicts performance in early-blind individuals'. In: *Plos biol* 3.2 (2005), e27.

[24] Barbara Hammer and Thomas Villmann. 'How to process uncertainty in machine learning?.' In: *ESANN*. Citeseer. 2007, pp. 79–90.

[25] Shawn Hershey et al. 'CNN architectures for large-scale audio classification'. In: *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE. 2017, pp. 131–135.

[26] Ursula Hess and Shlomo Hareli. 'The influence of context on emotion recognition in humans'. In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 3. IEEE. 2015, pp. 1–6.

[27] J. D. Hunter. 'Matplotlib: A 2D graphics environment'. In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[28] Dias Issa, M Fatih Demirci and Adnan Yazici. 'Speech emotion recognition with deep convolutional neural networks'. In: *Biomedical Signal Processing and Control* 59 (2020), p. 101894.

[29] A. K. Jain, R. P. W. Duin and J Mao. 'Statistical pattern recognition: A review'. In: *IEEE Transactions on pattern analysis and machine intelligence* 22.1 (2000), pp. 4–37.

[30] Joel Jogy. *How I Understood: What features to consider while training audio files?* https://towardsdatascience.com/how-i-understood-what-features-to-consider-while-training-audio-files-eedfb6e9002b. Sept. 2019.

[31] Pattern Recognition Journal. *Pattern Recognition*. https://www.journals.elsevier.com/pattern-recognition. 2020.

[32] Proedrou K. et al. 'Transductive confidence machines for pattern recognition'. In: *European Conference on Machine Learning*. Springer. 2002, pp. 381–390.

[33] Aggelos K Katsaggelos, Sara Bahaadini and Rafael Molina. 'Audiovisual fusion: Challenges and new approaches'. In: *Proceedings of the IEEE* 103.9 (2015), pp. 1635–1653.

[34] Hugo van Kemenade et al. *python-pillow/Pillow: 7.2.0*. Version 7.2.0. June 2020. DOI: 10.5281/zenodo.3923759. URL: https://doi.org/10.5281/zenodo.3923759.

[35] Alex Kendall and Yarin Gal. 'What uncertainties do we need in bayesian deep learning for computer vision?' In: *Advances in neural information processing systems*. 2017, pp. 5574–5584.

[36] Agnieszka Landowska. 'Uncertainty in emotion recognition'. In: *Journal of Information, Communication and Ethics in Society* (2019).

[37] Agnieszka Landowska, Grzegorz Brodny and Michal R Wrobel. 'Limitations of Emotion Recognition from Facial Expressions in e-Learning Context.' In: *CSEDU (2)*. 2017, pp. 383–389.

[38] Steven R. Livingstone. *RAVDESS Emotional speech audio*. https://github.com/CheyneyComputerScience/CREMA-D. 2018.

[39] Esma Mansouri Benssassi. 'Bio-inspired multisensory integration of social signals'. doctoral thesis. University of St Andrews, 2020.

[40] Sébastien Marcel and Yann Rodriguez. 'Torchvision the Machine-Vision Package of Torch'. In: *Proceedings of the 18th ACM International Conference on Multimedia*. MM '10. Firenze, Italy: Association for Computing Machinery, 2010, pp. 1485–1488. ISBN: 9781605589336. DOI: 10.1145/1873951.1874254. URL: https://doi.org/10.1145/1873951.1874254.

[41] Brian McFee et al. 'librosa: Audio and Music Signal Analysis in Python'. In: Jan. 2015, pp. 18–24. DOI: 10.25080/Majora-7b98e3ed-003.

[42] F. Michaud et al. 'Artificial emotion and social robotics'. In: *Distributed autonomous robotic systems 4*. Springer, 2000, pp. 121–130.

[43] Soonil Kwon Mustaqeem M. 'A CNN-assisted enhanced audio signal processing for speech emotion recognition'. In: *Sensors* 20.1 (2020), p. 183.

[44] Pratheeksha Nair. *The dummy's guide to MFCC*. https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd. July 2018.

[45] Timo Ojala, Matti Pietikäinen and David Harwood. 'A comparative study of texture measures with classification based on featured distributions'. In: *Pattern recognition* 29.1 (1996), pp. 51–59.

[46] Nicolas Papernot and Patrick McDaniel. 'Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning'. In: *arXiv preprint arXiv:1803.04765* (2018).

[47] F. Pedregosa et al. 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[48] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[49] Gerasimos Potamianos et al. 'Recent advances in the automatic recognition of audiovisual speech'. In: *Proceedings of the IEEE* 91.9 (2003), pp. 1306–1326.

[50] Ithai Rabinowitch and Jihong Bai. 'The foundations of cross-modal plasticity'. In: *Communicative & integrative biology* 9.2 (2016), e1002348.

[51] Nicolae-Cătălin Ristea, Liviu Cristian Duţu and Anamaria Radoi. 'Emotion Recognition System from Speech and Visual Information based on Convolutional Neural Networks'. In: *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE. 2019, pp. 1–6.

[52] Sourav Sahoo et al. 'A Segment Level Approach to Speech Emotion Recognition Using Transfer Learning'. In: *Asian Conference on Pattern Recognition*. Springer. 2019, pp. 435–448.

[53] Peter Salovey and John D Mayer. 'Emotional intelligence'. In: *Imagination, cognition and personality* 9.3 (1990), pp. 185–211.

[54] Nicu Sebe, Ira Cohen and Thomas SHuang. 'Multimodal emotion recognition'. In: *Handbook of Pattern Recognition and Computer Vision*. World Scientific, 2005, pp. 387–409.

[55] Mohammad Faridul Haque Siddiqui and Ahmad Y Javaid. 'A Multimodal Facial Emotion Recognition Framework through the Fusion of Speech with Visible and Infrared Images'. In: *Multimodal Technologies and Interaction* 4.3 (2020), p. 46.

[56] Kristen M Smith et al. 'Morphometric differences in the Heschl's gyrus of hearing impaired and normal hearing infants'. In: *Cerebral Cortex* 21.5 (2011), pp. 991–998.

[57] Cees GM Snoek, Marcel Worring and Arnold WM Smeulders. 'Early versus late fusion in semantic video analysis'. In: *Proceedings of the 13th annual ACM international conference on Multimedia*. 2005, pp. 399–402.

[58] TensorFlow. *crema$_d$*. https://www.tensorflow.org/datasets/catalog/crema_d. July 2020.

[59] P. Tzirakis et al. 'End-to-end multimodal emotion recognition using deep neural networks'. In: *IEEE Journal of Selected Topics in Signal Processing* 11.8 (2017), pp. 1301–1309.

[60] Tejal Udhan and Shonda Bernadin. 'Speaker-dependent low-level acoustic feature extraction for emotion recognition'. In: *The Journal of the Acoustical Society of America* 143.3 (2018), pp. 1747–1747.

[61] Stéfan van der Walt et al. 'scikit-image: image processing in Python'. In: *PeerJ* 2 (June 2014), e453. ISSN: 2167-8359. DOI: 10.7717/peerj.453. URL: https://doi.org/10.7717/peerj.453.

[62] Bo Wei et al. 'From real to complex: Enhancing radio-based activity recognition using complex-valued CSI'. In: *ACM Transactions on Sensor Networks (TOSN)* 15.3 (2019), pp. 1–32.

[63] Dongrui Wu. 'Fuzzy sets and systems in building closed-loop affective computing systems for human-computer interaction: Advances and new research directions'. In: *2012 IEEE International Conference on Fuzzy Systems*. IEEE. 2012, pp. 1–8.

[64] Jianhua Zhang et al. 'Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review'. In: *Information Fusion* 59 (2020), pp. 103–126.