



DEGREE PROJECT IN COMPUTER SCIENCE AND ENGINEERING,  
SECOND CYCLE, 30 CREDITS  
*STOCKHOLM, SWEDEN 2020*

# **Automated Multimodal Emotion Recognition**

**MARCOS FERNÁNDEZ CARBONELL**

**KTH ROYAL INSTITUTE OF TECHNOLOGY  
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE**



# **Automated Multimodal Emotion Recognition**

MARCOS FERNÁNDEZ CARBONELL

Master's Programme in Computer Science and Engineering

Date: August 31, 2020

Supervisor: Magnus Boman

Examiner: Jonas Beskow

School of Electrical Engineering and Computer Science

Host company: Stockholm University

Swedish title: Automatiserad multimodal känsligenkänning



## Abstract

Being able to read and interpret affective states plays a significant role in human society. However, this is difficult in some situations, especially when information is limited to either vocal or visual cues. Many researchers have investigated the so-called basic emotions in a supervised way. This thesis holds the results of a multimodal supervised and unsupervised study of a more realistic number of emotions. To that end, audio and video features are extracted from the GEMEP dataset employing openSMILE and OpenFace, respectively. The supervised approach includes the comparison of multiple solutions and proves that multimodal pipelines can outperform unimodal ones, even with a higher number of affective states. The unsupervised approach embraces a traditional and an exploratory method to find meaningful patterns in the multimodal dataset. It also contains an innovative procedure to better understand the output of clustering techniques.

**Keywords**— Multimodal Machine Learning, Emotion Recognition, Supervised Learning, Unsupervised Learning

## Sammanfattning

Att kunna läsa och tolka affektiva tillstånd spelar en viktig roll i det mänskliga samhället. Detta är emellertid svårt i vissa situationer, särskilt när information är begränsad till antingen vokala eller visuella signaler. Många forskare har undersökt de så kallade grundläggande känslorna på ett övervakat sätt. Det här examensarbetet innehåller resultaten från en multimodal övervakad och oövervakad studie av ett mer realistiskt antal känslor. För detta ändamål extraheras ljud- och videoegenskaper från GEMEP-data med openSMILE respektive OpenFace. Det övervakade tillvägagångssättet inkluderar jämförelse av flera lösningar och visar att multimodala pipelines kan överträffa unimodala sådana, även med ett större antal affektiva tillstånd. Den oövervakade metoden omfattar en konservativ och en utforskande metod för att hitta meningsfulla mönster i det multimodala datat. Den innehåller också ett innovativt förfarande för att bättre förstå resultatet av klusteringstekniker.

*Nyckelord*— Multimodal Maskininlärning, Känsligenkänning, Övervakad Inlärning, Oövervakad Inlärning

## Acknowledgments

I wish to thank my project leader, Petri Laukka, and my supervisor, Magnus Boman, for giving me the opportunity to be part of this fascinating project. They, together with Abubakr Karali, have guided me through the whole process, sharing their knowledge and expertise. I thank Prof. José Manuel Mossi García for advising on how to avoid image artifacts in our future recordings. I also wish to thank Blanca for her willingness and unconditional support. Lastly, thank you to my family for their endless love and support.





# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background . . . . .	1
1.2	Problem . . . . .	3
1.3	Purpose . . . . .	3
1.4	Benefits, Ethics, and Sustainability . . . . .	3
1.5	Research Methodology . . . . .	4
1.6	Thesis Outline . . . . .	4
<b>2</b>	<b>Extended Background</b>	<b>5</b>
2.1	Related Work . . . . .	5
2.2	Multimodality . . . . .	6
2.2.1	Audio Modality . . . . .	6
2.2.2	Video Modality . . . . .	7
2.2.3	Fusion Techniques . . . . .	8
2.3	Supervised Learning . . . . .	11
2.3.1	Elastic Net Regularization . . . . .	11
2.4	Unsupervised Learning . . . . .	12
2.4.1	Determining the Number of Clusters . . . . .	12
2.4.2	Dimensionality Reduction Methods . . . . .	14
<b>3</b>	<b>Methodology</b>	<b>15</b>
3.1	Research Method . . . . .	15
3.2	Method Evaluation . . . . .	15
3.3	Environment Setup . . . . .	17
3.4	Dataset . . . . .	17
3.4.1	Video Processing . . . . .	18
3.4.2	Audio Processing . . . . .	21
3.5	Implementation . . . . .	21
3.5.1	Supervised Learning . . . . .	22

3.5.2	Unsupervised Learning . . . . .	23
<b>4</b>	<b>Results</b>	<b>29</b>
4.1	Supervised Learning . . . . .	29
4.1.1	Unimodality . . . . .	29
4.1.2	Multimodality . . . . .	32
4.1.3	General Results . . . . .	35
4.2	Unsupervised Learning . . . . .	36
4.2.1	Traditional Approach . . . . .	36
4.2.2	Exploratory Approach . . . . .	39
<b>5</b>	<b>Discussion</b>	<b>47</b>
5.1	Supervised Learning . . . . .	47
5.2	Unsupervised Learning . . . . .	48
<b>6</b>	<b>Conclusions</b>	<b>50</b>
	<b>References</b>	<b>51</b>
<b>A</b>	<b>Supervised Learning</b>	<b>61</b>
A.1	Elapsed Modeling Time . . . . .	61
A.2	Feature Contributions . . . . .	63
<b>B</b>	<b>Unsupervised Learning</b>	<b>66</b>
B.1	Traditional Approach . . . . .	66

# Chapter 1

## Introduction

Whenever people communicate, they reveal emotions through their facial expressions, body gestures, and voice tone. Recognition of such nonverbal communication techniques is crucial for social interaction [1]; and has applications in many fields varying from psychotherapy [2] to human-computer interaction [3]. However, mastering this skill can sometimes be troublesome for some people, especially when information is limited to either verbal or nonverbal communication; or when distinct channels provide conflicting information [4] [5]. Something similar happens when it comes to recognizing emotions using Artificial Intelligence (AI) algorithms. Multimodal approaches (those which combine multiple inputs, e.g., audio, video, text) tend to outperform unimodal solutions [6] [7] [8]. Objective properties of expressions can be measured in many different ways (e.g., facial gestures, body movements, speech prosody, brain imaging) as they unfold over time. But do they have any multimodal cue patterns? This is one of the main questions contemplated within the interdisciplinary five-year research project led by Petri Laukka at the Psychology Department of Stockholm University. This thesis serves as part of a feasibility study before new videos are recorded for a novel database that will be employed in Laukka's research project.

### 1.1 Background

One of the first research efforts conducted in Multimodal Machine Learning (MML) was in 1989 and focused on audio-visual speech recognition [9] propelled by the McGurk effect [10] – an interaction between hearing and vision in speech perception. The phenomenon happens when the auditory component of one sound is combined with the visual component of another sound,

resulting in the perception of a third sound (e.g., a voiced */ba-ba/* with a visual */ga-ga/* is perceived as */da-da/* by most individuals). These results pushed many researchers in the field to investigate multimodal approaches. The vast majority of the unimodal speech recognition solutions were based on Hidden Markov Models (HMMs) back then [11] [12], and so were the early multimodal ones [13]. Even though the original purpose of using two different modalities was to improve speech recognition performance, the experimental results proved that the main benefit of using visual content was when the speech signal had a low signal-to-noise ratio [14] [15].

The study of human multimodal behaviors during social interaction was established in the early 2000s [16]. But it was not until the 2010s that the fields of emotion recognition and affective computing bloomed, thanks to technical improvements in automatic face detection, facial landmark detection, and facial expression recognition [17]. A meta-analysis conducted by D’Mello et al. in 2015 [7] disclosed that multimodal affect recognition leads to improvement when using more than one modality, nevertheless, this betterment is reduced when it comes to spontaneous emotions. Moreover, it has been shown that studies based on posed emotions are prone to outperform those solutions that use natural and induced datasets, where induced means it has been created by exposing subjects to a stimulus (e.g., watching a video) in a controlled environment [18].

Even though most of the affective applications have been unimodal, multimodal approaches have been investigated by countless researchers in the last few years [19] [20] [21] [22]. The principal modalities used in such applications are facial expression estimation, speech prosody (tone) analysis, physiological signal interpretation, and body gesture analysis [18]. A growing body of literature has evaluated the use of these cues in a supervised manner [23] [24] [25], yet few researchers have addressed the problem in an unsupervised way [26]. Besides, only a few studies have addressed the issue of classifying a wide variety of emotions for both facial [27] and vocal [28] expressions, as well as incorporating multimodal expressions [29] [30]; distancing their work from the commonly investigated basic emotions (anger, disgust, fear, happiness, sadness, and surprise).

Much work on the potential of emotion AI has been carried out, however, according to the 2019 Gartner hype cycle, this emergent technology will significantly impact on business, society, and people over the next five to ten years [31].

## 1.2 Problem

Past research on emotion recognition has tended to focus on supervised approaches rather than unsupervised solutions. Furthermore, most studies have relied on datasets with a limited number of so-called basic emotions, although everyday interactions are characterized by a wide variety of more subtle affective states. Hence, it remains unknown whether studying a broader number of emotions using a multi-solution strategy could lead to novel answers in the field. Therefore, the core problem of this thesis is *the apparent standardization of supervised strategies together with the conventional use of rather small emotion datasets*.

## 1.3 Purpose

The main goals of this thesis are to confirm that even with a higher number of emotions, a multimodal supervised machine learning solution defeats a unimodal one, and to see if an inductive data-driven tactic can steer toward unprecedented discoveries. Thus, the two research questions could be stated as follows:

- *Do supervised multimodal strategies outperform unimodal ones, even with a more realistic number of emotions?*
- *Can unsupervised algorithms reveal any meaningful structures in the multimodal dataset?*

## 1.4 Benefits, Ethics, and Sustainability

This thesis serves as a contribution to Laukka's research project as it represents an initial step toward being able to fully answer one of the big three research questions of the project, and it is part of a feasibility study. In the same way, it also contributes to the research community since some of the solutions and results could be used, either directly or indirectly, in many different multimodal scenarios.

Moreover, this work could end up being useful in areas such as healthcare, supporting people that can not properly read emotional cues (because of brain damage or disease); education, adjusting teaching methodologies according to the learners' affective states; entertainment, adapting game sounds in harmony

with the players' emotions; and industry, adjusting self-driving modes depending on the drivers' mood. Hence, this project may eventually play a part in some of the 2030 Sustainable Development Goals (SDGs)<sup>1</sup>. For instance, Goal 4: Quality Education, assuring non-violent and non-discriminative environments and helping learners with disabilities; and Goal 11: Sustainable Cities and Communities, improving road safety and inclusion of minority groups.

The results obtained in this thesis are reproducible and can be replicated by running the proper Jupyter<sup>2</sup> notebook available on the following GitHub repository<sup>3</sup>. Furthermore, all employed methods are compared using statistical measurements to be able to draw solid conclusions.

Bear in mind that this study has been carried out using a dataset with portrayed emotions, therefore, results may change depending on the employed dataset. No videos were recorded and no personal information was used in this investigation.

## 1.5 Research Methodology

In order to answer the above-stated research questions, the following quantitative research method was used. First, video and audio features were extracted from a multimodal dataset with a wide variety of acted emotions. Second, different unimodal and multimodal approaches were compared in terms of their performance. Lastly, two unsupervised solutions were utilized to find meaningful patterns in the multimodal dataset. Refer to Chapter 3 for further information.

## 1.6 Thesis Outline

This thesis is laid out as follows. Chapter 2 provides supplementary information about multimodality, audio and video features, and some of the employed machine learning algorithms. Chapter 3 presents a more exhaustive description of the methodology along with the deployed supervised and unsupervised solutions. Chapter 4 discloses the obtained results in a tabular and graphical way. Chapter 5 discusses the results presented in the previous section. Chapter 6 draws general conclusions and announces recommendations regarding the future directions of Laukka's project.

---

<sup>1</sup><https://www.undp.org/content/undp/en/home/sustainable-development-goals.html>

<sup>2</sup><https://jupyter.org/>

<sup>3</sup><https://github.com/marferca/automated-multimodal-emotion-recognition>

# Chapter 2

## Extended Background

### 2.1 Related Work

This thesis is a pre-pilot study before a more complicated research activity is started. Laukka’s project will involve the creation of a very structured dataset with a large number of emotions and information from different modalities. The GEMEP dataset [29] shares this idea and has been employed by many groups over the years [32] [33]. The openSMILE [34] and OpenFace [35] toolkits have been commonly utilized in the AVEC (Audio-Visual Emotion recognition Challenge) workshops to extract audio and video features [36] [37] [38] [39]. Future work of the project will be not only inspired by studies focused on audio and video features but also by what other groups have more recently attempted with other signals [40] [41] [42].

There are several examples of multimodal emotion recognition systems in the literature. These can be divided into traditional machine learning solutions, where features are “manually” extracted from the data and then input into a predictor [6]; and deep learning solutions, where the raw signals are input directly into the network so that it automatically extracts an intermediate representation of the input data [43]. Kessous et al. [6] proposed a multimodal system to recognize 8 different emotions based on the traditional solution. Facial expressions, body gestures, and acoustic analysis of speech were used to extract features, fuse them in a feature-level fashion, and then use them as input to a Bayesian classifier (BayesNet). With this configuration, they achieved an overall performance of 78.3%, reaching the highest accuracy for anger (96.7%) and the lowest for despair (53.3%). On the other hand, Tzirakis et al. [43] opted for the deep learning approach, using raw signals from the auditory and visual modalities as input data. In order to extract robust fea-

tures from the raw data, they utilized a convolutional neural network and a deep residual network of 50 layers (ResNet-50) [44] for the speech and the visual channels, respectively. The output of these two networks was then fused and fed to an LSTM to predict the affective states. Their system outperformed, in terms of the concordance correlation coefficient, traditional solutions based on handcrafted features (e.g., eGeMAPS [45] and the ones used in the ComParE challenges [46]) on the RECOLA database [47] of the AVEC'16 challenge. Their solution achieved the performance of 0.789 for arousal and 0.691 for valence.

## 2.2 Multimodality

The aggregation of multiple affective cues not only affords a more extensive compilation of data but also helps relieve the uncertainty effects of raw signals. In the end, these signals are acquired by imperfect sensors and frequently processed to extract features from them. Additionally, multimodality brings flexibility and robustness as it gives the possibility of using other modalities if one or more channels are suffering from a lack of meaningful information (e.g., visual occlusion or no aural content) [18]. In 1967, Albert Mehrabian, a body language pioneer researcher, discovered that 7% of the communication is verbal, 38% is vocal, and 55% is visual [48]. This explains why the bulk of studies use audiovisual content to study how emotion expressions are produced. More details regarding audio and video modalities and fusion techniques are given in the following subsections.

### 2.2.1 Audio Modality

Speech contains two interrelated channels of information: linguistic information that reflects the semantics of the message and paralinguistic information transmitted by prosody [18] – an oft-cited description from the book of Titze [49] is; “the non-lexical patterns of tune, rhythm, and timbre in speech; modulated by the implements of human vocal control: air pressure from the lungs, tension in the vocal cords and filtration through the throat, tongue, palate, cheeks, lips and nasal passages.”

As was reported earlier, verbal communication only constitutes a very small part of human communication. Also, the linguistic speech channel has some other drawbacks. First, it is not universal, hence a different natural language speech processor has to be developed for each dialect; second, it is sensitive to concealing since people are not always sincere about their feelings. On



the other hand, the paralinguistic channel seems to be a more informative and reliable source of information. As mentioned by Eyben et al. [45], “the underlying theoretical assumption is that affective processes differentially change autonomic arousal and the tension of the striate musculature and thereby affect voice and speech production on the phonatory and articulatory level and that these changes can be estimated by different parameters of the acoustic waveform [50].”

With the increasing interest in vocal expression, many researchers from different fields have been extracting a wide range of paralinguistic audio features, thereby hindering the comparison of methods and results across studies. To tackle this problem Eyben et al. [45] proposed two standard acoustic parameter sets for various areas of automatic voice analysis, such as paralinguistic or clinical speech analysis. The minimalistic version (GeMAPS) contains prosodic, excitation, vocal tract, and spectral descriptors, whereas the extended variant (eGeMAPS) adds a small set of cepstral descriptors (features based on the inverse Fourier transform of the log-magnitude of a signal spectrum). More details on this topic can be found in [45]. Both parameters sets can be extracted using the open-source toolkit openSMILE [34].

These features are used together with machine learning algorithms to predict affective states. According to a recent survey on AI-based multimodal methods for emotion detection [22], some of the most commonly employed classification methods are Artificial Neural Networks (ANNs), k-Nearest Neighbor (k-NN), Support Vector Machines (SVMs), and Decision Trees. The focus of the ML algorithms was put on classifiers since, to a large extent, this thesis centers on classifying emotions.

### 2.2.2 Video Modality

The visual modality is a rich source of information since it includes different types of non-verbal indicators such as body posture, gestures, facial expressions, eye gaze, etc. But, body movements and facial expressions are the most studied ones. Among both, the latter prevail as one of the most direct channels to transmit human emotions in non-verbal communication [51].

Darwin [52] was one of the first researchers to study the field of emotional expressions. In his work, he stated that facial expressions of emotion are universal, specific movements announce a particular emotion, and emotions should be seen as discrete entities and not unique to humans. Leaning on Darwin’s universality concept, Ekman and Friesen [53] defined the Facial Action Coding System (FACS), a method that objectively describes all visually

possible facial muscle activations. To do so, they divided facial expressions into separated components of muscle movements, called Action Units (AUs) (see Table 2.1).

There are some commercial tools for AU recognition, such as AFFDEX [54], Noldus FaceReader [55], and OKAO [56]. However, these solutions tend to be prohibitively costly and a bit of a black box, since they neither disclose the employed training data and algorithms, nor give open access to benchmarks. Fortunately, there are a few freely alternative toolkits that allow extracting these features, e.g., TAUD [57] and OpenFace [35]. The latter is not only the most recent free toolkit for such a feature extraction task but also utilizes computer vision algorithms that achieve state-of-the-art results on landmark detection, head pose estimation, facial action unit recognition, and eye gaze estimation (see Figure 2.1).

These video features have been commonly used as input data to ANNs, SVMs, and HMMs for classification tasks [18]. However, due to the fast-paced evolution of chip processing abilities, video frames have also been employed as raw input data to deep learning models [58].

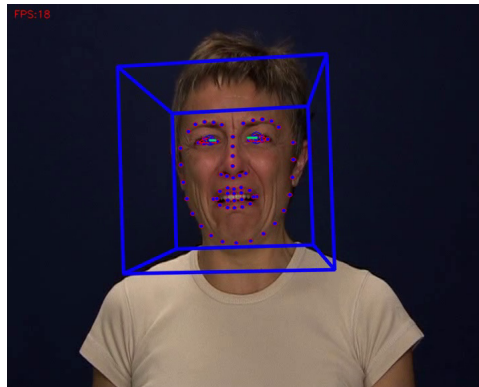


Figure 2.1: OpenFace 2.2.0 tracking results of a GEMEP [29] video portraying disgust. Points represent landmark detection, green lines the estimated eye gaze vectors, and the blue cube a 3D bounding box of the head.

### 2.2.3 Fusion Techniques

One of the key uses of multimodality is to fuse cues from two or more modalities to predict a result from a richer source of information. This topic has been widely studied by many researchers in the field, provoking different fusion technique classifications. Nonetheless, the majority of previous surveys [59] [7] [18] stick to the following grouping.


















AU	Description	Example
AU1	Inner brow raiser	
AU2	Outer brow raiser	
AU4	Brow lowerer	
AU5	Upper lid raiser	
AU6	Cheek raiser	
AU7	Lid tightener	
AU9	Nose wrinkler	
AU10	Upper lip raiser	
AU12	Lip corner puller	
AU14	Dimpler	
AU15	Lip corner depressor	
AU17	Chin raiser	
AU20	Lip stretched	
AU23	Lip tightener	
AU25	Lips part	
AU26	Jaw drop	
AU45	Blink	

Table 2.1: AU examples (adapted from [35], p. 63).

### Early Fusion

The easiest and most common solution lies in concatenating the feature vectors from each modality and then use them as input data to the predictor [60] [61]. However, there are more elaborate approaches where different algorithms are applied to discard the least significant features [19] [6].

Using this fusion technique provides some challenges. First, a larger number of features may lead to a lower accuracy if the training set is not large enough [62]. Second, there may be time resolution and format incompatibilities between fused signals. Finally, a higher-dimensional space increases the computational load.

### Late Fusion

This approach inputs the features from each modality into a distinct predictor and then merges their output into a final result. In this survey [63], Corneanu et al. grouped the most typical late fusion strategies for emotion recognition into the following categories (Figure 2.2 presents a practical example of how these fusion techniques work):

- Maximum rule: selects the maximum of all posterior probabilities.
- Sum rule: sums probabilities from each classifier and then picks the class with the highest value.
- Product rule: multiplies probabilities between classifiers and then chooses the class with the largest value.
- Weight criterion: results in a linear combination of the classifiers' output, where the constants are confidence rates of the predictors.
- Rule-based: selects a dominant modality for each class.
- Model-based: employs a machine-learning algorithm to fuse the output of the classifiers.

One of the biggest downsides of using this fusion tactic is the information loss of cross-modal correlation since using an individual model for each modality ignores the low-level interaction between modalities [59] [16].

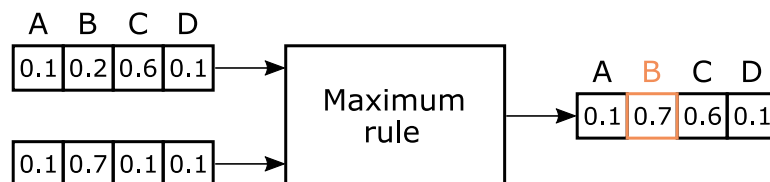


Figure 2.2: Maximum rule late fusion example: Imagine a scenario where information from two modalities is given to classify four classes (A, B, C, and D). Since it is a late fusion approach, two different classifiers are trained for each modality. Thus two prediction vectors are expected to input into the maximum rule fusion system. As illustrated in the figure, the system will return the maximum value per class and select the class with the highest value as the final decision. Note that the final output could be normalized, so that the total probability equals one.

## Hybrid Fusion

This solution attempts to bring the best of both fusion techniques by combining outputs from early fusion with individual unimodal predictors. However, this approach only makes sense when more than two modalities are utilized. For example, imagine a scenario in which three modalities are given: audio, video, and MRI (Magnetic Resonance Imaging) cues. Audio and video features could be concatenated and employed to train a classifier (early fusion). Simultaneously, another predictor could be used for MRI, to finally fuse the output from both classifiers (late fusion).

## 2.3 Supervised Learning

### 2.3.1 Elastic Net Regularization

In regression analysis, the goal is to find a good regression function (Equation 2.1) that minimizes a predefined loss function, such as the ordinary least squares (Equation 2.2).

$$\hat{f}(\mathbf{x}) = \mathbf{x}^T \hat{\boldsymbol{\beta}} \quad (2.1)$$

$$L_{OLS}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^n (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})^2 \quad (2.2)$$

However, if the values of  $\hat{\boldsymbol{\beta}}$  are unconstrained, they can grow rapidly and hence be vulnerable to high variance. To reduce this problem,  $\hat{\boldsymbol{\beta}}$  coefficients can be regularized imposing lasso ( $\ell_1$ -penalty) or ridge constraints ( $\ell_2$ -penalty). Their loss functions can be expressed as follows:

$$L_{\ell_1}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^n (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})^2 + \lambda \sum_{j=1}^p |\hat{\beta}_j| \quad (2.3)$$

$$L_{\ell_2}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^n (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})^2 + \lambda \sum_{j=1}^p \hat{\beta}_j^2 \quad (2.4)$$

On the one hand, lasso regression works best when the model contains several superfluous variables, creating an easier model to interpret. On the other hand, ridge regression performs best when most of the variables are useful. Therefore, it might be easy to choose between both methods when having a

vast knowledge of variables. Nevertheless, when dealing with a large number of variables coming from different modalities, this ends up being more complicated.

The elastic net regularization is a convex combination of both methods and has demonstrated superiority over the lasso [64]. Just like lasso and ridge regularization, its loss function (Equation 2.5) starts with least squares and then adds  $\ell_1$  and  $\ell_2$  penalties together with a ratio-variable that lets you choose between lasso ( $\alpha = 1$ ), ridge regularization ( $\alpha = 0$ ), or a combination of both ( $0 < \alpha < 1$ ).

$$L_{enet}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^n (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})^2 + \alpha \lambda \sum_{j=1}^p |\hat{\beta}_j| + (1 - \alpha) \lambda \sum_{j=1}^p \hat{\beta}_j^2 \quad (2.5)$$

## 2.4 Unsupervised Learning

### 2.4.1 Determining the Number of Clusters

Determining the number of clusters in a high-dimensional space is a hard task for humans. For this reason, some researchers have published different statistical techniques to tackle this problem.

#### CH Index

The CH index metric [65] rests upon two fundamental concepts:

- Between-Group Sum of Squares (BGSS): measures how separated the groups are from each other.
- Within-Group Sum of Squares (WGSS): measures how strongly grouped the clusters are.

When talking about BGSS, a greater value is better. But the problem is that when running a clustering algorithm for a group of different  $k$  values, and plotting the  $BGSS(k)$ ; the between-cluster variation keeps increasing. Something similar happens when it comes to WGSS. In this case, a lower value is better, but the within-cluster variation keeps decreasing as  $k$  increases. Therefore, the perfect value of  $k$  would be the one that simultaneously achieves a large BGSS and a small WGSS. That is the core idea of the CH index (Equation 2.6 and Equation 2.7).

$$CH(k) = \frac{BGSS(k)(n - k)}{WGSS(k)(k - 1)} \quad (2.6)$$

where:

$k$  = number of clusters  
 $n$  = number of instances

$$\hat{k} = \operatorname{argmax}_{k \in \{2, \dots, K_{max}\}} CH(k) \quad (2.7)$$

### Silhouette Score

The average silhouette width (also known as silhouette score<sup>1</sup>) might be used to select a suitable number of clusters [66]. To do so, first, the silhouette coefficient has to be calculated for each sample (Equation 2.8). Second, the average silhouette width is computed (Equation 2.9). Third, these two steps are repeated for different values of  $k$  to obtain  $s_{score}(k)$ . Lastly, the number of clusters is estimated by picking the value of  $k$  that has the largest  $s_{score}$  (Equation 2.10). As indicated by Rousseeuw in [66], the closer to 1 the score is, the better. By contrast, a value close to  $-1$  usually indicates that multiple samples have been assigned to the wrong cluster.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2.8)$$

where:

$a$  = mean intra-cluster distance  
 $b$  = mean nearest-cluster distance

$$s_{score}(i) = \frac{1}{n} \sum_{i=0}^{n-1} s(i) \quad (2.9)$$

$$\hat{k} = \operatorname{argmax}_{k \in \{1, \dots, K_{max}\}} s_{score}(k) \quad (2.10)$$

---

<sup>1</sup>Scikit-learn - Silhouette Score: [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html)

## 2.4.2 Dimensionality Reduction Methods

### PCA

Principal Component Analysis (PCA) is a classic linear method for dimensionality reduction. This technique tries to transform a set of probably correlated features into a new set of  $n < d$  uncorrelated features (where  $d$  is the original number of dimensions). To that end, the algorithm iteratively runs until it finds the  $n$  dimensions that achieve maximum variance, creating a new  $n$ -dimensional space. More details on PCA can be found in [67] and [68].

### t-SNE

The t-distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear dimensionality reduction technique that aims to take a set of points in a high-dimensional space and transform them into a revelatory two- or three-dimensional view. The algorithm can be summarized into two stages. First, each pair of data points in the high-dimensional space is modeled by a probability distribution. Second, a low-dimensional space that attempts to follow the previously obtained probability distributions is created. The main tunable parameter of the method is the perplexity, which in a nutshell, corresponds to the number of neighbors per data point. Please refer to [69] for a more in-depth explanation and to [70] for a more distilled idea of how to use and interpret this method.

### UMAP

Uniform Manifold Approximation and Projection (UMAP) is a non-linear dimensionality reduction technique that aims to create a low dimensional view of a high dimensional space that preserves relevant structure. Its creators claim that their algorithm is competitive with t-SNE and arguably maintains more of the global structure of the data in a more efficient way. The algorithm can be summarized into two steps. First, a fuzzy set topological representation is constructed. Second, this low-dimensional representation is optimized to reduce its fuzziness by minimizing the cross-entropy. Please refer to [71] for a mathematical description of the algorithm and to [72] for a reduced version of how it works.



# Chapter 3

## Methodology

### 3.1 Research Method

In order to study the stated research questions, the following quantitative research method was followed. First, audio and video features were extracted from a set of 1,260 video files and 18 emotions. Second, data was prepared for further use in a supervised and unsupervised approach. Third, the supervised solution evaluated whether multimodal pipelines could defeat unimodal ones even with a large number of emotions. To this end, the best unimodal and multimodal pipelines were compared in terms of their test AUC. These pipelines were composed of classic ML classification algorithms and different fusion techniques. Fourth, the unsupervised solution was employed to find meaningful patterns in the multimodal dataset. This approach lay in the use of a traditional solution, which involved the utilization of two conventional clustering techniques; and an exploratory solution, which entailed using an interactive toolkit. These four stages are explained in detail in the following subsections.

### 3.2 Method Evaluation

The use of typical metrics in skewed datasets can lead to suboptimal classification models and might conduce to deceiving conclusions. The area under the ROC curve (AUC) [73] is one of the most popular tools used in this type of scenario [74]. Besides, rank-based metrics, such as AUC, play an important role in applications where good class separation is needed [75].

The ROC curve is a diagnostic chart that summarizes the behavior of a model by computing the false positive rate (FPR) and the true positive rate

(TPR) for a set of given scores under distinct thresholds. The FPR and TPR can be calculated as in Equation 3.1 and Equation 3.2 respectively.

$$FPR = \frac{FP}{FP + TN} \quad (3.1)$$

$$TPR = \frac{TP}{TP + FN} \quad (3.2)$$

The AUC provides a summarized version of the ROC curve by just giving a single score, which makes it perfect for model evaluation. A value of 0.5 corresponds to a no skill classifier, whereas a value of 1.0 translates as an ideal classifier.

The dataset was randomly split into training (75%) and test (25%) subsets in a stratified fashion. Furthermore, 5-fold cross-validation was performed over the training set for hyperparameter tuning, obtaining more reliable performance results, and lastly, selecting between classifiers. The subset was divided into five folds, as  $k = 5$  and  $k = 10$  have been shown empirically to yield test error rate estimates that do not suffer from extremely high bias nor from very high variance. [76]. Unfortunately, the dataset was not divided on a per-subject basis since the ten actors neither portray the same emotions nor the same number of files per affective state (see Figure 3.1). Hence, it might be difficult to find a way to split the dataset in a per-subject basis, and even more complicated to preserve this condition when using k-fold cross-validation. Consequently, the fact of not using such a division technique may have resulted in higher classification rates.

Finally, the normalized gain [77] was used to check if the supervised multimodal strategies (early and late fusion) outperform the best unimodal one. This metric serves to normalize gain scores with regard to how much gain could have been achieved (Equation 3.3).

$$g = \frac{AUC_{test}^{mul.}(\%) - AUC_{test}^{uni.}(\%)}{100 - AUC_{test}^{uni.}(\%)} \quad (3.3)$$

where:

$AUC_{test}^{mul.}$  = AUC score over the test dataset for the multimodal solution

$AUC_{test}^{uni.}$  = AUC score over the test dataset for the best unimodal solution

Normalized Gain	Gain Level
$g < 0.30$	Low
$0.30 \leq g \leq 0.70$	Medium
$0.70 < g$	High

Table 3.1: Normalized gain and gain level according to [77].

### 3.3 Environment Setup

This study was mainly run on a MacBook Pro (Retina, 13-inch, Early 2015) with a 2.7 GHz Intel Core i5 processor, 8 GB 1867 MHz DDR3, Intel Iris Graphics 6100 1536 MB, APPLE SSD SM0128G drive, and macOS Mojave 10.14.6 as operative system. Regarding the programming environment, Python 3.7.6 [78], scikit-learn 0.22.1 [79], SciPy 1.4.1 [80], and Matplotlib 3.1.3 [81] were principally used. Besides, openSMILE 2.3.0 [34] and OpenFace 2.2.0 [35] were installed and used for feature extraction.

### 3.4 Dataset

The GEMEP (Geneva Multimodal Emotion Portrayals) Master Set [29] contains 1,260 audio and video files where ten professional actors coached by a professional director perform 18 affective states uttering two different pseudo-linguistic phoneme sequences or a sustained vowel “aaa”). The dataset includes the emotions presented in Table 3.2. These emotion IDs will be used from now on in some of the charts for making them look cleaner. This dataset was chosen among others because it contains a wider variety of both positive and negative emotions than other multimodal datasets [82].

As can be seen in Figure 3.1, the data is imbalanced since most of the portrayed emotions contain 90 files per class, yet others only have 30 records per class. Moreover, actors have not interpreted all the emotions, and in the majority of cases, the number of records per actor within each emotion is not proportionally divided.

Emotion	Emotion ID
Admiration	adm
Amusement	amu
Anger	col
Anxiety (worry)	inq
Contempt	mep
Despair	des
Disgust	deg
Interest	int
Irritation	irr
Joy (elation)	joi
Panic (fear)	peu
Pleasure (sensual)	pla
Pride	fie
Relief	sou
Sadness	tri
Shame	hon
Surprise	sur
Tenderness	att

Table 3.2: Portrayed emotions and their identifiers.

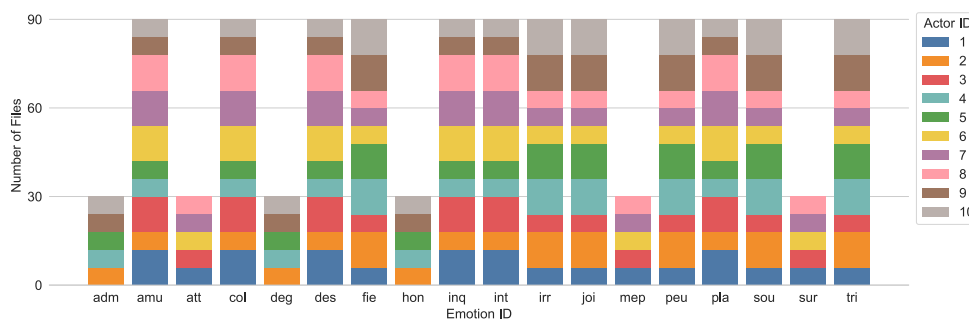


Figure 3.1: Number of files per emotion ID and actor ID.

### 3.4.1 Video Processing

#### Feature Extraction

The OpenFace toolkit was used for feature extraction over the 1,260 video files. The whole process took around an hour and generated the following outputs

per file:

- CSV file: contains basic details about the video feature extraction (e.g., frame number, timestamp, confidence, and success rates), together with information about eye gaze, pose, 2D landmark locations, 3D landmark locations, ridge and non-ridge shape parameters, and AUs.
- Similarity aligned face folder: includes images of the faces detected in the input video.
- HOG binary file: encloses the histogram of oriented gradients.
- Tracking video: displays landmarks, head pose and eye gaze tracking (see Figure 2.1).
- Metafile: includes information about the input and output data.

### **Data Cleaning**

From all the above-enumerated outputs, for this thesis, only data from the CSV file was used. In particular, the columns related to the basic details and the AUs. The confidence and success rates significantly helped with the data cleaning process. The first one tells about how reliable the tracking was (value from zero to one). The latter is used as a flag to inform users of face detection and successful tracking. Regarding the AUs, OpenFace can return the intensity, ranging from zero to five, of 17 facial muscle activations (see Table 2.1).

The video feature extraction toolkit returns multiple feature instances per file (one row per frame). A total number of 76,612 instances during the feature extraction process were created. As shown in Figure 3.2, the average number of instances per file is 50. However, there are a few files with around 200 and others with less than 20.

As illustrated in Figure 3.3 and Figure 3.4, only 0.58% of the instances were flagged as unsuccessful, and around 10% had a confidence rate lower than 0.98. Taking this into consideration, instances with a success value distinct from one or a confidence rate lower than 0.98 were dropped. The number of total instances decreased by 9.94% after the cleaning process and caused the erasure of an entire file.

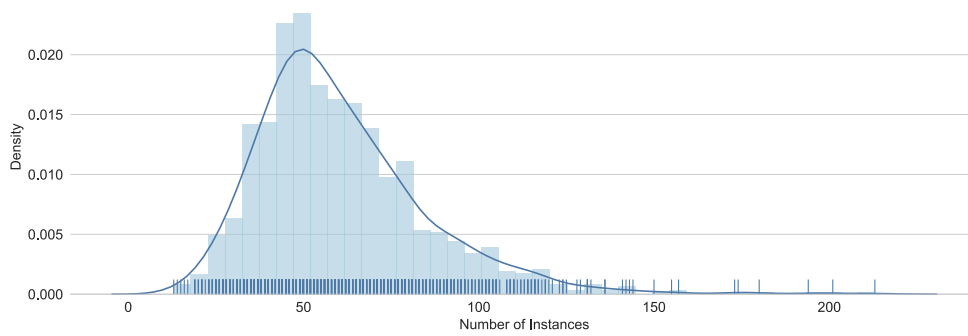


Figure 3.2: Kernel density estimate of the number of video instances per file with histogram and rug plot. The height of the curve is scaled, so that the total density is equal to one.

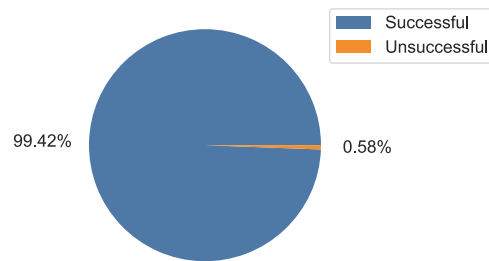


Figure 3.3: Distribution of the success field.

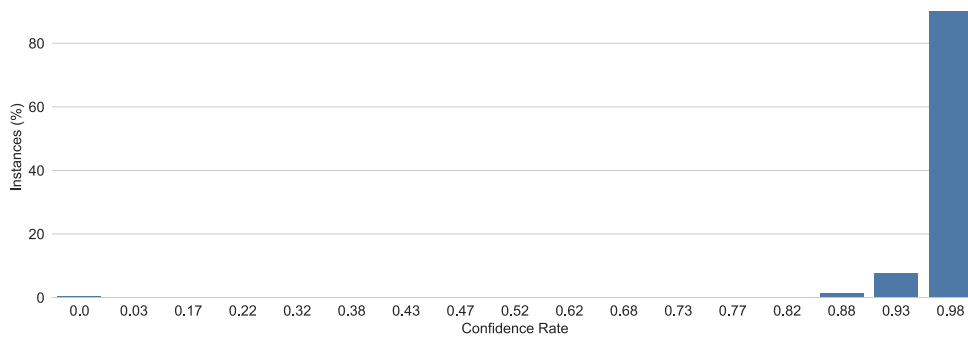


Figure 3.4: Percentage of instances per confidence rate.

### Data Preparation

Unlike the video toolkit, the audio software only returns one instance per file. Therefore, a way to achieve data consistency is needed. To do so, video in-

stances were grouped by file and averaged column-wise, reducing the entire dataset to 1,259 observations. Lastly, features values were divided by 5 for normalization purposes, and training and test subsets were created.

### 3.4.2 Audio Processing

#### Feature Extraction

In this part, two main steps were carried out. First, audio tracks were taken from the database videos by using FFmpeg<sup>1</sup>. Second, openSMILE was used to extract two different parameter sets. These two steps took 15 minutes. The GeMAPS set contains 62 features, while the extended version (eGeMAPS) 88. More details on the nature of the audio features can be found in Subsection 2.2.1 and in [45].

#### Data Cleaning

As was mentioned in Subsection 3.4.1, after the video data cleaning procedure, one file was deleted. Hence, the corresponding audio track had to be dropped from both audio feature sets, reducing that number of instances to 1,259.

#### Data Preparation

Once features were extracted and data was cleaned, both feature sets were separated into training and test sets in a stratified fashion. Lastly, all the parameters were normalized by first fitting a min-max scaler to the training sets and then applying the scales to the training and test sets. Thereby, test instances were adjusted as in a real-life scenario where new input values are modified according to the previously fitted scale, ensuring that new samples are always transformed into a decimal between zero and one.

## 3.5 Implementation

With the completion of the main data processing phase, the information from audio and video modalities was ready to be used for carrying out different solutions. In the following subsections, the implemented supervised and unsupervised approaches are thoroughly explained.

---

<sup>1</sup>FFmpeg software: <https://www.ffmpeg.org/>

### 3.5.1 Supervised Learning

This part of the study aims to answer one of the research questions of the thesis. In particular, whether multimodal supervised strategies can beat unimodal ones, even with a large number of emotions. To do so, different multimodal late and early fusion techniques were evaluated and compared to the best unimodal classifier.

The following multimodal pipelines utilize machine learning models such as Linear Classifiers with Elastic Net regularization, k-NN, Decision Tree, and Random Forest. The first three were used since they are some of the most commonly employed methods for emotion recognition (as was previously said in Subsection 2.2.1), whereas Random Forest was used because it is known as one of the best out-of-the-box classifiers [83]. Note that from now on, in this thesis, Linear Classifiers with Elastic Net regularization will be called Elastic Net classifiers.

#### Late Fusion

This approach can be summarized into three steps. First, audio and video classifiers were separately subjected to a modeling and selection process. Second, different techniques were tested for fusing the outputs of the audio and video classifiers. Third, the best late fusion pipeline was evaluated over the test set and compared to the best unimodal classifier. Note that the best unimodal classifier corresponds to the strongest model (in terms of their validation AUC) picked in the first step.

Next, a more deep explanation of the above-stated stages is given. The first step can be split, in turn, into two phases and was repeated for each modality. First, 5-fold cross-validation was employed for hyperparameter tuning over the training set. Second, once the best parameters for each classifier (Elastic Net, k-NN, Decision Tree, and Random Forest) were found, the average validation AUC was used to choose between types of machine learning classifiers. The second step followed the same nature as the previous one but evaluating different fusion methods, such as the maximum rule, sum rule, product rule, weight criterion, rule-based, and model-based (Elastic Net, k-NN, and Decision Tree). The last step consisted of comparing the best late fusion pipeline to the best unimodal classifier in terms of their test AUC and calculating its gain.



### Early Fusion

Before going into detail, this approach can be divided into three steps. First, audio and video instances were carefully concatenated. Second, different types of machine learning classifiers were subjected to a modeling and selection process. Third, the best early fusion pipeline was evaluated over the test set and compared to the best unimodal classifier.

Now, a more in-depth explanation of the above-stated stages is given. The first step lay in joining audio and feature instances on the "file\_id" field. The second can be subdivided into two phases. First, 5-fold cross-validation was used for hyperparameter tuning over the training set. Second, once the best parameters for each classifier (Elastic Net, k-NN, Decision Tree, and Random Forest) were obtained, the average validation AUC was used to choose between types of machine learning classifiers. Lastly, the latter stage involved comparing the best early fusion pipeline to the best unimodal classifiers in terms of their test AUC and calculating its gain.

### 3.5.2 Unsupervised Learning

This part of the thesis investigates the question of whether unsupervised algorithms can reveal any meaningful structures in the multimodal dataset. To that end, two different approaches were taken. On the one hand, a more traditional solution, which involved the use of k-Means and Hierarchical Clustering, was studied. On the other hand, a more exploratory and graphical method, which included the use of an open-source embedding projector tool, was investigated. But first, audio (eGeMAPS set) and video features were concatenated, resulting in a total dataset of 1,259 instances and 105 dimensions.

#### Traditional Approach

This solution lay in the use of k-Means and Hierarchical Clustering with and without dimensionality reduction. Furthermore, two ways of determining the number of clusters were evaluated. This subsection of the report only covers the estimation of the number of clusters as well as the dimensionality reduction process. The clustering results can be found in Subsection 4.2.1. Regarding the employed parameters for Hierarchical Clustering, the city-block distance metric (also known as Manhattan distance) was used due to the high dimensionality of the dataset [84], and three different distance methods were evaluated (simple, complete, and weighted). Please refer to [85] for further information on the nature of the parameters.

Figure 3.5 presents the obtained  $CH(k)$  and  $s_{score}(k)$  for k-Means before dimensionality reduction was applied. These two estimation methods were calculated as detailed in Subsection 2.4.1. Table 3.3 indicates that the estimated number of clusters was two for both techniques. On the other side, Figure 3.6 and Table 3.4 show the same information but for Hierarchical Clustering. In this case, the CH index demonstrated that the best number of clusters was two for single and complete distance methods, and four for the weighted one. However, the silhouette score selected two clusters as the best option for all of them.

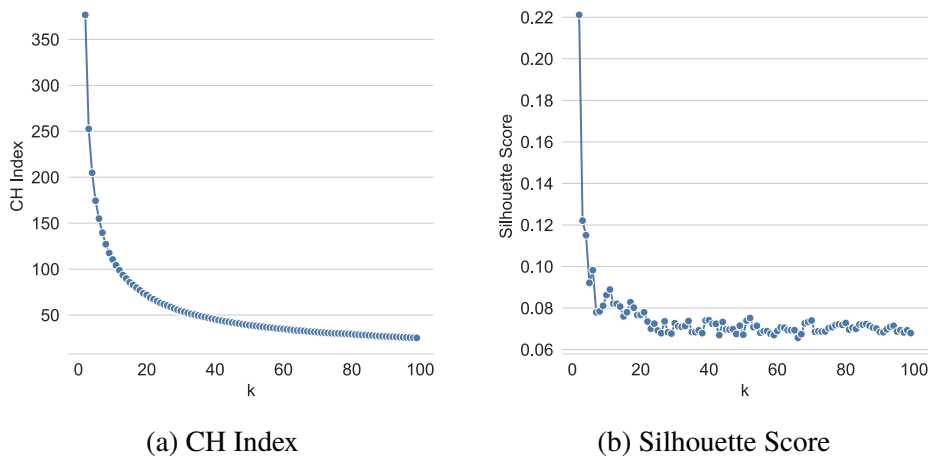


Figure 3.5: k-Means (before dimensionality reduction): Number of clusters estimation by using CH index and silhouette score, where  $k \in \{2, \dots, 100\}$ .

Method	$\hat{k}$
CH Index	2
Silhouette Score	2

Table 3.3: k-Means (before dimensionality reduction): Number of clusters estimation results.

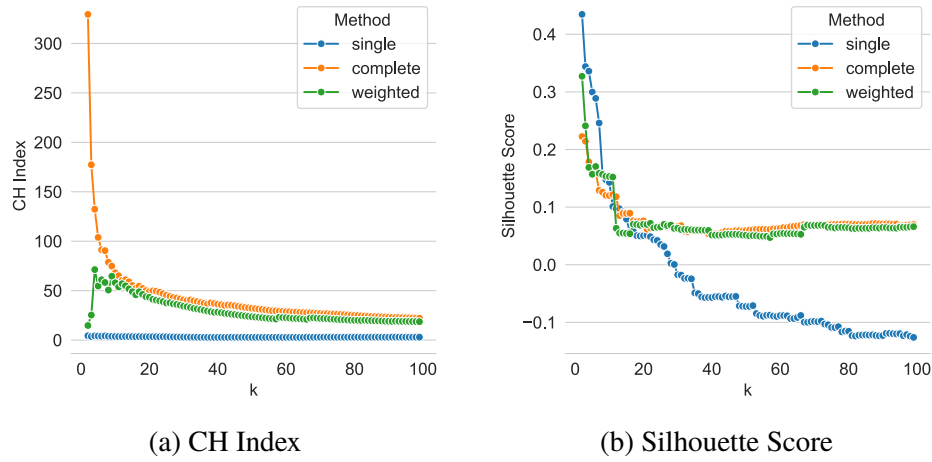


Figure 3.6: Hierarchical Clustering (before dimensionality reduction): Number of clusters estimation by using CH index and silhouette score, where  $k \in \{2, \dots, 100\}$ .

Method	Distance Method	$\hat{k}$
CH Index	single	2
	complete	2
	weighted	4
Silhouette Score	single	2
	complete	2
	weighted	2

Table 3.4: Hierarchical Clustering (before dimensionality reduction): Number of clusters estimation results.

Once the clustering without dimensionality reduction was performed, the dataset was inspected in search of weak and redundant fields. To that end, three feature reduction techniques were assessed. First, as illustrated in Figure 3.7, PCA revealed that the use of the three strongest singular values would only have explained 49% of the total variance of the data. Second, the standard deviation plot showed that there were not fields with zero variation, nor was an exaggerated drop of the variance present in the dataset (see Figure B.1). Lastly, the correlation matrix reports that there were some highly correlated features (see Figure B.2). Taking everything into consideration, the dimensionality of the multimodal dataset was diminished by dropping those fields that had a

correlation value greater than 0.9, decreasing the number of dimensions from 105 to 94 (10%). The use of this correlation threshold has been applied in many studies and has become a rule of thumb [86].

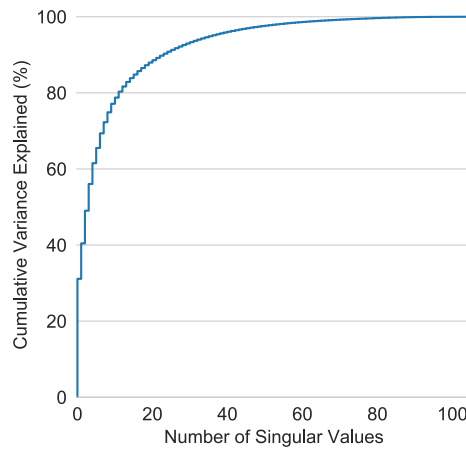


Figure 3.7: Percentage of cumulative variance explained per number of singular values.

After reducing the number of features, the CH index and silhouette score methods were used once again to determine the number of clusters. Both techniques were consistently saying that the best number of groups was two for k-Means and Hierarchical clustering (see Figure 3.8, Table 3.5, Figure 3.9, and Table 3.6).

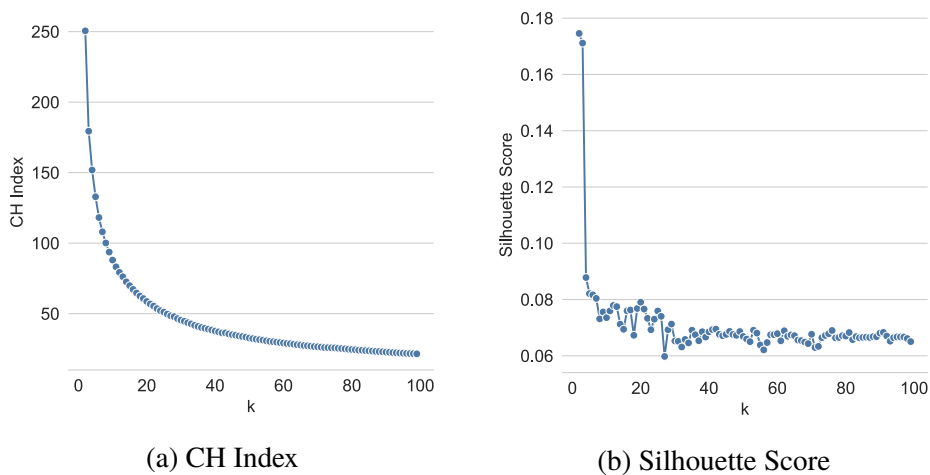


Figure 3.8: k-Means (after dimensionality reduction): Number of clusters estimation by using CH index and silhouette score, where  $k \in \{2, \dots, 100\}$ .

Method	$\hat{k}$
CH Index	2
Silhouette Score	2

Table 3.5: k-Means (after dimensionality reduction): Number of clusters estimation results.

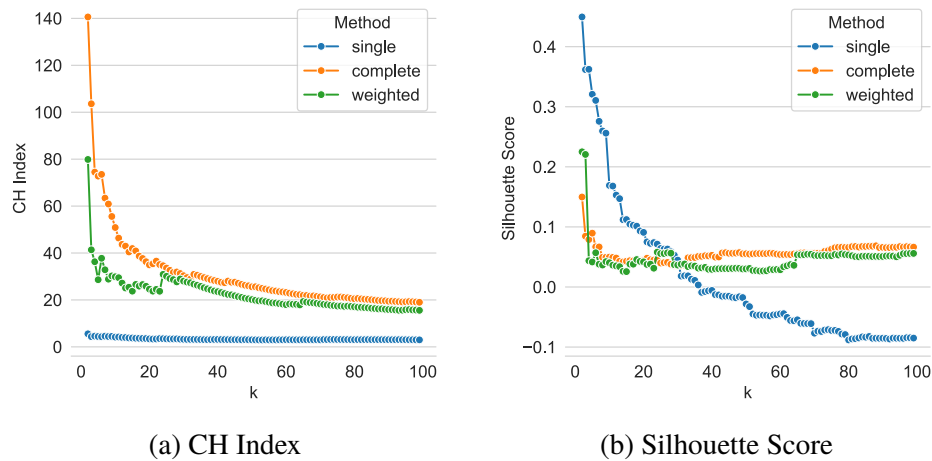


Figure 3.9: Hierarchical Clustering (after dimensionality reduction): Number of clusters estimation by using CH index and silhouette score, where  $k \in \{2, \dots, 100\}$ .

Method	Distance Method	$\hat{k}$
CH Index	single	2
	complete	2
	weighted	2
Silhouette Score	single	2
	complete	2
	weighted	2

Table 3.6: Hierarchical Clustering (after dimensionality reduction): Number of clusters estimation results.

Finally, k-Means and Hierarchical clustering were applied according to the obtained number of clusters. These results can be found in Subsection 4.2.1.

Additionally, to facilitate the interpretation of the clustering results, the problem was contemplated in a supervised manner, where the membership of the instances to the clusters corresponded to the target classes. To that end, a simple Decision Tree was trained, and the first decision nodes were analyzed.

### **Exploratory Approach**

This solution consisted of using the TensorFlow Embedding Projector<sup>2</sup>, a Google web application for interactive visualization and analysis of high-dimensional data [87]. This approach comprises two main stages. First, preparing the input data. Second, exploring the dataset. The first stage entailed the conversion of the non-reduced multimodal dataset into a TSV file, and the creation of a metadata file, which enclosed information such as the portrayed emotion, valence (positive or negative), actor id, and actor's sex. Once both files were loaded into the web application, data was ready for being explored. The system offers three different primary methods of dimensionality reduction (PCA, t-SNE, and UMAP) and can create two- and three-dimensional plots. For each of these techniques, parameters were tuned until any meaningful patterns were found. The results obtained can be found in Subsection 4.2.2.

---

<sup>2</sup>TensorFlow Embedding Projector: <https://projector.tensorflow.org/>

# Chapter 4

## Results

### 4.1 Supervised Learning

This section presents the obtained supervised learning results after following the implementation steps detailed in Subsection 3.5.1. First, the unimodal results are expounded. Second, the multimodal results are illustrated and interpreted. Lastly, the best unimodal and multimodal solutions are compared in terms of their test AUC. Bear in mind that the selection of the models was always made according to their average validation AUC, and the test AUC was only used to compare the performance between solutions.

#### 4.1.1 Unimodality

Table 4.1 lists the best audio classifiers after hyperparameter tuning was performed. As can be seen in this table, RF with the eGeMAPS parameter set outperformed the rest of the models with an average validation AUC of 0.8628. Furthermore, the table reveals two more things. First, the classifiers were not suffering from overfitting since the test AUC values were close to their correspondent validation ones on all occasions. Second, those models that used the eGeMAPS feature set always performed better. For this reason, only the extended version of the audio parameter set will be analyzed from now on.

Classifier	Average AUC (validation)	AUC (test)
Elastic Net (eGeMAPS)	0.8391	0.8504
Elastic Net (GeMAPS)	0.8237	0.8339
k-NN (eGeMAPS)	0.8032	0.8264
k-NN (GeMAPS)	0.8026	0.8212
Decision Tree (eGeMAPS)	0.7296	0.7175
Decision Tree (GeMAPS)	0.7141	0.7198
<b>Random Forest (eGeMAPS)</b>	<b>0.8628</b>	<b>0.8944</b>
Random Forest (GeMAPS)	0.8542	0.8770

Table 4.1: Unimodality: Best audio classifiers.

Figure 4.1 presents how the best unimodal audio classifier coped with the test set. The model performed better than chance for all the emotions, achieving the highest performance for amusement (amu). However, a large number of admiration (adm) instances were misclassified as relief (sou). Additionally, a more in-depth behavioral study of the classifier was conducted by computing the per-emotion feature contributions over the test set using the TreeInterpreter<sup>1</sup> package (see Figure A.1). It is apparent from this figure that the feature importances were dependent on the emotions since the most relevant parameters varied from one class to another. Furthermore, it is interesting to notice that the Hammarberg index (the difference between the energy of the highest spectral peak in the 0-2 kHz range and the one in the 2-5 kHz spectrum) [88] played a significant role in the classification of tenderness (att) and anger (col) samples.

<sup>1</sup>TreeInterpreter: <https://github.com/andosa/treeinterpreter>



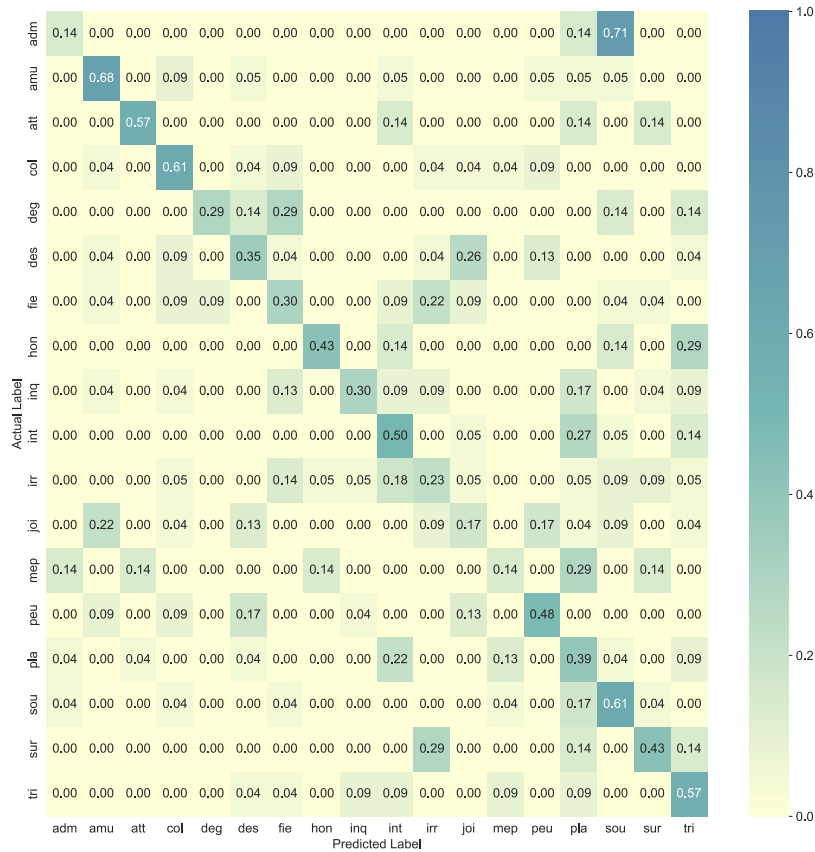


Figure 4.1: Unimodality (audio - eGeMAPS): Random Forest normalized confusion matrix (test set).

Table 4.2 encloses the best video classifiers after hyperparameter tuning was concluded. It is noticeable that RF did a better job than the rest of the classifiers, reaching an average validation AUC of 0.8323.

Classifier	Average AUC (validation)	AUC (test)
Elastic Net	0.8070	0.7845
k-NN	0.8046	0.7790
Decision Tree	0.7229	0.6941
<b>Random Forest</b>	<b>0.8323</b>	<b>0.8116</b>

Table 4.2: Unimodality: Best video classifiers.

Regarding its intraclass performance (Figure 4.2), the video classifier struggled to classify some of the emotions, especially interest (int), which was

mostly wrongly labeled as anxiety (inq) and despair (des). On the other hand, the model stood out in the prediction of amusement (amu) samples. Delving into the behavior of the classifier (Figure A.2) it can be seen that, once again, the importance of the features varied from one class to another. Furthermore, this figure reveals which AUs played a key part in detecting emotions. For example, AU6 (cheek raiser), AU10 (upper lip raiser), AU14 (dimpler), AU4 (brow lowerer), and AU12 (lip corner puller) were crucial in the detection of amusement (amu) portrayals.

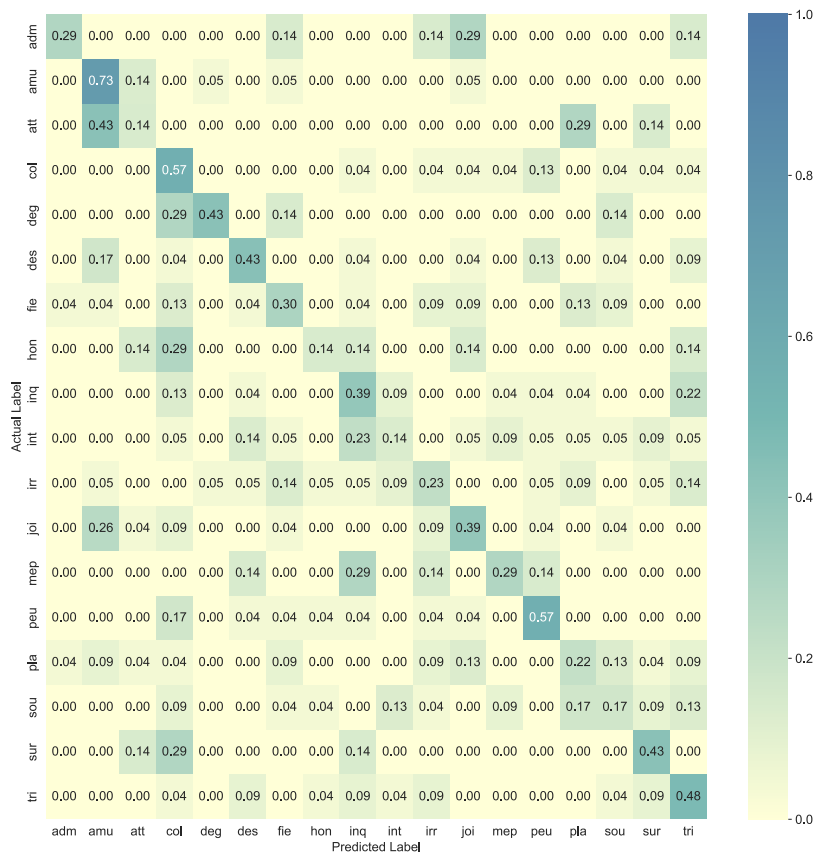


Figure 4.2: Unimodality (video): Random Forest normalized confusion matrix (test set).

## 4.1.2 Multimodality

### Late Fusion

Once the best audio and video unimodal classifiers were found, their outputs were merged by using different fusion techniques (see Table 4.3). This table

reveals that the product rule outperformed the rest of the methods, achieving an average validation AUC of 0.9078.

Fusion Technique	Average AUC (validation)	AUC (test)
Maximum Rule	0.8772	0.8980
Sum Rule	0.8972	0.9140
<b>Product Rule</b>	<b>0.9078</b>	<b>0.9195</b>
Weight Criterion	0.8973	0.9164
Rule-based	0.8671	0.8983
Elastic Net	0.8975	0.9202
k-NN	0.8078	0.8291
Decision Tree	0.6482	0.6460

Table 4.3: Multimodality (late fusion): Best fusion techniques.

The confusion matrix of the best late fusion pipeline (Figure 4.3) shows that the multimodal classifier performed better than chance for all the classes, achieving its highest performance for amusement (amu). However, some of the emotions experienced misclassification. For instance, admiration (adm) samples were mostly labeled as relief (sou), and shame (hon) as sadness (tri).

### Early Fusion

As was described in Subsection 3.5.1, audio (eGeMAPS parameter set) and video features were concatenated and used to train different classification methods. Table 4.4 details the best multimodal classifiers after hyperparameter tuning was performed. This table shows that RF defeated the rest of the models, scoring an average validation AUC of 0.9057. Moreover, it can be seen that the classifiers were able to generalize well since their test AUC scores were close to their validation ones.

Classifier	Average AUC (validation)	AUC (test)
Elastic Net	0.8823	0.8932
k-NN	0.8374	0.8490
Decision Tree	0.7450	0.7469
<b>Random Forest</b>	<b>0.9057</b>	<b>0.9320</b>

Table 4.4: Multimodality (early fusion): Best fusion classifiers.

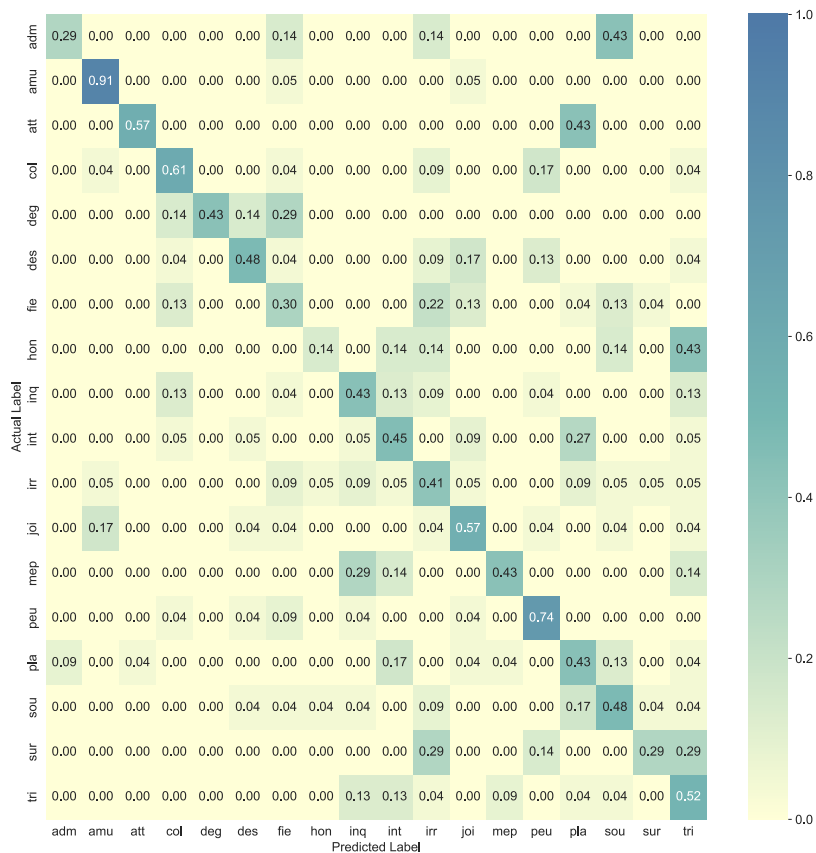


Figure 4.3: Multimodality (late fusion pipeline; product rule): normalized confusion matrix (test set).

As regards the best early fusion pipeline (Figure 4.4), its intraclass classification performance was better than random guessing for all the emotions. Nevertheless, two of them were mostly misclassified, as was the case for admiration (adm) labeled as relief (sou), and shame (hon) predicted as sadness (tri). Figure A.3 reveals the interaction of audio and video features for the best early fusion pipeline. For instance, AU6 (cheek raiser), AU12 (lip corner puller), and shimmer (difference of the peak amplitudes of consecutive  $F_0$  periods [45]) made a significant contribution to classify amusement (amu) samples, obtaining a correct classification rate of 86% (see Figure 4.4).

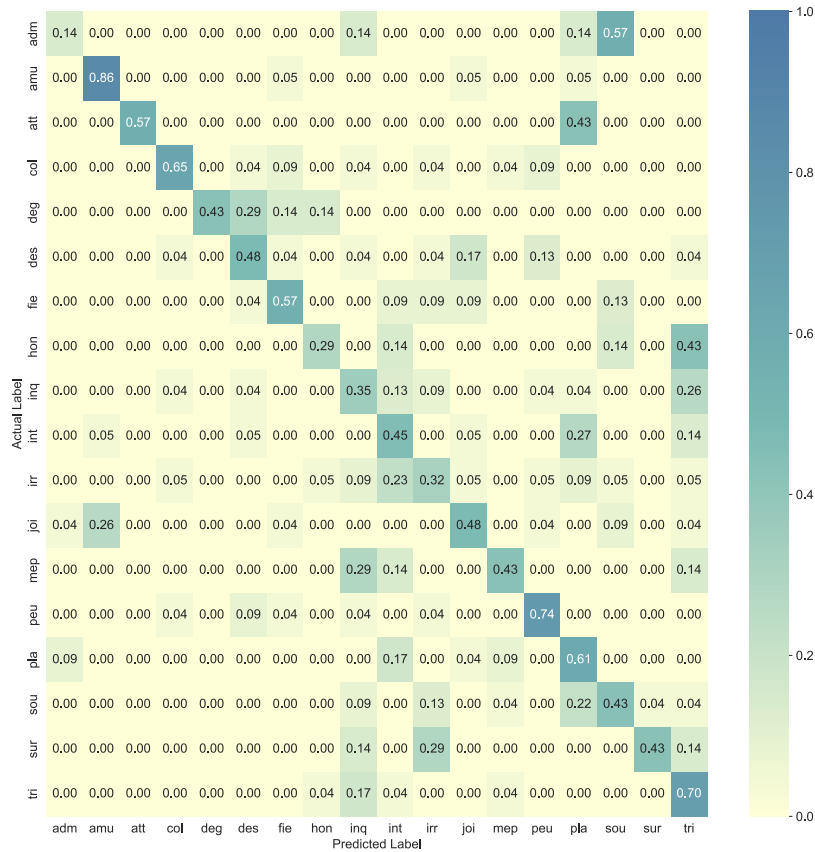


Figure 4.4: Multimodality (early fusion pipeline; Random Forest): normalized confusion matrix (test set).

### 4.1.3 General Results

After depicting the obtained results for the audio and video unimodal solutions and the late and early multimodal pipelines, this set the stage for comparing the two approaches. Table 4.5 lists and compares them in terms of their test AUC. The normalized gain was used to evaluate the superiority of the early and late multimodal pipelines over the best unimodal classifier (audio modality). This table indicates that both early and late fusion pipelines outperformed the best unimodal solution with a normalized gain of 0.3561 and 0.2377, respectively. According to Table 3.1, these values correspond to a medium gain level for the early fusion approach and a low gain level for the late one. Hence, these results favorably proved that supervised multimodal strategies outperform unimodal ones, even with a more realistic number of emotions. Lastly, Table 4.6 summarizes the total elapsed modeling time per approach. The entire process

involved the evaluation of nearly 200,000 models and took approximately three days and a half. Please refer to Appendix A.1 for a more detailed version of elapsed modeling times.

Unimodal AUC (test)		Multimodal AUC (test)		Normalized Gain [0,1]	
Audio	Video	Early	Late	Early	Late
0.8944	0.8116	0.9320	0.9195	0.3561	0.2377

Table 4.5: Single modality vs multimodality.

Approach	Number of Models	Elapsed Time (dd hh:mm:ss)
Unimodality (audio)	98,704	01 15:58:03
Unimodality (video)	49,352	00 17:07:51
Multimodality (early fusion)	49,352	00 20:39:33
Multimodality (late fusion)	2,254	00 12:37:17
<b>TOTAL</b>	<b>199,662</b>	<b>03 18:22:44</b>

Table 4.6: Supervised learning: Elapsed time.

## 4.2 Unsupervised Learning

This section encloses the unsupervised learning results after following the two approaches detailed in Subsection 3.5.2.

### 4.2.1 Traditional Approach

After determining the best number of clusters, these parameters were used as input to k-Means and Hierarchical Clustering. Both clustering techniques were evaluated over the multimodal dataset with and without dimensionality reduction.

#### Before Dimensionality Reduction

According to the results in Table 3.3, two was the best number of groups for k-Means without dimensionality reduction. Figure 4.5 illustrates the percentage

of files per emotion and cluster. This figure reveals that cluster zero principally included low-pitch emotions such as tenderness (att), shame (hon), interest (int), contempt (mep), pleasure (pla), relief (sou), and sadness (tri). In contrast, cluster one mainly contained high-arousal emotions such as amusement (amu), anger (col), despair (des), joy (joi), and panic (peu). Some of these findings were in good agreement with what was found after contemplating the clustering output in a supervised way. As shown in Figure B.3, the most relevant feature to distinguish between both groups was the “mfcc2”, a low-level spectral feature. According to the first decision node, those instances which have an “mfcc2” value greater than 0.685 were classified as cluster zero. Hence, the blue group was mainly characterized by low-pitch emotions. On the other hand, the use of complete-linkage hierarchical clustering had a similar effect on the dataset. As shown in Figure 4.6, low-pitch and high-arousal emotions were now part of cluster one and zero, respectively. Unfortunately, the other distance methods did not bring any significant results since most of the samples were in the same group.

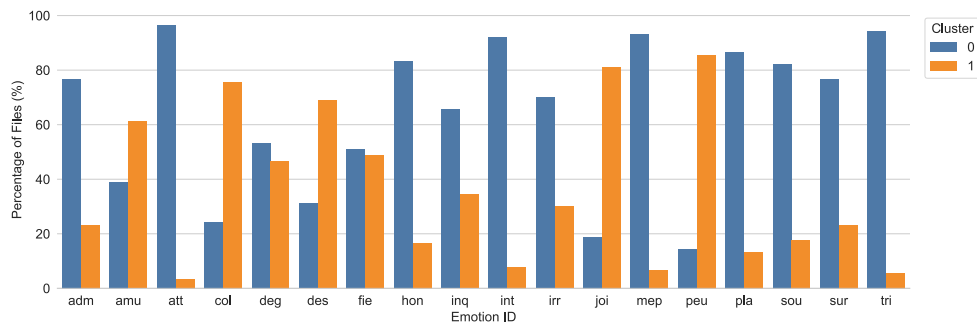


Figure 4.5: k-Means ( $k = 2$ ; before dimensionality reduction): Percentage of files per emotion and cluster.

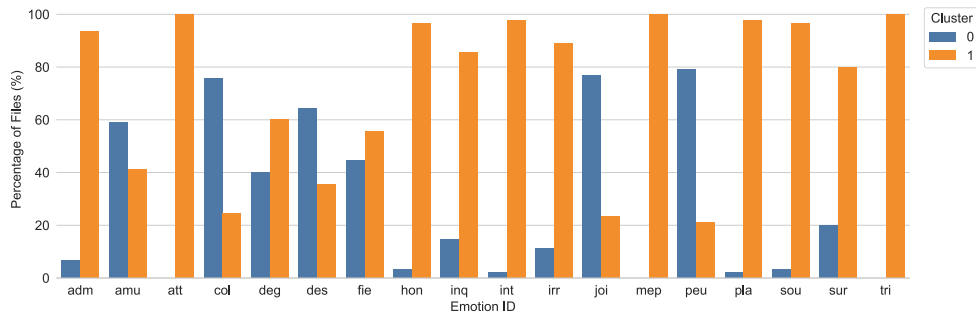


Figure 4.6: Hierarchical Clustering ( $k = 2$ , complete; before dimensionality reduction): Percentage of files per emotion and cluster.

### After Dimensionality Reduction

The obtained results after performing k-Means over the reduced dataset were in good agreement with the ones without reduction (see Figure 4.7 and Figure 4.5). The clustering method was able to preserve its behavior even with a smaller number of features. By contrast, the use of Hierarchical Clustering did not follow the same pattern (see Figure 4.8 and Figure 4.6). It can be seen in Figure 4.8 that this time most of the emotions were in cluster zero, except for panic (peu).

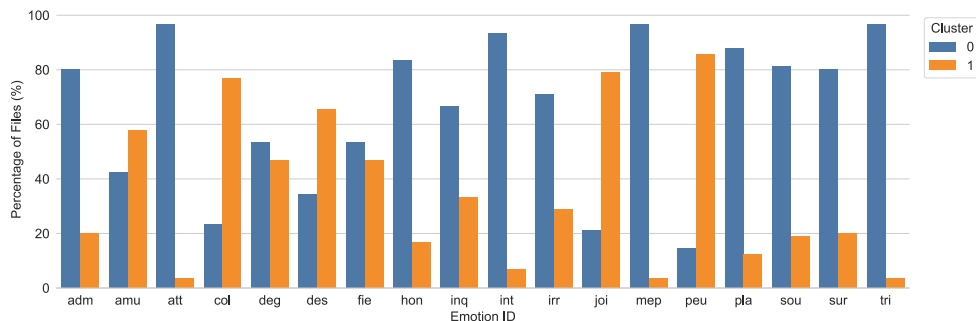


Figure 4.7: k-Means ( $k = 2$ ; after dimensionality reduction): Percentage of files per emotion and cluster.



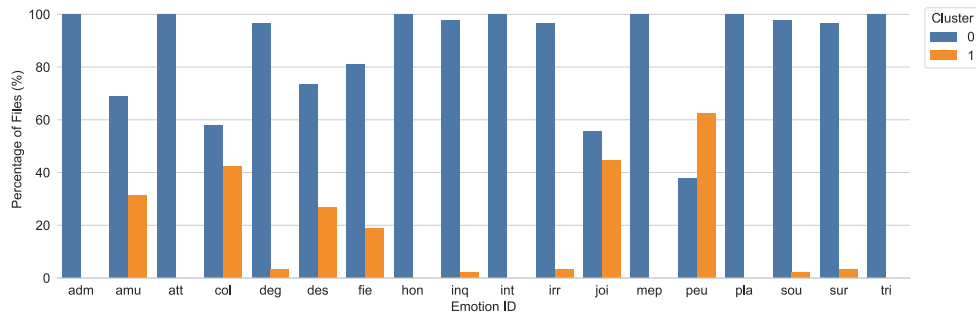


Figure 4.8: Hierarchical Clustering ( $k = 2$ , complete; after dimensionality reduction): Percentage of files per emotion and cluster.

## 4.2.2 Exploratory Approach

After preparing the features and metadata files of the non-reduced multimodal dataset (as was described in Subsection 3.5.2), the data was explored in search of meaningful patterns. To this end, three different dimensionality reduction techniques were employed via the TensorFlow embedding projector. The tunable parameters were manually adjusted until any interesting patterns were found. Next, the most significant findings are presented.

### PCA

Even though projecting the data into a two-dimensional space reduced the amount of explained variance to 37.4% (see Figure 3.7), some interesting patterns were detected. It is apparent from Figure 4.9 that the dataset could be split into two clusters. The left side mainly contained high-arousal emotions, such as samples of amusement, anger, despair, joy, and panic. On the other hand, the right side included low-pitch emotions, such as portrayals of tenderness, interest, contempt, pleasure, relief, sadness, and shame. These findings are consistent with what was found by using the traditional approach (see Figure 4.6).



Figure 4.9: PCA 2D visualization of the multimodal dataset colored by emotion. Note that 18 non-unique colors were used.

### t-SNE

The t-SNE dimensionality reduction technique was run until convergence (2,022 iterations) with a perplexity value of 68 and a learning rate of one. After this, the data points were colored by emotion, valence, actor, and actor’s sex. The algorithm grouped the data into four main clusters. As can be seen in Figure 4.10, emotions that are characterized by high-arousal were grouped together, whereas the others were split into three clusters. Figure 4.11 reveals that despite the fact that the data was not clearly divided by valence, positive and negative emotions tended to be close to each other. Something similar happened when coloring by actor, emotions portrayed by the same person tended to be nearby (see Figure 4.12). Finally, Figure 4.13 shows how the actor’s

sex played a significant role in clustering the data since the four groups were mainly composed of either female or male samples.



Figure 4.10: t-SNE 2D visualization of the multimodal dataset colored by emotion. Note that 18 non-unique colors were used.

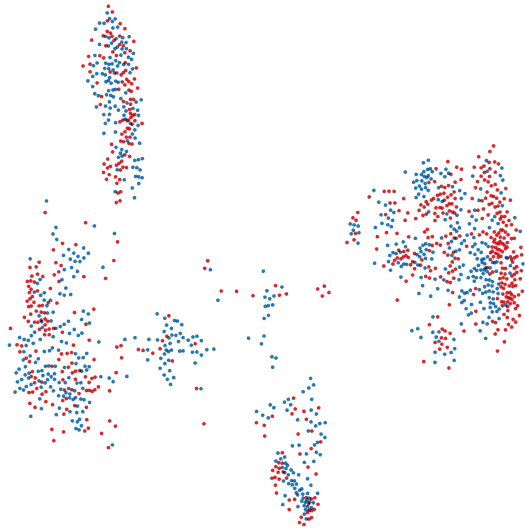


Figure 4.11: t-SNE 2D visualization of the multimodal dataset colored by valence (positive and negative).

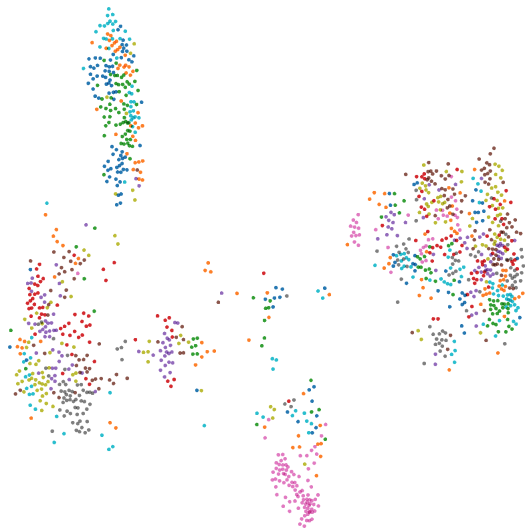


Figure 4.12: t-SNE 2D visualization of the multimodal dataset colored by actor.

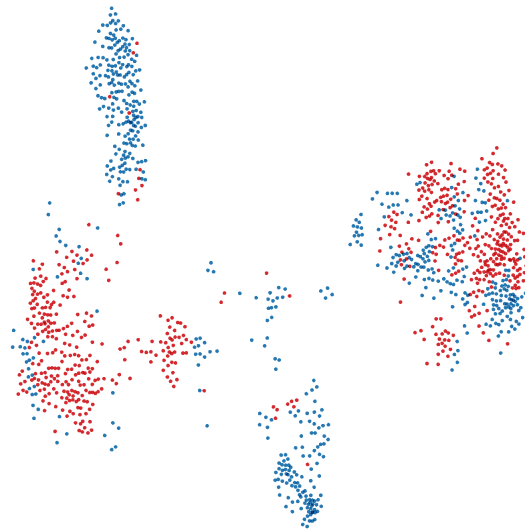


Figure 4.13: t-SNE 2D visualization of the multimodal dataset colored by actor's sex (female and male).

### UMAP

The UMAP algorithm was run for 500 epochs (it was not a tunable parameter) and 30 neighbors. Again, high-arousal and low-pitch emotions were close to each other (see Figure 4.14). However, positive and negative samples seemed to be randomly distributed this time (see Figure 4.15). On the other hand, as shown in Figure 4.16, emotions that were portrayed by the same person kept being nearby, especially for actor number three (pink dots). Lastly, similarly to t-SNE, Figure 4.17 reveals how the actor's sex played a part in the clustering results.



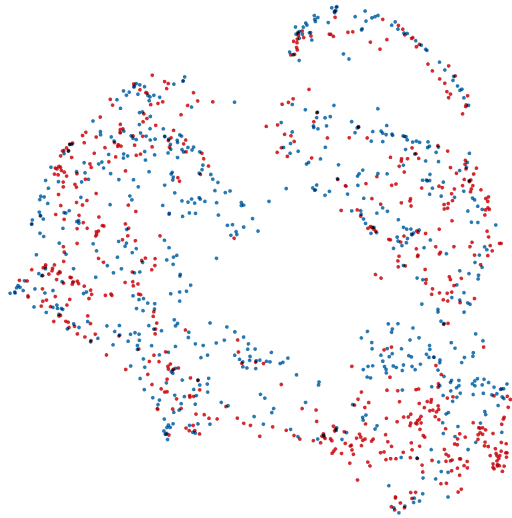


Figure 4.15: UMAP 2D visualization of the multimodal dataset colored by valence (positive and negative).



Figure 4.16: UMAP 2D visualization of the multimodal dataset colored by actor.

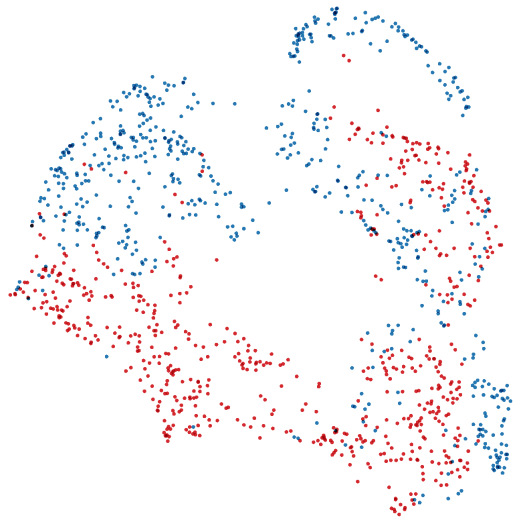


Figure 4.17: UMAP 2D visualization of the multimodal dataset colored by actor's sex (female and male).



# Chapter 5

## Discussion

### 5.1 Supervised Learning

The above-presented results confirmed that multimodal solutions could outperform unimodal ones, even with a large number of emotions. This concurs well with previous findings in the literature [6] [7] [8]. Additionally, our experiments showed that the gain of the early fusion pipeline to the best unimodal solution was significantly higher than the one obtained for the late fusion pipeline. In our opinion, this might be because the early fusion solution succeeded in exploiting the correlation between features from both modalities. This idea is in line with what many researchers in the MML field have characterized as one of the best advantages of employing early fusion pipelines [16] [59].

Regarding the performance of the classifiers, Random Forest beat the rest of the algorithms on all occasions. However, Elastic Net yielded promising results in less modeling time due to its reduced number of tunable parameters. Delving into the behavior of the classifiers, we noticed that admiration samples tended to be misclassified, whereas amusement instances were usually well classified. These findings are consistent with what was found by Bänziger et al. [29] since the average accuracy for the human recognition of admiration and amusement portrayals were one of the lowest and highest rates, respectively. Although the visual channel is known as the richest source of information [48], our results were not consistent with this idea since the audio classifiers generally performed better than the video classifiers. This may have happened because of the following reasons. First, audio signals were characterized by a larger number of features. Second, the use of average video features did not represent the visual channel well enough. Third, the exclusive

use of AUs was putting aside other informative non-verbal indicators, such as body movements. Fourth, the reliability of the extracted video features was lower than what we assumed since some videos were suffering from interlacing artifacts (see Figure 5.1a). The first two issues could have been improved by calculating not only the AUs' mean but also their standard deviation. The third one could have been addressed by employing an additional library (e.g., OpenPose [89]) to extract body posture descriptors and include them in the video feature set. The latter could have been mitigated by adding an extra pre-processing step to convert interlaced video to progressive [90]. However, as shown in Figure 5.1b, this conversion may produce shadows that could have also affected the feature extraction process. Having said that, we would like to stress the importance that calculating feature contributions had in the interpretation of the supervised classifiers. Furthermore, some of the combinations of AUs per emotion correlated favorably with the ones stated in [53]. Finally, it remains unknown whether the use of time-series features could have led to new discoveries.

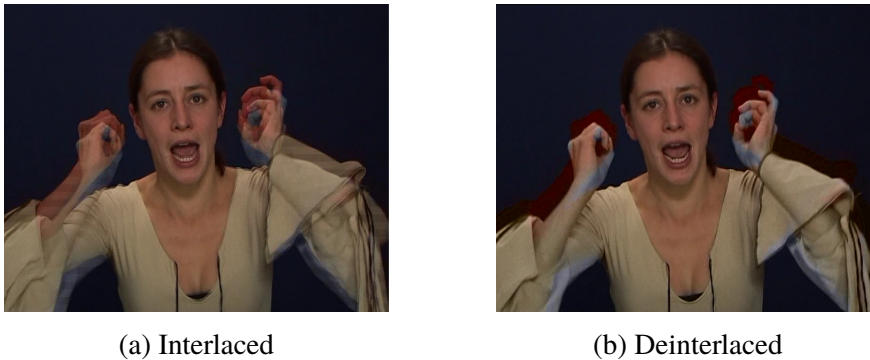


Figure 5.1: Example of interlacing artifacts of a GEMEP [29] video portraying anger. (a) Video frame suffering from interlacing artifacts. (b) Video frame after converting to progressive video.

## 5.2 Unsupervised Learning

The use of unsupervised techniques led us to find meaningful structures in the multimodal dataset. The traditional approach showed that the data could be divided into two groups and that the patterns we found were consistent across solutions. For example, one cluster was mainly constituted by low-pitch emotions, whereas the other one by high arousal portrayals. To estimate the number of clusters, we used two different clustering validation techniques

that were in good agreement in most of the cases. We recently read a paper on this topic where the authors compare eleven internal validation measures in five different scenarios and demonstrate that the S Dbw index strongly outperforms the others [91]. Thus, we will consider the use of this measure in future investigations. Additionally, we believe we employed an innovative method to determine the features that the groups were reflecting. Lastly, even though the exploratory approach had its downsides (e.g., the lack of reproducibility of the results and the reduced flexibility of the parameters), it let us discover interesting patterns, such as the role played by the emotion, emotion's valence, actor's ID, and the actor's sex. Furthermore, the output of PCA was consistent with what was found in the traditional approach. This approach has offered us a broader perspective and make us aware of the impact that the actor and actor's sex may have had on our results. Hence, future work will concentrate on finding a way to split the dataset in a per-actor basis and investigate a method to normalize spectra for anatomical differences in male and female speakers.

# Chapter 6

## Conclusions

In this thesis, we have investigated a multimodal dataset with a large number of emotions from two perspectives, breaking with the conventional supervised study of the so-called basic emotions. On the one hand, our supervised learning approach has shown that multimodal solutions can outperform unimodal ones, even with a more realistic number of emotions. Besides, the calculation of feature contributions has facilitated the interpretation of models. On the other hand, our unsupervised solution has not only led us to meaningful patterns but also to an innovative way of understanding the output of clustering techniques. Moreover, we have critically analyzed our results and gave possible solutions to our limitations.

The present study has only investigated average time-series features from acted emotions. Therefore, it remains unknown whether the use of temporal measurements and spontaneous expressions could lead to new answers. Future work will concentrate on the creation of a novel multimodal database. This will contain a wide variety of acted and spontaneous affective states conveyed with different levels of emotional intensity. The dataset will include behavioral information and images of the brain activity in response to the perception of emotion expressions. Lastly, the database will be utilized together with some of the methods used in this thesis to study how emotion expressions are produced and perceived.

# References

- [1] P. Ekman. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Henry Holt and Company, 2003. ISBN: 9780805072754.
- [2] P. P. P. Machado, L. E. Beutler, and L. S. Greenberg. “Emotion recognition in psychotherapy: impact of therapist level of experience and emotional awareness”. In: *Journal of Clinical Psychology* 12.1 (1999), pp. 39–57. DOI: 10.1002/(SICI)1097-4679(199901)55:1<39::AID-JCLP4>3.0.CO;2-V.
- [3] R. Cowie et al. “Emotion recognition in human-computer interaction”. In: *IEEE Signal Processing Magazine* 18.1 (Jan. 2001), pp. 32–80. ISSN: 1558-0792. DOI: 10.1109/79.911197.
- [4] P. Ekman. *Emotion in the Human Face*. 2nd ed. Cambridge University Press, 1983. ISBN: 9780521239929.
- [5] P. Ekman and W. V. Friesen. “Detecting deception from the body or face”. In: *Journal of Personality and Social Psychology* 29.3 (1974), pp. 288–298. DOI: 10.1037/h0036006.
- [6] L. Kessous, G. Castellano, and G. Caridakis. “Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis”. In: *Journal on Multimodal User Interfaces* 3 (2010), pp. 33–48. DOI: 10.1007/s12193-009-0025-5.
- [7] S. K. D’mello and J. Kory. “A Review and Meta-Analysis of Multimodal Affect Detection Systems”. In: *ACM Comput. Surv.* 47.3 (Feb. 2015), pp. 43–79. ISSN: 0360-0300. DOI: 10.1145/2682899.
- [8] S. Poria et al. “A Review of Affective Computing”. In: *Inf. Fusion* 37.C (Sept. 2017), pp. 98–125. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2017.02.003.

- [9] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski. “Integration of acoustic and visual speech signals using neural networks”. In: *IEEE Communications Magazine* 27.11 (Nov. 1989), pp. 65–71. ISSN: 1558-1896. DOI: 10.1109/35.41402.
- [10] H. McGurk and J. Macdonald. “Hearing lips and seeing voices”. In: *Nature* 264.5588 (Nov. 1976), pp. 746–748. DOI: 10.1038/264746a0.
- [11] L. R. Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”. In: *Proceedings of the IEEE* 77.2 (Feb. 1989), pp. 257–286. ISSN: 1558-2256. DOI: 10.1109/5.18626.
- [12] J. Picone. “Continuous speech recognition using hidden Markov models”. In: *IEEE ASSP Magazine* 7.3 (July 1990), pp. 26–41. ISSN: 1558-1284. DOI: 10.1109/53.54527.
- [13] M. Brand, N. Oliver, and A. Pentland. “Coupled hidden Markov models for complex action recognition”. In: *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. June 1997, pp. 994–999. DOI: 10.1109/CVPR.1997.609450.
- [14] M. Gurban et al. “Dynamic Modality Weighting for Multi-Stream HMMs in Audio-Visual Speech Recognition”. In: *Proceedings of the 10th International Conference on Multimodal Interfaces*. ICMI ’08. Chania, Crete, Greece: Association for Computing Machinery, 2008, pp. 237–240. ISBN: 9781605581989. DOI: 10.1145/1452392.1452442.
- [15] J. Ngiam et al. “Multimodal Deep Learning”. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML’11. Bellevue, Washington, USA: Omnipress, 2011, pp. 689–696. DOI: 10.5555/3104482.3104569.
- [16] T. Baltrušaitis, C. Ahuja, and L. Morency. “Multimodal Machine Learning: A Survey and Taxonomy”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (Feb. 2019), pp. 423–443. DOI: 10.1109/TPAMI.2018.2798607.
- [17] F. De la Torre and J.F. Cohn. “Guide to Visual Analysis of Humans: Looking at People”. In: Springer, 2011. Chap. Facial Expression Analysis. ISBN: 978-0-85729-997-0.
- [18] H. Al Osman and T. Falk. “Multimodal Affect Recognition: Current Approaches and Challenges”. In: *Emotion and Attention Recognition Based on Biological Signals and Images*. Ed. by Seyyed Abed Hosseini. InTechOpen, Feb. 2017. Chap. 5, pp. 59–86. ISBN: 978-953-51-2915-8. DOI: 10.5772/65683.

- [19] C. Busso et al. “Analysis of emotion recognition using facial expressions, speech and multimodal information”. In: Jan. 2004, pp. 205–211. DOI: 10.1145/1027933.1027968.
- [20] M. Pantic et al. “Affective Multimodal Human-Computer Interaction”. In: *Proceedings of the 13th Annual ACM International Conference on Multimedia*. MULTIMEDIA '05. Hilton, Singapore: Association for Computing Machinery, 2005, pp. 669–676. ISBN: 1595930442. DOI: 10.1145/1101149.1101299.
- [21] M. Paelari, B. Huet, and R. Chellali. “Towards Multimodal Emotion Recognition: A New Approach”. In: *Proceedings of the ACM International Conference on Image and Video Retrieval*. CIVR '10. Xi'an, China: Association for Computing Machinery, 2010, pp. 174–181. ISBN: 9781450301176. DOI: 10.1145/1816041.1816069.
- [22] C. Marechal et al. “Survey on AI-Based Multimodal Methods for Emotion Detection”. In: *High-Performance Modelling and Simulation for Big Applications: Selected Results of the COST Action IC1406 cHiPSet*. Ed. by J. Kołodziej and H. González-Vélez. Cham: Springer International Publishing, 2019, pp. 307–324. ISBN: 978-3-030-16272-6. DOI: 10.1007/978-3-030-16272-6\_11.
- [23] F. Canento et al. “Multimodal biosignal sensor data handling for emotion recognition”. In: *SENSORS, 2011 IEEE*. Limerick, Ireland: IEEE, Oct. 2011, pp. 647–650. DOI: 10.1109/ICSENS.2011.6127029.
- [24] E. Ghaleb, M. Popa, and S. Asteriadis. “Multimodal and Temporal Perception of Audio-visual Cues for Emotion Recognition”. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. Cambridge, United Kingdom: IEEE, Dec. 2019, pp. 552–558. DOI: 10.1109/ACII.2019.8925444.
- [25] W. Xun, Z. Wei-Long, and L. Bao-Liang. *Investigating EEG-Based Functional Connectivity Patterns for Multimodal Emotion Recognition*. Apr. 2020. arXiv: 2004.01973 [cs.HC].
- [26] B. Azari et al. *Comparing supervised and unsupervised approaches to emotion categorization in the human brain, body, and subjective experience*. Mar. 2020. DOI: 10.31234/osf.io/egh2t.
- [27] D. T. Cordaro et al. “Universals and cultural variations in 22 emotional expressions across five cultures”. In: *Emotion* 18.1 (2018), pp. 75–93. DOI: 10.1037/emo0000302.

- [28] A. S. Cowen et al. “The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures”. In: *Nature Human Behaviour* 3 (Mar. 2019), pp. 369–382. DOI: 10.1038/s41562-019-0533-6.
- [29] T. Bänziger, M. Mortillaro, and K. R. Scherer. “Introducing the Geneva Multimodal Expression Corpus for Experimental Research on Emotion Perception”. In: *Emotion* 12 (2012), pp. 1161–1179. DOI: 10.1037/a0025827.
- [30] X. Zhang et al. “BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database”. In: *Image and Vision Computing*. Vol. 32. 10. ELSEVIER, Oct. 2014, pp. 692–706. DOI: 10.1016/j.imavis.2014.06.002.
- [31] K. Panetta. *5 trends appear on the Gartner hype cycle for emerging technologies, 2019*. [Online]. Available: <https://www.gartner.com/smarterwithgartner/5-trends-appear-on-the-gartner-hype-cycle-for-emerging-technologies-2019> [Accessed: 28 June 2020]. Aug. 2019.
- [32] M. F. Valstar et al. “Meta-Analysis of the First Facial Expression Recognition Challenge”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.4 (2012), pp. 966–979. DOI: 10.1109/TSMCB.2012.2200675.
- [33] B. Schuller et al. “Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge”. In: *Computer Speech & Language* 53 (2019), pp. 156–180. ISSN: 0885-2308. DOI: 10.1016/j.csl.2018.02.004.
- [34] F. Eyben et al. “Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor”. In: *Proceedings of the 21st ACM International Conference on Multimedia*. MM ’13. Barcelona, Spain: Association for Computing Machinery, Oct. 2013, pp. 835–838. ISBN: 9781450324045. DOI: 10.1145/2502081.2502224.
- [35] T. Baltrusaitis et al. “OpenFace 2.0: Facial Behavior Analysis Toolkit”. In: *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. Xi’an, China: IEEE, May 2018, pp. 59–66. DOI: 10.1109/FG.2018.00019.



- [36] Michel Valstar et al. “AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge”. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. AVEC '16. Amsterdam, The Netherlands: Association for Computing Machinery, 2016, pp. 3–10. ISBN: 9781450345163. DOI: 10.1145/2988257.2988258.
- [37] Fabien Ringeval et al. “AVEC 2017: Real-Life Depression, and Affect Recognition Workshop and Challenge”. In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. AVEC '17. Mountain View, California, USA: Association for Computing Machinery, 2017, pp. 3–9. ISBN: 9781450355025. DOI: 10.1145/3133944.3133953.
- [38] Fabien Ringeval et al. “AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition”. In: *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. AVEC'18. Seoul, Republic of Korea: Association for Computing Machinery, 2018, pp. 3–13. ISBN: 9781450359832. DOI: 10.1145/3266302.3266316.
- [39] Fabien Ringeval et al. *AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition*. 2019. arXiv: 1907.11510 [cs.HC].
- [40] H. Becker et al. “Emotion Recognition Based on High-Resolution EEG Recordings and Reconstructed Brain Sources”. In: *IEEE Transactions on Affective Computing* 11.2 (2020), pp. 244–257. DOI: 10.1109/TAFFC.2017.2768030.
- [41] B. H. Kim and S. Jo. “Deep Physiological Affect Network for the Recognition of Human Emotions”. In: *IEEE Transactions on Affective Computing* 11.2 (2020), pp. 230–243. DOI: 10.1109/TAFFC.2018.2790939.
- [42] Philipp V. Rouast, Marc Adam, and Raymond Chiong. “Deep Learning for Human Affect Recognition: Insights and New Developments”. In: *IEEE Transactions on Affective Computing* (2019). DOI: 10.1109/taffc.2018.2890471.
- [43] 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. DOI: 10.1109/JSTSP.2017.2764438.
- [44] K. He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

- [45] F. Eyben et al. “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing”. In: *IEEE Transactions on Affective Computing* 7.2 (Aug. 2016), pp. 190–202. DOI: 10.1109/TAFFC.2015.2457417.
- [46] Björn Schuller et al. “The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism”. In: Aug. 2013, pp. 148–152.
- [47] F. Ringeval et al. “Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions”. In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 2013, pp. 1–8. DOI: 10.1109/FG.2013.6553805.
- [48] A. Mehrabian and S.R. Ferris. “Inference of attitudes from nonverbal communication in two channels”. In: *Journal of consulting psychology* 31.3 (June 1967), pp. 248–252. DOI: 10.1037/h0024648.
- [49] I. R. Titze. *Principles of Voice Production*. 1st ed. Prentice Hall, Mar. 1994. ISBN: 013717893X.
- [50] K.R. Scherer. “Vocal affect expression: A review and a model for future research”. In: *Psychological Bulletin* 99.2 (Mar. 1986), pp. 143–165. DOI: 10.1037/0033-2909.99.2.143.
- [51] N. Ambady and M. Weisbuch. “Nonverbal behavior”. In: *Handbook of Social Psychology*. Ed. by S.T. Fiske, D.T. Gilbert, and Lindzey G. 5th ed. John Wiley & Sons, Inc., Feb. 2010, pp. 464–497. ISBN: 978-0-470-13747-5.
- [52] C. Darwin. *The Expression of the Emotions in Man and Animals*. 4th ed. Originally published in 1872. Oxford University Press, 2009.
- [53] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto: Consulting Psychologists Press, 1977.
- [54] D. McDuff et al. “AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit”. In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’16. San Jose, California, USA: Association for Computing Machinery, 2016, pp. 3723–3726. DOI: 10.1145/2851581.2890247.
- [55] M. J. den Uyl and H. van Kuilenburg. “The FaceReader: Online facial expression recognition”. In: *Proceedings of Measuring Behavior 2005*. Wageningen, 2005, pp. 589–590.

- [56] *OKAO Vision Software Library*. [Online]. Available: <https://www.components.omron.com/sensors/image-sensing/solution/software-library>. [Accessed: 28 June 2020]. OMRON Corporation. Japan, 2009.
- [57] B. Jiang, M. F. Valstar, and M. Pantic. “Action unit detection using sparse appearance descriptors in space-time video volumes”. In: *Face and Gesture 2011*. Santa Barbara, CA, USA: IEEE, 2011, pp. 314–321. DOI: 10.1109/FG.2011.5771416.
- [58] S. Li and W. Deng. “Deep Facial Expression Recognition: A Survey”. In: *IEEE Transactions on Affective Computing* (Mar. 2020), pp. 1–1. DOI: 10.1109/TAFFC.2020.2981446.
- [59] Z. Zeng et al. “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.1 (Jan. 2009), pp. 39–58. DOI: 10.1109/TPAMI.2008.52.
- [60] M. Wöllmer et al. “LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework”. In: *Image and Vision Computing* 31.2 (2013). Affect Analysis In Continuous Input, pp. 153–163. DOI: 10.1016/j.imavis.2012.03.001.
- [61] S. Chen and Q. Jin. “Multi-Modal Dimensional Emotion Recognition Using Recurrent Neural Networks”. In: *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. AVEC ’15. Brisbane, Australia: Association for Computing Machinery, 2015, pp. 49–56. DOI: 10.1145/2808196.2811638.
- [62] G. Hughes. “On the mean accuracy of statistical pattern recognizers”. In: *IEEE Transactions on Information Theory* 14.1 (Jan. 1968), pp. 55–63. DOI: 10.1109/TIT.1968.1054102.
- [63] C. A. Corneanu et al. “Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.8 (Jan. 2016), pp. 1548–1568. DOI: 10.1109/TPAMI.2016.2515606.
- [64] H. Zou and T. Hastie. “Regularization and variable selection via the Elastic Net”. In: *Journal of the Royal Statistical Society, Series B* 67.2 (2005), pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.
- [65] T. Caliński and J. Harabasz. “A dendrite method for cluster analysis”. In: *Communications in Statistics* 3.1 (Sept. 1974), pp. 1–27. DOI: 10.1080/03610927408827101.

- [66] P. J. Rousseeuw. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (Nov. 1987), pp. 53–65. ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7.
- [67] S. Jonathon. *A Tutorial on Principal Component Analysis*. 2014. arXiv: 1404.1100 [cs.LG].
- [68] R. Garreta and G. Moncecchi. *Learning Scikit-Learn: Machine Learning in Python*. Packt Publishing, 2013. ISBN: 1783281936.
- [69] L.v.d. Maaten and G. Hinton. “Visualizing data using t-SNE”. In: *Journal of Machine Learning Research* 9 (Nov. 2008), pp. 2579–2605.
- [70] M. Wattenberg, F. Viégas, and I. Johnson. “How to Use t-SNE Effectively”. In: *Distill* (Oct. 2016). DOI: 10.23915/distill.00002.
- [71] L. McInnes, J. Healy, and J. Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2018. arXiv: 1802.03426 [stat.ML].
- [72] L. McInnes, J. Healy, and J. Melville. *How UMAP Works*. [Online]. Available: [https://umap-learn.readthedocs.io/en/latest/how\\_umap\\_works.html](https://umap-learn.readthedocs.io/en/latest/how_umap_works.html). [Accessed: 28 June 2020]. 2018.
- [73] C. E. Metz. “Basic principles of ROC analysis”. In: *Seminars in Nuclear Medicine* 8.4 (Oct. 1978), pp. 283–298. DOI: 10.1016/S0001-2998(78)80014-2.
- [74] P. Branco, L. Torgo, and R. Ribeiro. *A Survey of Predictive Modelling under Imbalanced Distributions*. 2015. arXiv: 1505.01658 [cs.LG].
- [75] C. Ferri, J. Hernández-Orallo, and R. Modroiu. “An experimental comparison of performance measures for classification”. In: *Pattern Recognition Letters* 30.1 (Jan. 2009), pp. 27–38. DOI: 10.1016/j.patrec.2008.08.010.
- [76] G. James et al. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN: 1461471370. DOI: 10.1007/978-1-4614-7138-7.
- [77] R. Hake. “Interactive-Engagement Versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses”. In: *American Journal of Physics* 66.1 (Jan. 1998), pp. 64–74. DOI: 10.1119/1.18809.
- [78] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

- [79] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [80] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.
- [81] 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.
- [82] E. G. Krumhuber et al. “A Review of Dynamic Datasets for Facial Expression Research”. In: *Emotion Review* 9.3 (2017), pp. 280–292. DOI: 10.1177/1754073916670022.
- [83] B. Sjardin, L. Massaron, and A. Boschetti. *Large Scale Machine Learning with Python*. Packt Publishing, 2016. ISBN: 9781785888021.
- [84] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. “On the Surprising Behavior of Distance Metrics in High Dimensional Space”. In: *Database Theory — ICDT 2001*. Ed. by J. Van den Bussche and V. Vianu. Vol. 1973. Berlin, Heidelberg: Springer Berlin Heidelberg, Oct. 2001, pp. 420–434. DOI: 10.1007/3-540-44503-X\_27.
- [85] SciPy. *Hierarchical clustering (scipy.cluster.hierarchy.linkage)*. [Online]. Available: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html>. [Accessed: 28 June 2020]. Dec. 2019.
- [86] M.H. Katz. *Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers*. Cambridge University Press, 2011. ISBN: 9781139500319.
- [87] D. Smilkov et al. *Embedding Projector: Interactive Visualization and Interpretation of Embeddings*. 2016. arXiv: 1611.05469 [stat.ML].
- [88] B. Hammarberg et al. “Perceptual and Acoustic Correlates of Abnormal Voice Qualities”. In: *Acta Oto-Laryngologica* 90.1-6 (1980), pp. 441–451. DOI: 10.3109/00016488009131746.
- [89] Z. Cao et al. “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019). DOI: 10.1109/TPAMI.2019.2929257.
- [90] G. De Haan and E. B. Bellers. “Deinterlacing-an overview”. In: *Proceedings of the IEEE* 86.9 (Sept. 1998), pp. 1839–1857. DOI: 10.1109/5.705528.

- [91] Y. Liu et al. “Understanding of Internal Clustering Validation Measures”. In: *2010 IEEE International Conference on Data Mining*. Sydney, NSW, 2010, pp. 911–916. doi: 10.1109/ICDM.2010.35.

# Appendix A

## Supervised Learning

### A.1 Elapsed Modeling Time

Classifier	Number of Models	Elapsed Time (dd hh:mm:ss)
Elastic Net (eGeMAPS)	400	00 00:12:03
Elastic Net (GeMAPS)	400	00 00:10:12
k-NN (eGeMAPS)	7,952	00 04:29:40
k-NN (GeMAPS)	7,952	00 03:04:43
Decision Tree (eGeMAPS)	36,000	00 07:06:47
Decision Tree (GeMAPS)	36,000	00 09:14:40
Random Forest (eGeMAPS)	5,000	00 08:24:41
Random Forest (GeMAPS)	5,000	00 07:15:17
<b>TOTAL</b>	<b>98,704</b>	<b>01 15:58:03</b>

Table A.1: Single modality: Elapsed time audio classifiers.

Classifier	Number of Models	Elapsed Time (dd hh:mm:ss)
Elastic Net	400	00 00:08:33
k-NN	7,952	00 03:10:44
Decision Tree	36,000	00 06:55:41
Random Forest	5,000	00 06:52:53
<b>TOTAL</b>	<b>49,352</b>	<b>00 17:07:51</b>

Table A.2: Single modality: Elapsed time video classifiers.

Classifier	Number of Models	Elapsed Time (dd hh:mm:ss)
Elastic Net	400	00 00:12:20
k-NN	7,952	00 03:04:30
Decision Tree	36,000	00 07:56:13
Random Forest	5,000	00 09:26:30
<b>TOTAL</b>	<b>49,352</b>	<b>00 20:39:33</b>

Table A.3: Multimodality (early fusion): Elapsed time fusion techniques.

Fusion Technique	Number of Models	Elapsed Time (dd hh:mm:ss)
Maximum Rule	1	-
Sum Rule	1	-
Product Rule	1	-
Weight Criterion	20	00 00:08:39
Rule-based	1	-
Elastic Net	400	00 02:55:24
k-NN	930	00 04:49:45
Decision Tree	900	00 04:43:29
<b>TOTAL</b>	<b>2,254</b>	<b>00 12:37:17</b>

Table A.4: Multimodality (late fusion): Elapsed time fusion techniques.





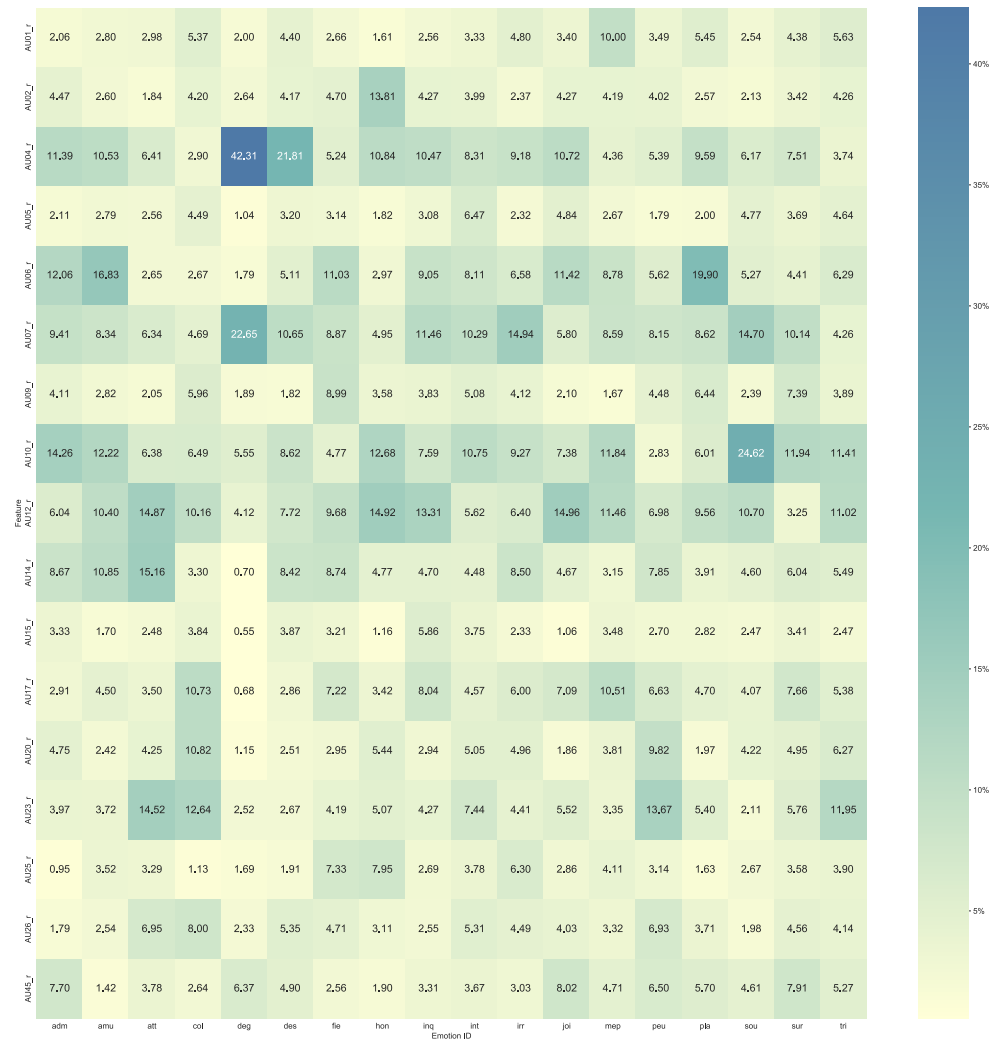


Figure A.2: Unimodality (video): Random Forest average feature contributions per emotion (%).

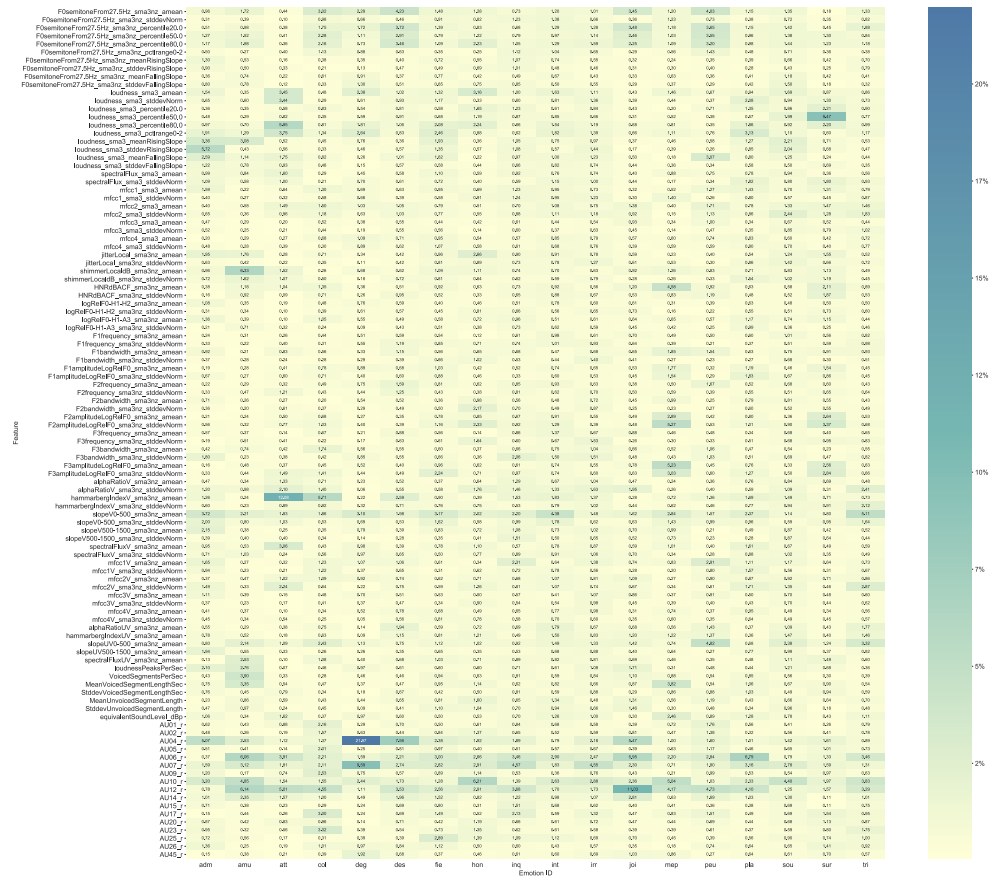


Figure A.3: Multimodality (early fusion pipeline; Random Forest): average feature contributions per emotion (%).

# **Appendix B**

## **Unsupervised Learning**

### **B.1 Traditional Approach**

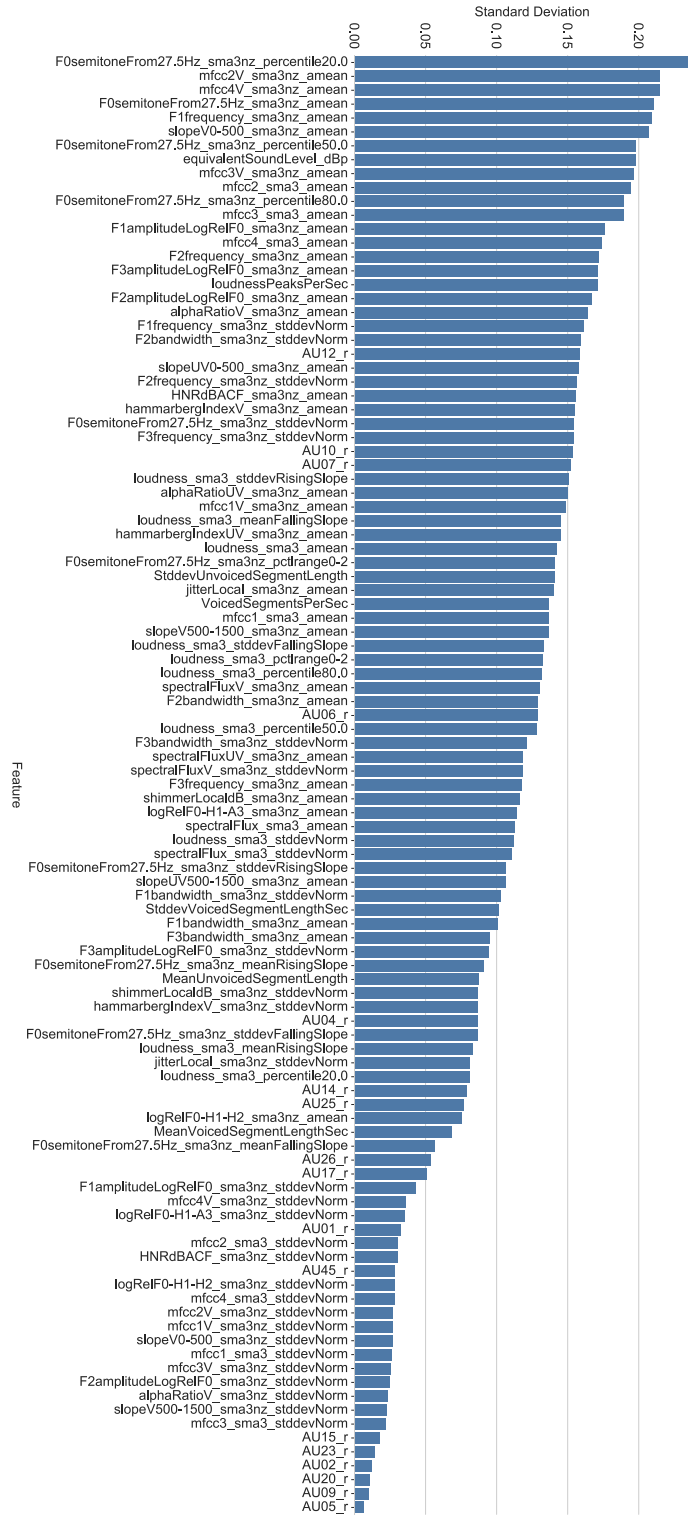


Figure B.1: Standard deviation per feature.

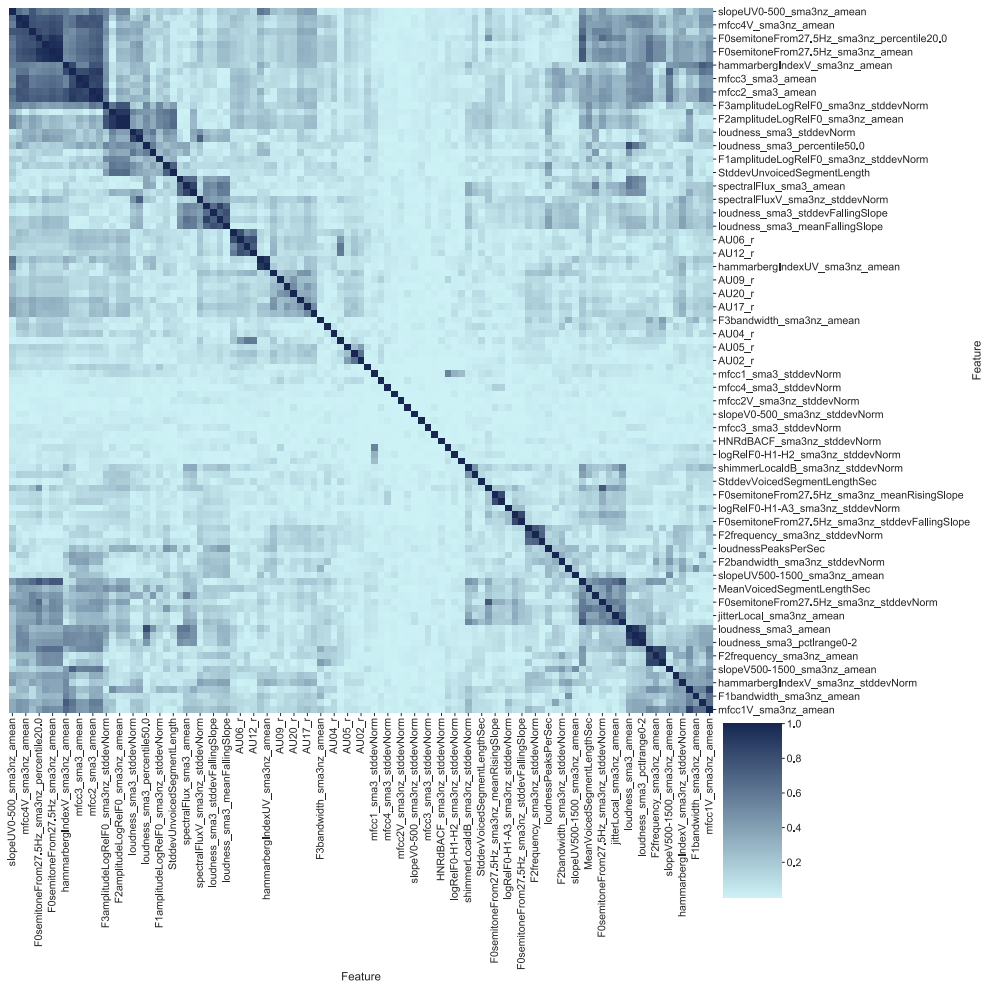


Figure B.2: Correlation matrix before dimensionality reduction. Note that not all the feature labels could be displayed.

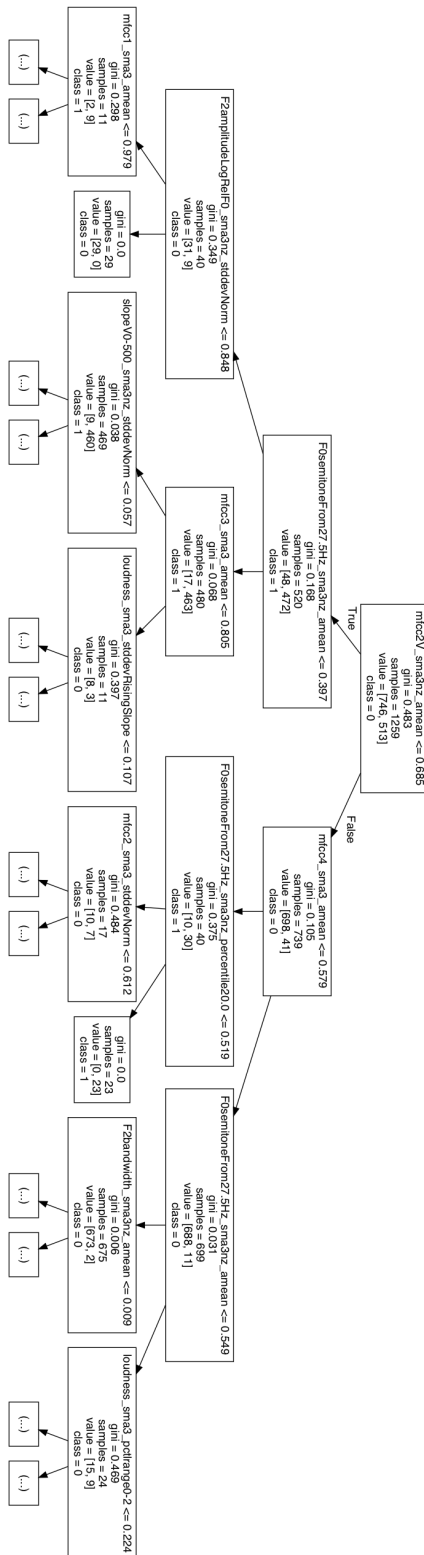


Figure B.3: Fragment of the Decision Tree used to interpret the clustering. The model was trained on the output of the k-Means ( $k = 2$ ; before dimensionality reduction).







TRITA -EECS-EX