Université Pierre et Marie Curie

École Doctorale numéro 130 :
Informatique, Télécommunications et Électronique de Paris

Multimodal detection of stress: evaluation of the impact of several assessment strategies

Thèse de doctorat

Jonathan Aigrain

Date de soutenance : 05/12/2016

Composition du jury :

Directrice de thèse :	Séverine Dubuisson	UPMC Sorbonne Universités, ISIR
Encadrants :	Mohamed Chetouani	UPMC Sorbonne Universités, ISIR
	Marcin Detyniecki	AXA Assurances
Rapporteurs :	Jean-Claude Martin	Université Paris Sud, LIMSI-CNRS
	Alessandro Vinciarelli	University of Glasgow
Examinateurs :	Catherine Pélachaud	UPMC Sorbonne Universités, ISIR
	Lionel Prevost	ESIA, LRD
	Dominique Vaufreydaz	Université Grenoble Alpes, LIG

Abstract

It is now widely accepted that stress plays an important role in modern societies. It impacts the body and the mind at several levels and the association between stress and disease has been observed in several studies. However, there is no consensual definition of stress yet, and therefore there is no consensual way of assessing it either. Thus, although the quality of assessment is a key factor to build robust stress detection solutions, researchers have to choose among a wide variety of assessment strategies. This heterogeneity impacts the validity of comparing solutions among them.

In this thesis, we evaluate the impact of several assessment strategies for stress detection. We first review how different fields of research define and assess stress. Then, we describe how we collected stress data along with multiple assessments. We also study the association between these assessments. We present the behavioural and physiological features that we extracted for our experiments. Finally, we present the results we obtained regarding the impact of assessment strategies on 1) data normalization, 2) feature classification performance and 3) on the design of machine learning algorithms.

Overall, we argue that one has to take a global and comprehensive approach to design stress detection solutions.

Résumé

Il est maintenant largement accepté que le stress joue un rôle important dans les sociétés modernes. Le stress impacte en effet le corps et l'esprit à différents niveaux. De plus, le lien entre stress et maladie a été observé dans plusieurs études. Cependant, il n'y a pas encore de définition consensuelle du stress, et par conséquent il n'y a pas de manière consensuelle de le mesurer. Ainsi, bien que la qualité de la mesure joue un rôle majeur dans la réalisation de solutions robustes de détection du stress, les chercheurs doivent choisir une stratégie de mesure parmi un grand nombre de possibilités. Cette hétérogénéité impacte la validité des comparaisons faites entre les différentes solutions.

Dans cette thèse, nous évaluons l'impact de plusieurs stratégies de mesure pour la détection du stress. Dans un premier temps, nous résumons comment différents domaines de recher-

che définissent et mesurent le stress. Nous décrivons ensuite comment nous avons collecté des données de sujets en situation stressante ainsi que plusieurs mesures du stress. Nous étudions également les liens entre ces différentes mesures. Par la suite, nous présentons les descripteurs comportementaux et physiologiques que nous avons extraits pour nos expériences. Enfin, nous présentons les résultats obtenus concernant l'impact des stratégies de mesure sur 1) la normalisation de données, 2) la performance des descripteurs pour la classification et 3) sur la conception d'algorithmes d'apprentissage automatique.

De manière générale, nous défendons l'idée qu'il faut adopter une approche globale pour concevoir une solution de détection du stress.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Context and motivation

In 1997, Rosalind Picard pioneered the expression "affective computing" in her book of the same name [99]. In this book, she discussed why computers would need to be able to express and recognize affect and emotion. According to her, it would be beneficial for the quality of human-computer interactions and also help advance emotion and cognition theory. 20 years later, affective computing is now a dynamic field of research that proposes frameworks for diverse problematics such as emotionnal speech synthesis, virtual avatars or depression detection. In 2003, Picard listed the main challenges of affective computing [98]. Recognizing affect, emotion and mental state is one of them.

In this thesis, we focus on automatic stress detection. It is now widely accepted that stress plays an important role in today's people lifestyles. In *The Global Burden of Disease* [82], which was published in 1996 by the Wold Health Organization (WHO), it is estimated that depression, stress and anxiety disorders will become the second most frequent disabilities behind heart diseases. Kalia evaluated the economic impact of stress in [61] and reported that the Mental Health Foundation estimates that stress costs 3 billion pounds per year to the British industry. Overall, stress is omnipresent in our society and can be of various intensity, from being late to a meeting because of traffic jam to post-traumatic stress disorders (PTSD).

However, stress as we know it today is still a relatively new concept. Hans Selye is often considered to be the one who pioneered modern research on stress with his book *The Stress*

of Life [112], published in 1956. He defined stress as the "nonspecific response of the body to noxious stimuli". However, as we will see later in this document, this definition has been criticized and reviewed in several ways. Nowadays, there are still researchers who work on refining the stress concept, such as Koolhaas *et al.* in 2011 [70].

Overall, we can see that, despite being omnipresent, stress is a complex concept that we do not fully understand yet. We believe that, because of this observation, developing solutions for automatic stress detection is particularly relevant. Being able to detect stress will indeed significantly improve the quality of human computer interaction and of healthcare systems since stress is so present in our society. Several works have already studied how automatic stress detection may improve the safety of drivers [42, 54], the robustness of speech-based interfaces [39, 138] or the prevention of stress-related health problems [133]. In addition, studying stress in different contexts will help understand its effect on the body and the mind. Therefore, this work has two main objectives:

- Evaluate in a systematic way the impact of stress on physiology and behaviour while taking into account the variety of assessment strategies.
- Tackle the methodological problems one faces when designing an emotion/mental state recognition system.

1.2 Challenges

Automatic recognition of affect is a popular subject of study in the affective computing community [98]. Figure 1.1 displays the main steps that compose a traditional emotion/mental state recognition system. Being a wide, multidisciplinary domain of research, it faces different kinds of issues. Each step faces specific challenges:

- **definition:** The definition of many mental states are still debated. For instance, Andrews *et al.* have proposed to rename the generalized anxiety disorder as the generalized worry disorder [5]. Panksepp and Russell discussed the categorical and dimensional models of affect in [136]. Regarding Social Signal Processing (SSP), the definitions of several concepts such as synchrony and engagement are still being refined [33, 108]. This lack of

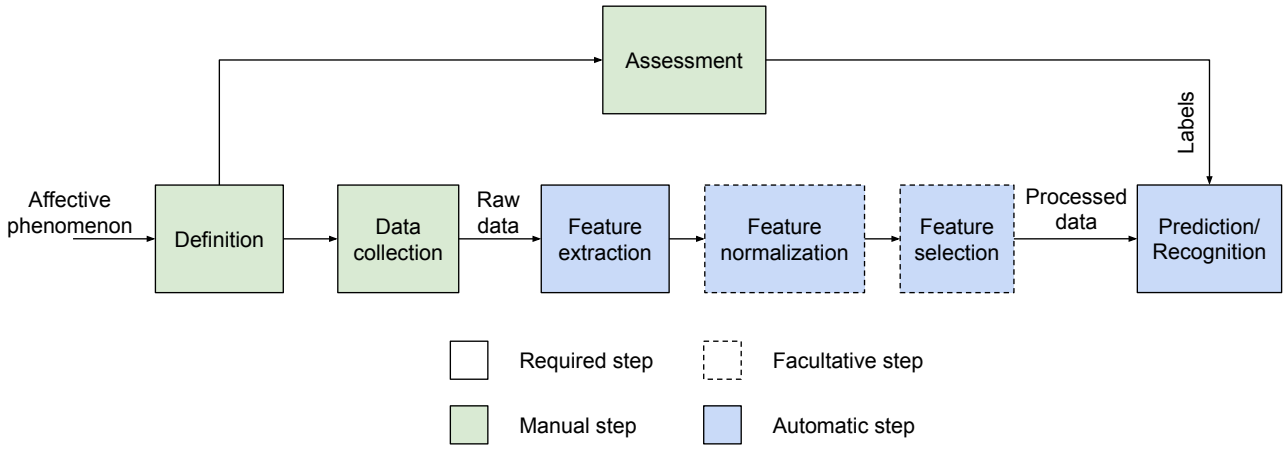

Figure 1.1: Summary of the different steps that compose an usual affect recognition system.

definition is an important issue since it impacts how a mental state can be elicited and assessed.

- **data collection:** Collecting enough data of good quality is a required step to obtain robust analysis and conclusions. There are 3 common ways to obtain this data: designing an emotion/mental state elicitation procedure, using actors and collecting data in naturalistic settings. However, each of these 3 solutions faces some issues. Regarding elicitation procedures, the biggest challenge is to design an experiment that limits the impacts of the "Hawthorne effect" [1]. This effect describes how experimental subjects may modify their behaviour because of the presence of an observer. In addition, there are also ethical limits with elicitation procedures, especially when working with negative emotions/mental states. Regarding the usage of actors, there are concerns about the quality of the collected data and about the generalization power of the models built from these data. Moreover, it can only be used for certain modalities such as speech and body language, but is unrelevant with physiology. Finally, regarding data collection in naturalistic settings, real-lived emotion can be rare and difficult to assess in an accurate way. In addition, the sensors used to collect signals must be as unobtrusive as possible or it may very well change the subject's behaviour as in the "Hawthorne effect".
- **data annotation:** Unlike several pattern recognition applications such as character recognition or face detection, there is often no consensual way to assess and annotate an affective phenomenon, as it is closely linked to its definition. Thus, although the quality

of assessment is as important as the quality of data to build robust models for emotion recognition, researchers have to choose among a wide variety of assessments strategies: self-assessment [14], external perception [10], physiological markers [11], etc. This heterogeneity also impacts the validity of comparing methods among them.

- **feature extraction and selection:** Emotion expression in human beings is highly multimodal [130], as emotions may be manifested through facial expressions, gestures, physiology and/or speech. Thus, it becomes complex to extract and select all the necessary signals and features to build an effective recognition model.
- **feature normalization:** In addition to being highly multimodal, emotion expression is also highly person-dependent. Indeed, people may express the same emotion in very different ways. However, human beings are still able to recognize these emotions despite these interindividual differences and the irrelevant information they bring. Regarding automatic recognition systems, feature normalization has been commonly used to try to limit the impact of interindividual differences. However, the effectiveness of these normalization methods are dependent of the features and of the assessments considered. Thus, researchers must carefully chose how to normalize their data.
- **automatic prediction/recognition:** The automatic recognition of emotion/mental states faces specific issues compared to other pattern recognition problems. First, the features extracted are often multimodal as we have discussed before. Thus, the machine learning algorithms must be able to use features that may have different distributions and temporal dynamics. Then, the labels used in the training phase are often noisy and uncertain. Machine learning must therefore also be able to take into account this uncertainty in order to build models with good generalization properties.

1.3 Contributions

In this work, we made contributions for several of the steps that compose an usual emotion/mental state recognition system. We strongly believe that all these steps are closely linked to each other and that one has to take an integrative approach in order to make significant improvements. For instance, we will see in Chapter 5 that choosing the best feature normaliz-

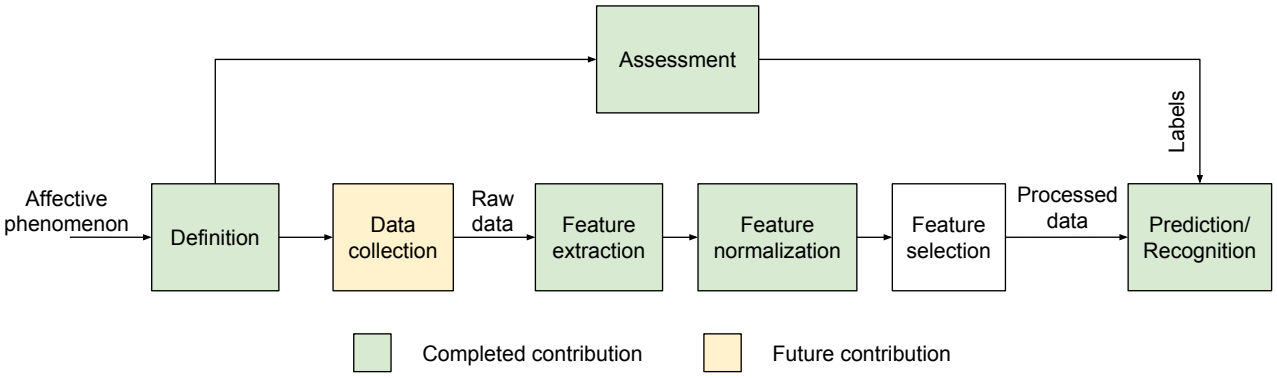

Figure 1.2: Summary of contributions.

ation method depends on the way stress is assessed. Figure 1.2 summarizes the contributions made in this work. This document present these contributions as follows:

- **Definition and assement:** In Chapter 2, we review and summarize how 3 perspectives - the biological perspective, the phenomenological perspective and the behavioural perspective - define and assess stress. We also review several automatic stress detection frameworks.
- **Data collection:** In Chapter 3, we describe how we collected behavioural data as well as 3 different assessments for 44 people. We also collected physiological signals for 25 of them. Our objective is to make this database public in a close future. In this chapter, we also study the association between the 3 collected assessments and discuss the results.
- **Feature extraction:** In Chapter 4, we describe how to extract original body language features for stress detection such as finger rubbing or periods of high body activity. We also describe the facial and physiological features that we extracted.
- **Feature normalization:** In Chapter 5, we evaluate 5 different feature normalization methods on two different stress assessments. We show that the efficiency of these methods is highly dependent on the way stress is assessed.
- **Prediction/recognition:** In Chapter 6, we propose a methodology to evaluate features from multiple potentially biased assessments in order to obtain more robust findings. We use this methodology to evaluate the relevance of 101 behavioural and physiological features for stress detection. Also, we propose in Chapter 7 adaptations for 4 classic machine

learning algorithms so that they can better handle labels collected through crowdsourcing.

Finally, we give a concluding discussion in Chapter 8. We summarize the thesis achievements and we discuss the applications and perspectives of this work.

Chapter 2

State of the art

2.1 Background

Stress is a complex phenomenon that impacts the body and the mind at several levels. Short-term and long-term stress exposure affects digestive functions [91], blood volume pressure [22], skin conditions [43], eating habits [2], performance [84], decision making [63] and health in general [27]. It is also widely considered as one of the biggest issue of western culture lifestyle, especially at work. Indeed, several papers investigate how working conditions may induce stress and/or propose strategies to reduce job related stress [30, 68, 74, 88].

Being both complex and omnipresent in our society, stress has been a popular topic of research. It has been studied for a long time from different perspectives. In this chapter, we first present how three different fields of research define and assess stress: the biological perspective focuses on the impact of stress on the body, the phenomenological perspective on the impact on the mind and the behavioural perspective on the impact on behaviour. Then, we present previous automatic stress detection systems by describing how these works faced some of the challenges presented in Section 1.2. We present the stimuli used for stress elicitation, how stress is assessed, the collected signals and the machine learning methods used in these works.

2.2 Stress definition

Although researchers have studied the topic for more than a century, the stress definition is still debated [70] and can be studied from different perspectives. In this thesis, we focus on 3

perspectives: the biological perspective, the phenomenological perspective and the behavioural perspective.

2.2.1 Biological perspective

The biological perspective aims at understanding how the body responds to a stressful stimulus. It was pioneered by Hans Selye [111], who defined stress as the non-specific neuroendocrine response of the body to a demand placed on it, such as extreme temperatures [111, 122]. The body responds to a stressful stimulus by the activation of the hypothalamo-pituitary-adrenal (HPA) pathway and the autonomic nervous system (ANS) that mediates the general adaptation syndrome [113]. After the stimulus, a neuroendocrine chain reaction begins in the brain. Recent neuroimaging studies support evidence of the major implication of some cerebral structures with a multiroad processing system of stress [75, 97]. At a peripheral level, adrenal glands respond by the release of epinephrin and cortisol into the bloodstream with an effect on cardiovascular, musculoskeletal, gastrointestinal, nervous and endocrine systems.

This physiological cascade can be measured through salivary or blood sampling with biomarkers such as cortisol. It can also be measured with wearable sensors [38] via valuable signals, such as skin conductance [58] or heart-rate variability (HRV) measures [123, 126]. Using filtering techniques, the sympathovagal balance can be directly calculated as the ratio of low and high frequencies of HRV [114]. Previous experimental studies suggest that the stress response is linked to a modulation in spectral density of the low frequency band of HRV (HRV-LF) [92].

2.2.2 Phenomenological perspective

The phenomenological perspective considers that self-perception is the key aspect. This vision has been supported within the Cannon-Bard theory: the authors state that stress can occur even when the body changes are not present because the physiological response of the body is more slowly recognized by the brain compared to its function to release an emotional response [19, 31]. A major contribution to the field of research on stress was described within Lazarus' theory of cognitive appraisal: stress is a two-way process which includes both the stressor and the individual assessment of resources required to minimize, tolerate or eradicate the stressor and the stress it produces. Experimental studies confirmed recently that stress experience is

moderated by the ability of a human subject to feel his body signals such as the heartbeat [110]. Lazarus states that *"Stress occurs when an individual perceives that the demands of an external situations are beyond his or her ability to cope with them"* [72, 73].

Since this definition focuses mainly on individual perception, stress can be measured by questionnaires [28, 86, 87], Likert and visual analogue scales [79].

2.2.3 Behavioural perspective

The behavioural perspective investigates the impact of stress on human and animal behaviour both at individual and group levels [127, 90]. Both transfer of ethological research to human behaviour and social signal processing lead researchers to a promising approach of behavioral measure of stress in non human primates [127] and more recently in human subjects [90, 131]. Engaging in displacement behaviors such as scratching, face touching and lip biting have been associated with stressfull experiences and may give more valuable information about the subject's emotional state than verbal statements and verbal expressions [127]. Authors suggest that these behaviors could impair cognitive performance by "cutting-off" attention temporarily from stressful or threatening stimuli. This short term diversion of attention could reduce the ability to deal with a mentally challenging or stressful task [23, 89].

In this perspective, behavior characteristics do not infer internal subjective feelings but are used as external marker for behavior adaptation. Therefore, stress can be assessed on the basis of behavior modifications (such as the appearance of certain gestures [77] or voice modification [138]) when a subject is exposed to a stressor.

2.3 Automatic stress detection

In the last decade, several works have studied the feasibility of automatic stress detection. Table 2.1 displays the papers we have reviewed for this thesis. In this section, we present how these works faced the challenges listed in Section 1.2: how stress is elicited and assessed, which features are extracted and how stress is detected.

Study	Stimulus	Stress assessment	Signals	ML algorithms
Ahmed *et al.* [3]	Memory task + Stroop test + Dual task + Mirror tracing task + Public speaking task	Experimental conditions + self-assessment + assessment from respiration	HR, GSR and respiration	Generalized estimating equation
Barreto *et al.* [6]	Stroop test + noise exposure	Experimental conditions	BVP, GSR, skin temperature and pupil diameter	Naïve Bayes + Decision tree + SVM
Chen *et al.* [25]	Trier Social Stress Test + memory task	Assessment from cortisol	Tissue oxygen saturation	Binary classifier
Fernandez *et al.* [39]	Mental arithmetic task while driving	Experimental conditions	Speech	Bayesian network + Mixture of Hidden markov models
Gao *et al.* [42]	Acted	Acting instructions	Facial expressions	SVM
Giakoumis *et al.* [45]	Stroop test	Assessment from GSR + self-assessment	GSR, ECG, body movement, head position, posture and occurence of specific gestures	Linear discriminant analysis classifier
Healey *et al.* [54]	Driving in several conditions	Experimental conditions	ECG, EMG, GSR and respiration	Linear discriminant function
Lefter *et al.* [77]	Acted	Assessment from external observers	Speech, movement and gestures valence and arousal	Bayes Net
Lefter *et al.* [78]	Naturalistic settings	NA	Speech	Gaussian mixture model + SVM
Plarre *et al.* [100]	Public speaking task + mental arithmetic task + cold pressor	Experimental conditions	ECG, respiration and self-reports	Decision Tree + SVM
Sharma *et al.* [115]	Video exposure	Experimental conditions	facial data in thermal and visible spectrums	SVM
Shi *et al.* [116]	Public speaking task + Mental arithmetic task + cold pressor	Self-assessment	HR, ECG, respiration, GSR, skin temperature	SVM
Tartarisco *et al.* [124]	Naturalistic settings	Assessment from expert clinicians	ECG and 3–axis accelerometer	Neural network + Fuzzy logic rules
Wijsman *et al.* [133]	Mental arithmetic + logical puzzle + memory tasks	Self-assessment	HR, ECG, EMG and respiration	Linear Bayes normal classifier + quadratic Bayes normal classifier + KNN + Fisher's least square linear classifier
Zhou *et al.* [138]	Riding a roller-coaster + acted	Experimental conditions + acting instructions	Speech	HMM

Table 2.1: Stimuli, stress annotations, signals and machine learning (ML) algorithms for some automatic stress detection systems.

2.3.1 Stress elicitation

One of the most influential works that have been done regarding stress elicitation is the meta-review of Dickerson and Kemeny [35]. They reviewed 208 studies on the impact of psychological stressors on cortisol response, which is correlated with stress as reported in Section 2.2.1. They conclude that the nature of the stressor has a major impact on the amplitude of the cortisol response. Among 4 categories of stressor - cognitive tasks, public speaking, noise exposure and emotion induction - only the categories cognitive tasks and public speaking are associated with an increased cortisol level. The most effective stressors are those that combine cognitive load and public speaking. The authors conclude that the possibility of being negatively judged by others on task performance is an effective stressor.

Given these findings, it is not surprising to see that most of the stimuli used in the reviewed papers are based on cognitive tasks and/or social evaluation. In [25], Chen *et al.* explicitly designed their experimental stressor according to Dickerson and Kemeny's work. They used a modified version of the Trier Social Stress Test (TSST), which was first introduced by Kirschbaum *et al.* in [65]. This test is composed of 2 tasks: a mental arithmetic task and a public speaking task. In their modified version, Chen *et al.* added a memory task and social-evaluative characteristics. In other works, exercises such as mental arithmetic [39, 100, 116, 133], memory task [3, 133] and stroop test [6, 45] are often used to induce cognitive load. Two works also used driving as a cognitive task [39, 54], as their objective was to study driver's stress. Public speaking tasks are also a popular mean to combine cognitive load and social evaluation [3, 100, 116]. In these tasks, the subject is usually asked to prepare a speech in a limited amount of time, and then to deliver the presentation in front of a small audience.

However, other works used different kinds of stressors than those recommended by Dickerson and Kemeny. In addition to cognitive tasks, Plarre *et al.* [100] and Shi *et al.* [116] also used a cold pressor stressor [100, 116]. This was done in order to study both psychological and physical stress. Sharma *et al.* used exposure to videos with stressful content (suspense with jumpy music) as the stimulus [115]. In [138], Zhou *et al.* used the SUSAS database (Speech Under Simulated and Actual Stress) [52] in which part of the data were obtained from subjects

riding a roller-coaster.

Two works used data obtained in naturalistic settings [78, 124]. In [78], Lefter *et al.* used genuine recordings from emergency call-centres from the South-African database. In [124], Tartarisco *et al.* designed a mobile architecture to monitor stress related physiological signals.

Finally, several works use actors instead of stress elicitation experiments or naturalistic settings [42, 77, 138]. Using actors has advantages and drawbacks, as explained in [77]:

+ It is more ethical than elicitation procedures.
+ Some real-lived emotions are rare, making it difficult to gather enough data.
− Using actors is irrelevant when studying the impact of stress (or any other emotion/mental state) on physiology.
− Acted data may be too prototypical to be useful for pratical applications.
− Using professional actors is sometimes necessary in order to ensure a better quality of data.

In [138], Zhou *et al.* also used the simulated data from SUSAS database. In [42], Gao *et al.* asked subjects to make facial expressions associated with the six basic emotions (i.e. anger, disgust, fear, happiness, sadness and surprise) and the neutral expression. In [77], Lefter *et al.* asked professional actors to improvise interactions at a service desk.

2.3.2 Stress assessment

As seen in Section 2.2, stress can be assessed in several ways depending on the chosen definition: using questionnaires [28, 86, 87], biomarkers [58, 123, 126] or through changes in one's behaviour [90, 131]. However, the assessment choice will greatly determine the findings of stress prediction models. Lutchyn *et al.* suggest that, regarding automatic stress detection, *"inconsistent results reported in some areas of research can be partially explained by the choice of measurements that capture different manifestations of affective phenomena, or focus on different elements of affective processes"* [85]. Thus, it might be necessary to consider several assessments in order to analyze the results of stress prediction models in a more comprehensive way.

However, most of the reviewed papers use only one stress assessment strategy among a wide variety of them:

- **Experimental conditions.** In several works [6, 39, 42, 54, 100, 115, 138], stress presence is inferred from experimental conditions. Thus, data obtained during a condition considered as stressful is labelled as "stressed" while the others are labelled as "relax". This method is risky since it does not take into account one's appraisal of the stressor, although it has been shown to be an important aspect of stress [70, 73, 85, 125].
- **Self-assessment.** In [116, 133], stress is assessed using self-reports and questionnaires.
- **Assessment from external observers.** Stress is assessed by external observers who look at different kinds of information. Thus, this assessment allows to study how stressed one appears, but not how stressed one feels. The external annotators can either be experts [77, 124] or novices (see Chapter 7). In [77], two experts judge whether the actor appeared stressed by looking at her behaviour (speech + body language). In [124], expert clinicians provide stress assessments by looking at physiological data.
- **Acting instructions.** In [42], stress is inferred from instructions regarding facial expressions. If the subject is asked to act angry or disgusted, the corresponding data is labelled as "stressed". In [138], part of the data used in the experiment is also labelled from acting instructions.

It is however noteworthy that a multi-assessment approach has been adopted in some recent works [3, 45]. In [3], Ahmed *et al.* propose a method to relabel stress/relax examples using respiration signals. Then, they compare the performances of 3 models training using 3 different assessments: experimental conditions (i.e. considering that some conditions are always stressful and some are always relaxing), self-assessment, and re-assessment from respiration. The last model obtains the best results. Giakoumis *et al.* used self-assement and assessment from GSR for their classification experiments [45].

2.3.3 Feature extraction

Automatic stress detection systems used to mainly extract 2 categories of features: physiological features and speech features. Using physiological features is intuitive since stress has

many measurable effects on the body, as discussed in Section 2.2.1. Thus, many works use features extracted from cardiac signals (such as blood volume pulse, heart-rate, electrocardiography, etc.) [3, 6, 45, 54, 100, 116, 124, 133], respiration [3, 54, 100, 116, 133], electromyography [54, 133], skin temperature [6, 116] and skin conductance [3, 6, 45, 54, 116]. Although physiological signals are usually collected through wearable sensors, some works have used hyperspectral imaging techniques to obtain signals such as the tissue oxygen saturation [25] and skin temperature [115] in an unobtrusive way.

Speech has also been used in early works on stress detection. Indeed, as explained in [138], stress introduces variabilities on the acoustic speech signal that reduce speech recognition accuracy. In [138], Zhou *et al.* studied how features based on the Teager energy operator [60] performed for stress detection compared to traditional features such as pitch and mel-frequency cepstrum coefficients features. The Teager energy operator is also used by Fernandez *et al.* in [39]. In this paper, the authors worked on modeling driver's speech under stress in order to improve the interaction with a speech interface. In [78], Lefter *et al.* use prosodic and spectral features to detect stress in emergency calls.

Recent works have also extracted visual features from body language [45, 77] and facial expressions [42]. In [45], the authors show that using behavioural features such as body movement or head position enhances the performance of traditional physiology-based stress detection systems. In [77], Lefter *et al.* use low-level features extracted from speech and gestures to detect intermediate level variables such as speech valence and arousal, gesture valence and arousal, etc. These intermediate variables are then used to detect stress. Gao *et al.* [42] extract global and local facial descriptors such as SIFT descriptors [83] to detect the 6 basic emotions. Anger and disgust are then merged as a single stress class.

Overall, features extracted from physiology and speech have provided good results for stress detection for more than a decade. More recently, features extracted from body language and facial also provided promising results. However, these findings greatly depend on the way stress is assessed. This point is discussed in more depth in Chapter 6.

2.3.4 Stress detection

As seen in Section 2.3.2, Lutchyn *et al.* suggest in [85] that inconsistent results may be partially explained by the different methods used to assess stress. The heterogeneity of evaluation processes of stress detection systems may also contribute to this phenomenon. Indeed, independently of the machine learning algorithm used for prediction, evaluation processes can vary in many ways. These factors can impact the performance of stress detection models and the findings associated with them:

- **Prediction resolution.** Machine learning algorithms can be applied for different kinds of problems with different levels of difficulty:
  - Binary classification problems. Examples are from one of 2 classes: positive and negative (stress and non-stress in the case of stress detection). The objective is to predict the correct class of testing examples. The majority of the reviewed papers present frameworks for this kind of problem [3, 6, 25, 42, 45, 78, 100, 115, 116, 133, 138].
  - Multi-class classification problems. Examples are from one of n classes, with $n > 2$. The objective is to predict the correct class of testing examples. Several reviewed papers present frameworks for this kind of problem. In [39], Fernandez *et al.* try to predict the driving conditions of each example: slow driving and slow-paced arithmetic task (SS), slow driving and fast-paced arithmetic task (SF), fast driving and slow-paced arithmetic task (FS) and fast driving and fast-paced arithmetic task (FF). Healey *et al.* and Lefter *et al.* try to predict 3 different stress levels [54, 78] and Tartarisco *et al.* try to predict 4 different stress levels [124].
  - Regression problems. Examples are associated with a continuous value. The objective is to estimate the associated value of testing examples. None of the reviewed papers present frameworks for this kind of problem.
- **Performance metrics.** There are several ways to measure and present the results obtained by classification systems. Since most of the reviewed papers describe frameworks for binary classification problems, we present performance metric formulas for this kind

	Predicted: Stress	Predicted: Non-Stress
Actual: Stress	TP	FN
Actual: Non-Stress	FP	TN

Table 2.2: Example of a confusion matrix in the case of binary classification. TP means true positives, FN means false negatives, FP means false positives and TN means true negatives.

of problem. To do so, we will use Table 2.2, which is an example of a confusion matrix in the case of binary classification.

- Classification accuracy, which is used in most of reviewed works [3, 6, 39, 45, 54, 77, 100, 115, 124, 133, 138], gives the proportion of correctly classified examples.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- Weighted classification accuracy gives the average of correctly classified examples per class [45, 54, 77]:

$$weighted\_accuracy = 0.5 \times \left(\frac{TP}{TP + FN} + \frac{TN}{FP + TN} \right)$$

- Precision and recall are usually used on the positive class (i.e. the stress class for stress detection systems) [116]. The precision (also called the True Positive Rate or TPR) gives the proportion of examples classified as positive which are correctly classified:

$$precision = \frac{TP}{TP + FP}$$

  Recall gives the proportion of postive examples which are correctly classified:

$$recall = \frac{TP}{TP + FN}$$

- ROC curve plots the TPR against the False Positive Rate (or FPR = $\frac{FP}{FP+TN}$) for various decision threshold values [25, 100].
- The F1-score or F-measure is the harmonic mean of precision and recall [42]. Thus, it is also usually used on the positive class:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

- **Subject-dependent models vs subject-independent models.** In most of the reviewed papers, and in many works in the affective computing community, a subject provides several examples. Indeed, subjects usually face several experimental conditions (resting and stressful conditions [3], an increasing complexity for the cognitive task [39], etc.), resulting in several examples that are going to be treated separately in the classification task. However, these examples are not independent since they are collected from the same subject. In subject-dependent models, examples from a same subject are used both to train the model and to evaluate it [6, 25, 39, 42, 45, 54, 77, 100, 133]. Thus, these models cannot be used to detect stress in any subject, but only for one or some subjects. In subject independent models, examples from a same subject are either used to train the model or to evaluate it, but not both [3, 39, 78, 115, 116, 124]. In this case, these models can be used on new subjects.

Regarding the ML algorithms in the reviewed papers, the most used one is the Support Vector Machine (or SVM) [6, 42, 78, 100, 115, 116]. SVM is a traditional machine learning framework for binary classification [12]. It aims at finding an optimal hyperplane that separates by the widest margin points from 2 classes. Points can be projected into a transformed feature space in order to perform nonlinear classification. Other traditional ML algorithms have been applied: Hidden Markov Models (or HMM) [39, 138], decision trees [6, 100], Bayesian networks [39, 77], etc.

2.4 Conclusion

Stress is a complex phenomenon that has serious impacts on health and on productivity at work. It has been studied by several fields of research which gave different definitions and/or ways to assess stress:

- For the biological perspective, stress is the physiological response of the body to a stressful stimulus. It can be assessed through biomarkers such as hormone levels or physiological changes.
- For the phenomenological perspective, stress is a process that occurs when an individual

perceives a stimulus as threatening and that demands of the situation are beyond her abilities. This perspective focuses on individual perception and assess stress mainly using questionnaires.

- For the behavioural perspective, stress can be assessed through changes in the bahaviour.

In the last decade, several works have presented frameworks for automatic stress detection. In these works, laboratory experiments are usually used to collect data in stressful situations. Previous works showed that the most effective way to elicit stress during laboratory experiments is to combine cognitive load and social evaluation.

We have presented how stress was assessed in the reviewed papers. We have seen that various methods are applied: assessment from physiology, self-assessment, assessment from experimental conditions, assessment from external observers, ... The assessment choice has a huge impact on how a model is trained, on its results and thus on how these results are interpreted. The fact that stress is assessed in such various ways may partially explain inconsistent results. We have seen that one way to limit this phenomenon is to consider multiple assessments.

We have seen that stress detection systems used to extract features mainly from physiology and speech. Lately, features extracted from body language and facial expressions have been included in multimodal frameworks. One work concludes that adding behavioural features enhances the performance of traditional physiology-based frameworks.

Finally, we have discussed how evaluation processes may vary from one work to another. We have described different prediction problems (binary classification problems, multi-class classification problems and regression problems) and different performance metrics. We also discussed the difference between subject-dependent models and subject-independent models.

Chapter 3

Acquisition of stress data with multiple assessments

3.1 Introduction

One of the main conclusions of Chapter 2 is stress detection systems are heterogeneous in many ways. The heterogeneity of assessment methods is especially important, as it may explain the inconsistent findings reported in the literature [85]. One of the contributions of this thesis is that we study the stress phenomenon in a more comprehensive way by considering the results obtained from different annotations. Figure 3.1 is an extension of Brunswik Lens and summarizes how we address the issue of the annotation choice. The Brunswik Lens [16] is used in the affective computing literature to illustrate the difference between self-assessment and external assessment for phenomena such as personality [129] and stress [77]. We extend it by adding the assessment provided by a physiology expert. Thus, we annotate stress in 3 different ways:

- Following the behavioural perspective, we gather external observer assessments (EOA) using a crowdsourcing platform.
- Following the phenomenological perspective, we ask the subject to provide her self-assessment (SA).
- Following the biological perspective, a physiology expert assesses the presence of stress from the percentage of low frequencies in the heart rate variability (PEA).

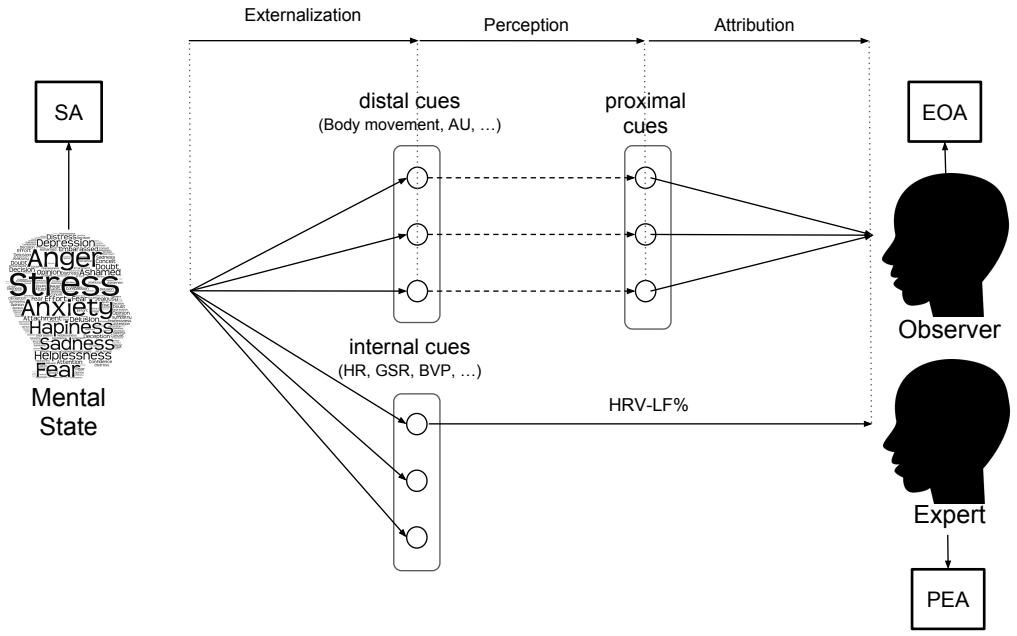

Figure 3.1: Data is annotated in 3 different ways. First, following the phenomenological perspective, we ask the subject to provide her Self-Assessment (SA). Then, following the behavioural perspective, we ask external observers (recruited using a crowdsourcing platform) assessments (EOA). Finally, following the biological perspective, a physiology expert assesses the presence of stress from the percentage of low frequencies in the heart rate variability (PEA).

Designing a new experimental protocol was necessary to collect the required data to analyze stress from multiple assessments, since to our knowledge, no available dataset provided these assessments. Giraud *et al.* presented a multimodal stress corpus in [46]. A public speaking task is used as the stimulus. Physiological and behavioural data are collected along with self-assessed stress, personality and coping strategies. Koldijk *et al.* described the SWELL dataset in [67]. They collected behavioural and physiological data from 25 subjects while they were performing typical work tasks (making presentations, reading e-mails, etc.). Again, self-assessment was used as ground truth.

In this chapter, we first present the experiment we designed to obtain behavioural and physiological data in a stressful situation. We present the stimulus, the setup, the post-experiment questionnaires and the acquired datasets: *Dataset-44* and *Dataset-21*. Then, we describe the 3 collected stress assessments: the External Observers Assessment (EOA), the Self-Assessment (SA) and the Physiology Expert Assessment (PEA).

3.2 Experimental stressor

3.2.1 Stimulus

As explained in Section 2.3.1, Dickerson and Kemeny state in [35] that the best way to increase the cortisol level of a subject is to induce cognitive load while socially evaluating her. Based on this work, we designed a time-constrained mental arithmetic test as the stress-induction stimulus of the experiment (Figure 3.2). Subjects were told that the objective of the experiment is to estimate their developmental age and to correlate it with their academic and professional careers. It made them believe that they were socially evaluated while keeping hidden the stress induction aim of the experiment. This way, stress induction occurred as naturally as possible.

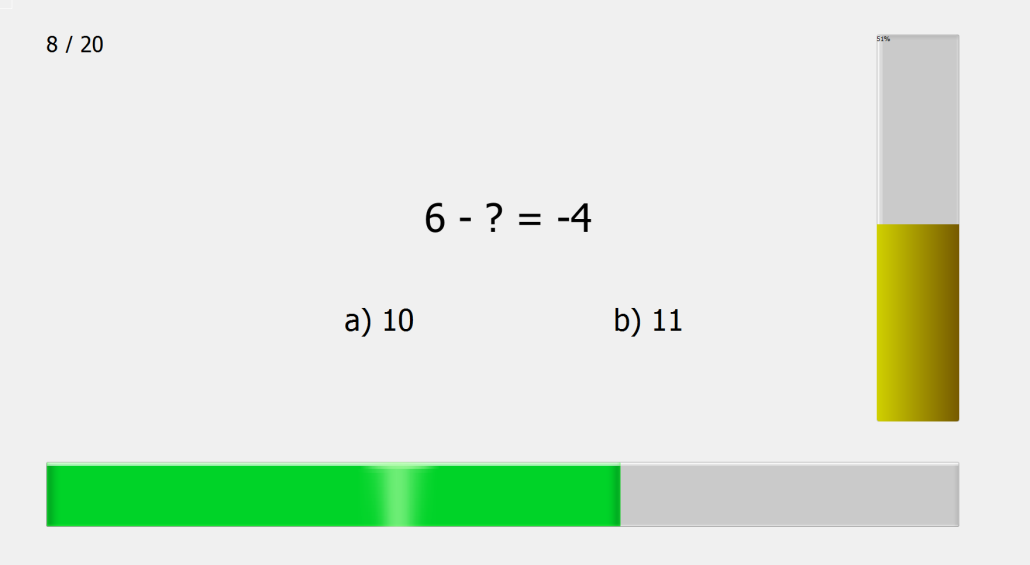

Figure 3.2: Screenshot of the test software used for the study. The question asked is shown in the middle of the screen. The two possible answers are below the question. At the bottom, the remaining time is displayed using a progress bar. On the right, the color of the score bar provides a feedback regarding the performance of the subject: green means "above average", yellow means "average" and red means "below average".

In our protocol, the subject is first briefed about the fake objective of the experiment and asked to sign the consent form and the release waiver. We also informed the subject that she could stop the experiment at anytime. Then, the physiological sensors are installed, and the subject starts taking the test. The test is composed of 6 steps of increasing difficulty. There is a break period of 5 seconds between 2 steps. The subject is told that both quickness and correctness of her answers are taken into account to compute her score. In reality, the values of

the score bar are set in advance. It displays an "above average" score at the beginning, so that the participant finds the test easy enough and feels like she should succeed. Then, the score drops to "average" and "below average" levels, giving the participant the feeling she is actually failing. Overall, the score bar, the presence of the 2 people who ran the experiment and the fake objective induce the feeling of social evaluation while the questions and the time bar induce cognitive load. Once the test is finished, the real objective is revealed and the experiment is debriefed.

3.2.2 Setup

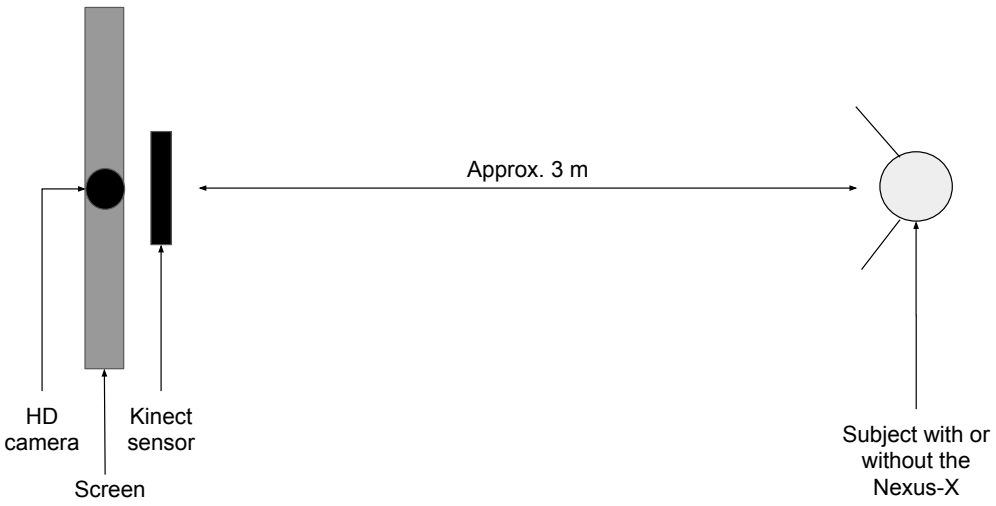

Figure 3.3: Setup of the experiment

Figure 3.3 shows the setup of the experiment. The subject is standing approximately 3 meters away from the screen where the test is displayed. Video and skeleton data were recorded using a Microsoft Kinect. Since the resolution (640×480 pixels) of the video recorded by the Kinect is too low for an accurate facial expression analysis, we also recorded video data of the subject's face using the optic zoom of a high definition (1440×1080 pixels) camera. Physiological data was recorded with a Nexus-10 portable device (MindMedia B.V., The Netherlands) with a measurement of EMG, GSR, skin temperature, respiration, BVP and HR.

3.2.3 Post-experiment questionnaires

After the experiment, we present 4 questionnaires: a personality test, the State-Trait Anxiety Inventory (STAI), questions about the impact of the experiment on stress and questions about self-assessment of stress. The description of self-assessment will be presented in Section 3.3. In this section, we present the questions and we compute the answer distribution of the first 3 questionnaires.

Personality test

The personality test we used is the 10-item version of the Big-Five Inventory [102]. This test aims at describing human personality by evaluating 5 factors: openness, conscientiousness, extraversion, agreeableness and neuroticism. Each factor is represented by 2 questions. All 10 questions are Likert-scaled questions, ranging from 1 to 5. To evaluate a factor, one sums the values of the 2 related questions. Thus, scores for each factor range from 2 to 10. Figure 3.4 presents the distribution of each factor among the subjects. It is noteworthy that values of neuroticism are fairly well distributed. It is important since the relation between stress and neuroticism have been studied in several works [34, 41, 93]. Thus, it seems that there is no bias regarding neuroticism among the subjects of our experiment.

The State-Trait Anxiety Inventory (STAI)

The State-Trait Anxiety Inventory is a questionnaire designed by Charles Spielberger usually composed of 40 sef-report items aiming at measuring anxiety [120]. It evaluates two types of anxiety: state anxiety and trait anxiety. State anxiety corresponds to the anxiety felt at a specific moment, while trait anxiety corresponds to one's relatively enduring disposition to feel anxious or not. In our experiment, we only evaluate trait anxiety. Consequently, the questionnaire is composed of 20 Likert-scaled questions, ranging from 1 to 4. Thus, STAI scores range from 20 to 80. Figure 3.5 presents the distribution of STAI scores among the subjects.

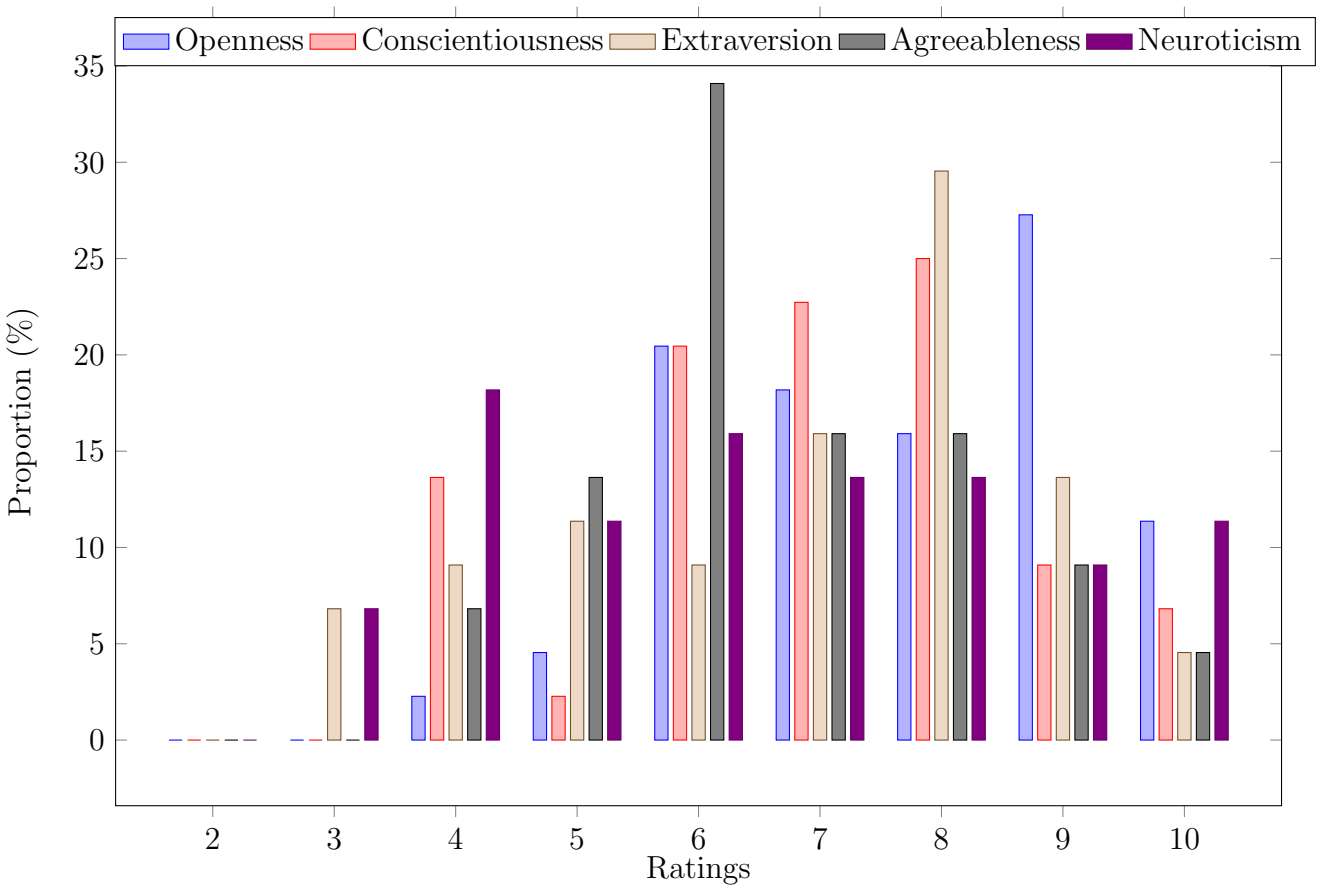

Figure 3.4: Values distribution for each personality traits assessed with the Big-Five. Best viewed in color.

Impact of cognitive load and social evaluation on stress

As explained in Section 3.2.1, we decided to elicit stress by combining cognitive load and social evaluation. In the experiment we designed, question complexity and the allowed answer time are used to induce cognitive load, while the score and the presence of the 2 people running the experiment are used to induce social evaluation. During the questionnaire, we ask 4 Likert-scaled questions (1 - 5) about the impact of each of these 4 elements on stress. Figure 3.6 presents the mean values of the self-reported rating of the impact of each element on stress. We can see that all 4 elements induce relatively high level of stress, since their mean values are all above 3.5. According to the subjects' answers, the most stressful element is the limited answer time for each question. It is not surprising since the relation between time pressure, decision making and stress has been studied and accepted for a long time [96, 121]. Regarding the 3 other elements, there is no statistical significant difference among them. Overall, it appears

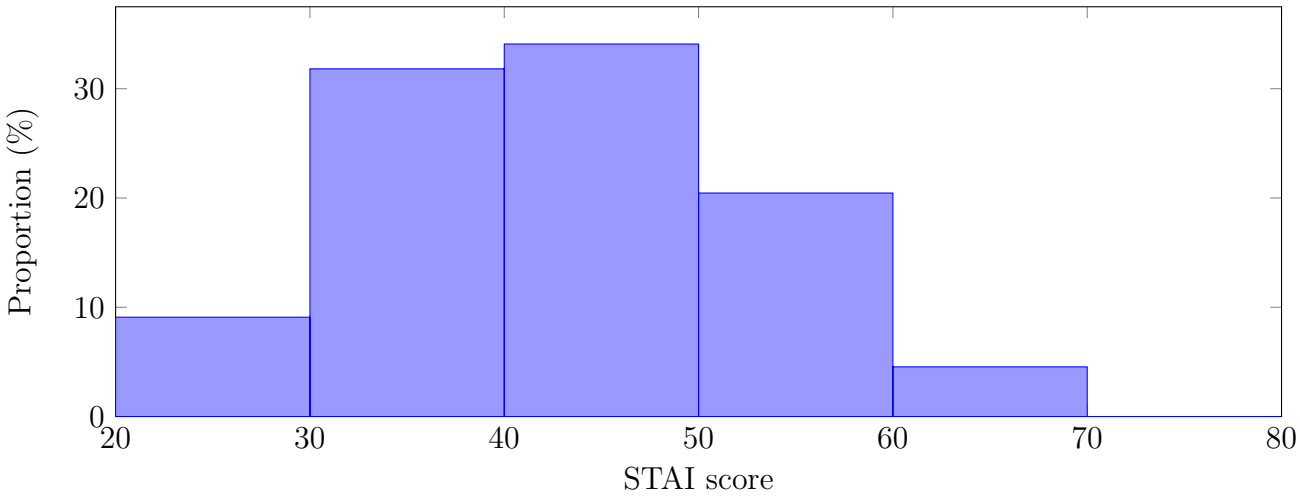

Figure 3.5: STAI value distribution.

that cognitive load induced more stress than social evaluation in our experiment. However, it is important to note that people's presence may have been more impactful if we chose to adopt a judgemental or cold attitude rather than being natural.

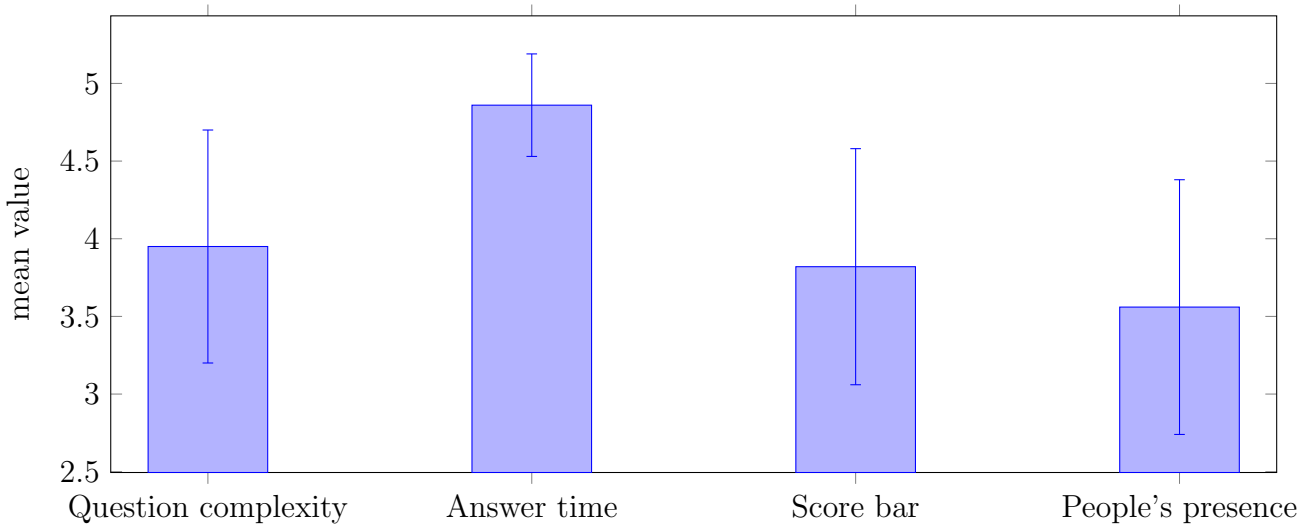

Figure 3.6: Mean value of the self-reported impact of each element of the experiment on stress.

3.2.4 Acquired Datasets

Overall, 44 people recruited among medical and computer science students participated in our experiment. However physiological signals were not recorded for all the participants. Consequently, we have 2 datasets:

- **Dataset-44** for which we only have Kinect and HD video data. All 44 participants are included in this dataset: 25 men and 19 women. On average, the subjects are 25.4 ±

3.7 years old. For each of those subjects we process each of the 6 steps independently, making a total of $6 \times 44 = 264$ examples.

- **Dataset-21** for which we have Kinect, HD video and physiological data. A subset of 21 participants are included in this dataset: 15 women and 6 men. On average, the subjects are 26.3 ± 4.6 years old. In total, we process $6 \times 21 = 126$ examples.

3.3 Description of external observer, self and physiology expert assessments of stress

As shown on Figure 3.1, each example is annotated in 3 different ways, one for each perspective we presented in Section 2.2:

- External Observers Assessment (EOA)
- Self-Assessment (SA)
- Physiology Expert Assessment (PEA, available only for Dataset-21)

In this section, we describe in detail these 3 annotations and study their correlations.

3.3.1 Description of External Observers Assessment (EOA)

We used the crowdsourcing platform CrowdFlower[1] to obtain annotations from external observers. It allows to easily obtain a large amount of annotations while providing some quality control mechanisms.

Crowdsourcing acquisition procedure

We presented the video of the body recorded by the Kinect for all 264 examples. Three questions were asked for each video (Figure 3.7):

- Do you think this person is stressed? Answers: not stressed/stressed *(Q1)*
- How stressed is the person in this video? Answers: Likert scale 1-5 *(Q2)*
- How confident are you on your ratings? Answers: Likert scale 1-5 *(Q3)*

[1] www.crowdflower.com

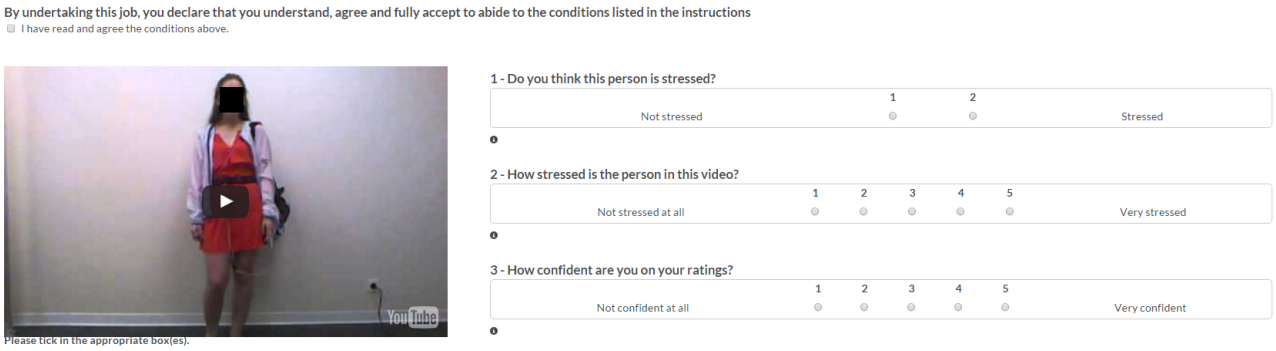

Figure 3.7: Screenshot of the CrowdFlower platform.

Regarding the instructions given to the annotators, we told them that they were shown videos of people taking a cognitive test. This was done so that they would have enough knowledge about the context to provide accurate ratings. However, we did not mention the mental arithmetic nature of the test in order to avoid social projection [106], which may have impacted the workers' answers. To obtain acceptable statistical significance, we requested 10 annotations per video.

Crowdsourcing annotation quality control

Even if, for a subjective question such as "do you think this person is stressed?", there is no such thing as a wrong answer, it is still important to try to avoid spammers and malicious workers in order to collect good quality data. To do so, we used 3 mechanisms to ensure the annotation quality.

- CrowdFlower proposes 3 categories of workers according to their performances on the platform. We have chosen workers from the highest ranked category.
- We have set the minimum amount of time a worker should take to answer the questions of one video. Since we want workers to watch the videos until the end and since the shortest video lasts 50 seconds, we have chosen this duration as the threshold. If a worker takes less time to annotate one video, her answers are discarded and she cannot work for this task anymore.
- During a pilot experiment, we selected examples which are considered prototypical videos of stressed and non-stressed people. To do so, 28 people answered the same question Q for a subset of 15 videos that we had previously selected. Then, we discarded 4 videos

that achieved an agreement rate lower than 90% for question Q. These 11 videos were used by the CrowdFlower platform to filter the workers: the platform randomly chose 5 videos to create a quizz. If a worker successfully answers Q for less than 4 videos out of 5, she is not allowed to participate in the task. In addition, CrowdFlower randomly inserts "Test Questions" among the videos. Test Questions are videos for which there is a set of acceptable answers. If a worker misses too many Test Questions, she is not allowed to provide new ratings and her previous ratings are marked as unreliable. This method is risky since we may discard honest workers who just gave their opinion, however it is also the most effective way to discard spammers who would be aware that there is a minimum amount of time to spend on each video.

Crowdworkers

259 people annotated an average of 11.90 ± 9.48 videos. Their repartition over the continents is presented in Table 3.1. We can see that most of them are from western culture, as the subjects of the experiment are. This is important since stress may be expressed and perceived differently depending on one's culture. Analysis on the impact of culture on stress expression and perception is important but is out of the scope of this thesis.

Continent	EU	SA	AS	NA	AF	OC
Number of annotators	133	52	49	24	1	1

Table 3.1: Repartition of the annotators over the continents (EU = Europe, SA = South America, AS = Asia, NA = North America, AF = Africa, OC = Oceania).

Annotations aggregation

Since we have several annotations per video, we have to use an aggregation method in order to assign a single label to each video. To do so, we chose the Honeypot method [101]. First, we remove untrustworthy annotators using answers to Test Questions. Then, we assign the majority decision to each video: if more than half of the remaining annotations are Stress answers, we assign the Stress label, otherwise we assign the Non-Stress label.

Label	Non-Stress	Stress
Dataset-44	46.2%	53.8%
Dataset-21	39.7%	60.3%

Table 3.2: EOA label distribution for both datasets.

3.3.2 Description of Self-Assessment (SA)

Self-assessment of stress was conducted during the debriefing of the experiment. The subjects answered a Likert-scaled (1-5) question about how stressed they felt during each step. To limit memory bias, they watched their own videos before providing their answers. Then, in order to obtain binary labels, we use a threshold on the stress level: Non-Stress = {1, 2} and Stress = {3, 4, 5}. This threshold has been chosen regarding the repartition of the answers to *Q2* according to the answer given for *Q1*. As shown in Figure 3.8, it appears that stress levels 1 and 2 are associated with Non-Stress, while stress levels 3, 4 and 5 are associated with Stress.

Label	Non-Stress	Stress
Dataset-44	34.1%	65.9%
Dataset-21	25.4%	74.6%

Table 3.3: SA label distribution for both datasets.

3.3.3 Description of Physiology Expert Assessment (PEA)

For PEA, presence of stress was assessed based on clinician expertise on the physiological impact of stress. We used the percentage of low frequency in the heart-rate variability (HRV-LF%) measure provided by the Nexus-10. HRV is considered to be a reliable indicator to assess the presence of physiological stress [123, 126] and its percentage of low frequency is seen as a valuable marker [11, 26, 92]. It also has the advantage to be a fast physiological marker of the activation of the HPA pathway and the ANS. Thus, it gives a fast image of the impact made by the stressor, unlike cortisol which is released with a 5 to 20 minutes delay [13]. In order to obtain binary labels, we compare the values obtained with the average of HRV-LF% over all the examples. For each example, if the HRV-LF% observed is above the computed average, then

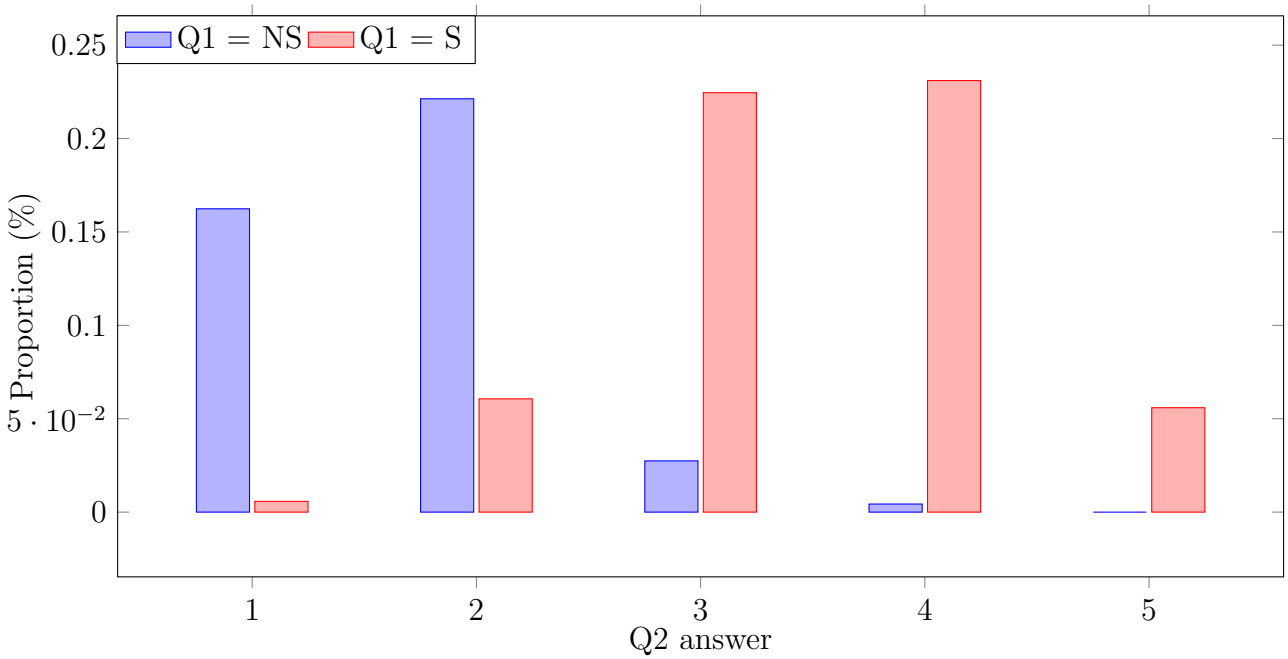

Figure 3.8: Answer distribution to *Q2* according to the answer given for *Q1* (Stress or Non-Stress). Best viewed in color.

the example is associated with the Stress label. Otherwise, it is associated with the Non-Stress label. Consequently, we obtain the distribution shown in Table 3.4.

Label	Non-Stress	Stress
Dataset-21	52.4%	47.6%

Table 3.4: PEA label distribution for Dataset-21.

3.3.4 Are PEA, SA, and EOA significantly associated ?

Assessment sets	SA×EOA	SA×PEA	EOA×PEA
Dataset-44	0.25	NA	NA
Dataset-21	0.38	-0.08	-0.07

Table 3.5: Cohen's Kappa for each combination of 2 assessment sets for both datasets.

To assess how PEA, SA, and EOA were associated, we performed two analyses. First, we calculated Cohen's Kappa to assess their agreement based on binary labels. Second we used correlation analysis based on non-binary values. Table 3.5 shows the Cohen's kappa scores for each combination of 2 assessment sets. The only scores which are considered as fair by the guidelines given by Landis *et al.* in [71] are obtained by the pair SA×EOA for both datasets.

This could be explained by the fact that the subjects were asked to watch their own videos before providing their self-assessment. Thus, they looked at the same distal cues as the external observers before judging whether they felt stressed or not. The low kappa scores obtained by PEA with SA and EOA may be explained by the differences in distribution: PEA is more balanced (Table 3.4) than SA (Table 3.3) and EOA (Table 3.2). Since the kappa scores are impacted by the choice of a specific threshold for each assessment, we also used correlation analysis on non-binary values. Table 3.6 presents the correlation coefficient between the non-binary values associated with each assessment:

- EOA: the proportion of annotators that answered "Stressed" for *Q1*.
- SA: the self-reported answers to the Likert-scaled (1-5) question asked during the debriefing of the experiment.
- PEA: the HRV-LF% values.

Assessment sets	SA×EOA	SA×PEA	EOA×PEA
Dataset-44	0.32*	NA	NA
Dataset-21	0.41*	-0.11	-0.06

Table 3.6: Correlation coefficients for each combination of 2 assessment sets for both datasets. Significant correlations ($p < 0.05$) are marked with *.

The correlation coefficients are very similar to the kappa scores: the only significant correlations are obtained by the pair SA×EOA. The correlation coefficients - 0.32 and 0.41 - indicate a modest correlation. There is no significant correlations between PEA and EOA and between PEA and SA.

Overall, the kappa scores and the correlation coefficients give similar conclusions. The lack or limited correlation found supports: (1) the idea that stress is a complex phenomenon which can be expressed through one's body, behaviour and/or mind. (2) Physiological parameters may differ in timing for stress induction compared to behavioral cues ; (3) despite the correlation between EOA and SA, it appears that the two phenomena have both common basis and separate cues. Thus, it is important to assess stress in several ways because of the diversity of its expression.

3.4 Conclusion

We have presented the stress elicitation experiment we designed, which combines cognitive load and social evaluation as recommended by Dickerson and Kemeny in [35]. We have acquired Kinect and HD video data from 44 subjects. We refer to this dataset, which is composed of 264 examples, as *Dataset-44*. We also have acquired physiological data from a subset of 21 subjects. We refer to this dataset, which is composed of 126 examples, as *Dataset-21*. For each example, stress is assessed in several ways:

- Workers from the crowdsourcing platform Crowdflower provide External Observers Assessment (EOA)
- Subjects of the experiment provide Self-Assessment (SA)
- A physiology expert provides Physiology Expert Assessment (PEA, available only for Dataset-21)

We will use these datasets and assessments to study several aspects of automatic stress detection in the following chapters of this thesis.

Chapter 4

Feature extraction for automatic stress detection

4.1 Introduction

In this chapter, we present the features we extract for stress detection. As explained in Section 2.3.3, most frameworks for automatic stress detection use features extracted from speech and/or physiological signals [6, 25, 39, 54, 116, 133, 138]. However, there are some contexts where speech and physiology are not the most suited sources of information for stress detection. In [39] and [54], stress is detected during a real-world driving-task. However, drivers do not always talk, and physiological sensors are obtrusive. In this case, cues extracted from the body language are likely to be the most convenient sources of information to assess stress since they can be captured by camera sensors.

Recent works have studied the contribution of behavioural features such as movement and posture, with various results [45, 77, 119]. Giakoumis *et al.* showed that using behavioural features enhances the performance of traditional physiology-based stress detection system [45]. Behavioural features have also been used to recognize or to synthesize someone's affective state [21, 46, 47].

In this thesis, we propose original body features extracted from Kinect data. We also extract facial and physiological features in order to study multimodal features for stress detection. Overall, we extract 101 features from 3 sources: 15 body features from the Kinect data, 24 facial features from the HD video and 62 physiological features from the signals provided by

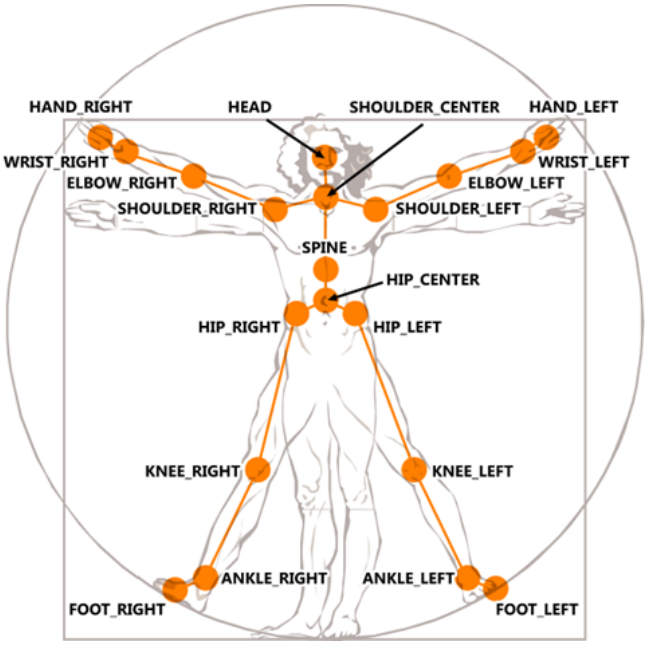

Figure 4.1: The skeleton joints extracted by the Kinect[1]

the Nexus-10. Body and facial features are presented in Table 4.1 gathered as behavioural features. Physiological features are presented in Table 4.2.

4.2 Body features

4.2.1 Quantity of Movement

The main body activity feature extracted is the Quantity of Movement (QoM). We compute it in two ways: using the RGB video (IQoM), and using the skeleton joints (SQoM) (Figure 4.1). IQoM is the number of pixels that changed between two successive frames.

$$IQoM(i) = Card(\{p_i | abs(p_i - p_{i-1}) > t\})$$

with p_i the RGB vector of pixel p in the i^{th} frame and t a threshold. SQoM is the sum of the displacements of the skeleton joints between two frames.

$$SQoM(i) = \sum_{j \in joints} \sqrt{(v_{j_i} - v_{j_{i-1}})^2}$$

with v_{j_i} the position vector of the joint j in the i^{th} frame. Each method has its advantages and drawbacks. SQoM enables us to detect slight movements in the camera axis. However, as the

[1]Image retrieved from https://msdn.microsoft.com/en-us/library/microsoft.kinect.jointtype.aspx

Kinect skeleton can be unstable during the recordings, IQoM is also used in order to extract a less noisy quantity of movement. For both IQoM and SQoM, we compute their average value over all the Kinect video frames of the protocol step. Then, in order to make these features invariant to the size of a person and to the distance between her and the camera, we normalize them with respect to the surface of the box bounding the person. We also compute the SQoM only for the head joint (HeM) and isolate its movement along the camera-axis (HeMZ).

4.2.2 Periods of high body activity

We make the hypothesis that periods of high body activity characterize an increasing uncomfortability. These periods are extracted by detecting the peaks in the IQoM signal (Figure 4.2). We use the number of periods extracted (HAPC), their average duration (HAPMD) and their average intensity (HAPMV) as features.

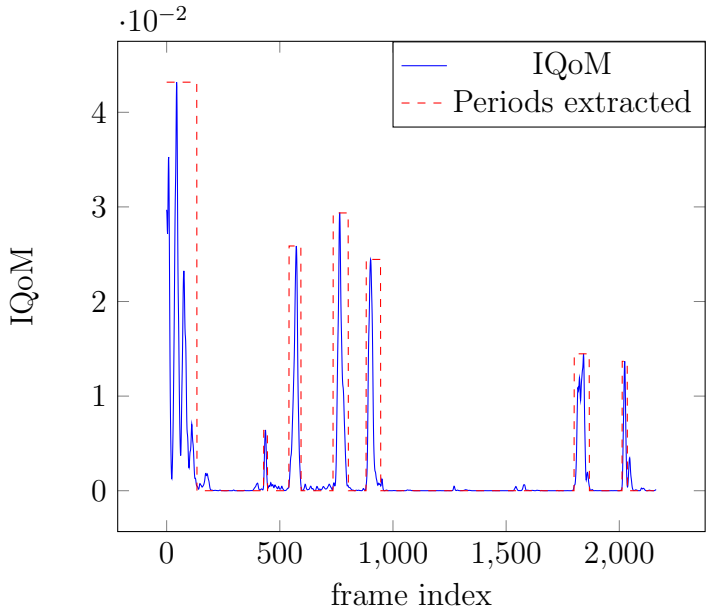

Figure 4.2: Extraction of periods of high body activity from the IQoM. Blue line: IQoM values per frame. Red dashed line: periods of high activity extracted. Best viewed in color.

4.2.3 Posture changes

As for periods of high body activity, posture changes may reveal uncomfortability. Giraud *et al.* concluded in [46] that the variability of the center of gravity displacements is related to negative emotions, such as stress. In this thesis, we use the number of posture changes (PCC) that occur during the experiment as a feature. Because of the skeleton stability issues in the

recordings, especially when the participant crosses her arms, we use the periods of high body activity described previously to extract the posture changes. For each period, we compare the first frame and the last frame by computing their difference (Figure 4.3). If the number of pixels divided by the surface of the bounding box is above a given threshold, we consider that there is a posture change.

Figure 4.3: Example of detection of a posture change. From left to right: first frame of the period, last frame and their thresholded absolute difference.

4.2.4 Detection of self-touching

Harrigan suggests in [53] that self-touching can be an indicator of negative affect. We detect two types of self-touching: face touching, which is part of the displacement behaviours described by Troisi [128], and rubbing fingers together. Since detecting self-touching requires a precise tracking of the hand, we use skin detection to refine the hand joint location provided by the Kinect. Starting from the position given by the Kinect, we look for the closest skin pixel detected. This becomes the new position of the hand. Figure 4.4 shows an example of the refinement of the hand location.

To determine if a person is touching her face, we compute the hand-head and the hand-neck distances. If one of these distances is below a given threshold, we consider that the person is face touching (Figure 4.5). The number of occurences (FTC) and the average duration (FTMD) are used as features. Similar features are extracted when the person is self-touching her head with two hands (FT2HC and FT2HMD).

Figure 4.4: Example of the refinement of the hand joint location. The red squares represent the location given by the Kinect. The blue circles represent the new location after refinement. Best viewed in color.

To detect gestures such as rubbing fingers together, the Kinect skeleton is not sufficient since it does not provide joints for the fingers. Thus, using the hand positions, we first extract the sub-image of each hand region. Then, we compute the IQoM between successive extracted sub-frames. We only compute it when the person is not moving her hand since the IQoM can be affected by changes in the background. This feature is computed for each hand separately (LHM for the left hand, RHM for the right one) and for both (HM).

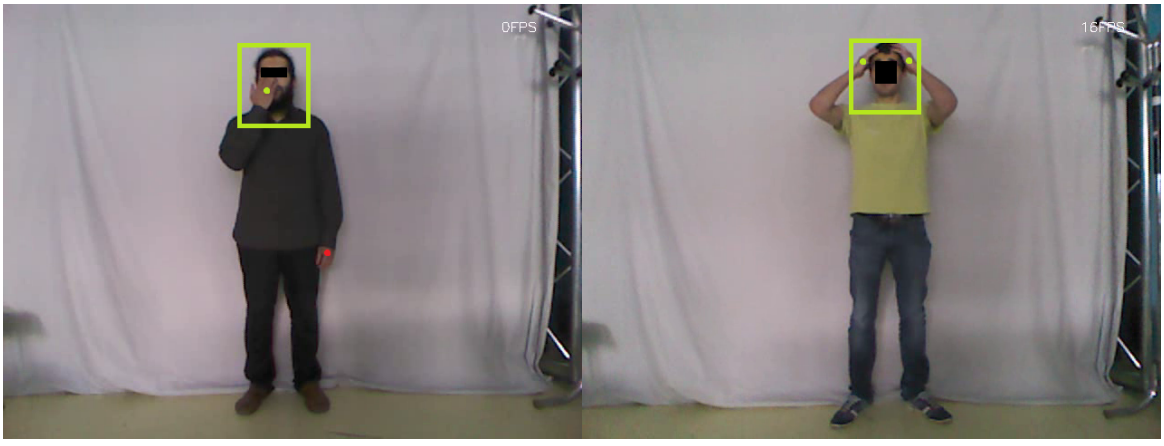

Figure 4.5: Examples of detections of face touching. Left image: face touching with one hand. Right image: face touching with two hands.

4.3 Facial features

For facial expressions, we extract the activation levels of 12 Actions Units (AU). Paul Ekman presented AUs as part of the Facial Action Coding System (FACS) [37]. This system encodes movements of individual facial muscles in order to characterize facial expressions in a systematic way. Each AU has 5 possible activation levels: trace, slight, marked, severe and maximum.

Figure 4.6: Example of Action unit activation.

To extract these activation levels, we use the method presented in [94], which proposes a multi-task extension for a subspace learning algorithm called Metric Learning for Kernel Regression. Once we have extracted the activation level of each AU for each HD video frame, we compute the average and the standard deviation and use them as features.

4.4 Physiological features

The Nexus-10 device is used to extract physiological features classically associated in the literature with stress. The BVP sensor allows the monitoring of cardiac signals such as the Blood Volume Pulse (BVP) or the Heart Rate (HR). The Heart Rate Variability (HRV), which corresponds to the variation in the time between heartbeats (also called R-R or N-N intervals), is also acquired. From these signals, the Blood Volume Pulse Amplitude (BVPA), the Heart

Rate Variability Amplitude (HRVA), the spectral density of the low frequency band of the HRV (HRV-LF%), the square root of the mean squared difference between successive N-N intervals (HRV-RMSSD) and the standard deviation of N-N intervals (HRV-SDNN) are computed and used as features. The respiration sensor monitors the abdominal breathing signal (RSP). From this signal, we compute the respiration rate (RSPR) and the respiration amplitude (RSPA). Also, the level of coherence between the respiration and the heart rate (RSP+HR) is computed from both signals. The temperature sensor monitors skin temperature (TMP). The skin conductance sensor monitors the galvanic skin response (GSR). Finally, the EXG sensor allows the monitoring of the EMG signal from the sternocleidomastoid and upper trapezius muscles (EMG and EMG2). We compute the mean frequency (EMGMF) and the amplitude (EMGA) of the mean signal between EMG and EMG2. For most of these signals, we use the mean value, the standard deviation, the min and the max values as features.

Figure 4.7: Images of the sensors used to capture physiological signals. Top row, from left to right: BVP sensor, respiration sensor and temperature sensor. Bottom row, from left to right: skin conductance sensor and EXG sensor[2].

4.5 Conclusion

In this thesis, we have extracted original body features from Kinect data for stress detection such as periods of high activity, posture changes, face-touching and fingers rubbing. We also extracted the level of activation of 12 Action Units as facial features, and physiological features

[2]Image retrieved from http://www.mindmedia.info/CMS2014/en/products/sensors

from BVP, GSR, EMG, respiration and skin temperature signals. Overall, we have extracted 39 behavioural features - 15 body features and 24 facial features - and 62 physiological features. We use these features for the experiments on automatic stress detection we present in the following chapters.

Feature	Description
AU1	Inner Brow Raiser
AU2	Outer Brow Raiser
AU4	Brow Lowerer
AU5	Upper Lid Raiser
AU6	Cheek Raiser
AU9	Nose Wrinkler
AU12	Lip Corner Puller
AU15	Lip Corner Depressor
AU17	Chin Raiser
AU20	Lip Stretcher
AU25	Lips Part
AU26	Jaw Drop
SQoM	QoM computed with the skeleton
IQoM	QoM computed with the RGB frames
HAPC	Number of periods of high activity
HAPMD	Mean duration of periods of high activity
HAPMV	Mean highest value of periods of high activity
PCC	Number of posture changes
FTC	Number of times face touching with one hand occured
FTMD	Mean duration of face touching with one hand
FT2HC	Number of times face touching with two hands occured
FT2HMD	Mean duration of face touching with two hands
LHM	QoM for the left hand
RHM	QoM for the right hand
HM	QoM for both hands
HeM	QoM for the head
HeMZ	QoM for the head only along Z-axis

Table 4.1: List of the extracted behavioural features.

Feature	**Description**
BVP	Blood Volume Pulse
BVPA	Blood Volume Pulse Amplitude
EMG	Electromyographic activity of the sternocleidomastoid and upper trapezius - channel 1
EMG2	Electromyographic activity of the sternocleidomastoid and upper trapezius - channel 2
EMGMF	Electromyographic activity of the sternocleidomastoid and upper trapezius Mean Frequency
EMGA	Electromyographic activity of the sternocleidomastoid and upper trapezius Amplitude
GSR	Galvanic Skin Response
HR	Heart Rate
HRVA	Heart Rate Variability Amplitude
HRV-LF%	Heart Rate Variability Low Frequency band
HRV-RMSSD	Heart Rate Variability square root of the mean squared difference between adjacent N-N intervals
HRV-SDNN	Heart Rate Variability Standard Deviation of Normal to Normal intervals
RSP	Abdominal Respiration
RSPA	Abdominal Respiration Amplitude
RSPR	Abdominal Respiration Rate
RSP+HR	Level of coherence between the Respiration and the
TMP	Temperature

Table 4.2: List of the extracted physiological signals.

Chapter 5

Handling interindividual differences for automatic stress detection

5.1 Introduction

As explained in Section 1.2, one of the biggest challenges in affective computing and in social signal processing is to develop solutions to handle interindividual differences. Indeed, as said in [9], *"Different people tend to display the same emotion in very different ways"*. This heterogeneity is found in several modalities such as speech [18, 137], physiology [7, 17, 81], facial expressions [94, 137] and body language [9].

In pattern recognition, one common method to handle interindividual differences is data normalization. The objective of normalization techniques is to reduce the impact of interindividual variability while preserving the differences between classes. It can also be used in machine learning to modify data distribution or scale [8]. These methods can either be person-specific or generic. Person-specific normalizations apply a different transformation to each subject's data, while generic normalizations apply the same transformation.

In this chapter, we evaluate 5 different normalization techniques for stress detection: mean-centering, range normalization, standardization, baseline comparison and Box-Cox transformation. To do so, we use the *Dataset-44* introduced in Chapter 3, thus working only on subjects' behaviour. We first introduce the normalization methods, then we present the results of the evaluation of each normalization method on 2 assessments: self-assessment and external observer assessments. We will see that the effectiveness of each normalization method also depends

on the chosen assessment strategy.

5.2 Normalization methods

5.2.1 Mean-centering (MC)

Using the idea presented in [9], we make the hypothesis that a person's behaviour biases are affected by constant factors, such as gender or physical build. Therefore, we model their impact as an additive constant, which can be computed as the average behaviour of the person over all the steps:

$$x'_{ps} = x_{ps} - \bar{x_p}$$

with x_{ps} the value of a given feature for person p during step s, $\bar{x_p}$ the average value of the given feature over all the steps for person p and x'_{ps} the feature normalized value. This person-specific method is used in the work presented in [100]. The authors stated that using this method improved the classification accuracies of 2 machine learning algorithms: J48 decision tree and SVM.

5.2.2 Range normalization (RN)

In the same fashion as mean-centering, one can make the hypothesis that a person's behaviour biases influence how much stress impact her behaviour. As we previously modeled constant factors by an additive constant, we can model by a multiplicative constant the rate at which stress influences the behaviour:

$$x'_{ps} = \frac{x_{ps}}{max(x_p) - min(x_p)}$$

with $max(x_p)$ and $min(x_p)$ the maximum and minimum values over all the steps for person p.

5.2.3 Standardization (ST)

This method combines the hypotheses of the two previous methods. We model constant factors by an additive constant and how much stress impact one's behaviour a multiplicative constant:

$$x'_{ps} = \frac{x_{ps} - \bar{x_p}}{\sigma_{x_p}}$$

with σ_{x_p} the standard deviation for person p. This person-specific normalization method is used in the following papers [3, 133].

5.2.4 Baseline comparison (BL)

Another way to look at personal biases is to consider that everybody has a usual behaviour which can be different from each other. For instance, some people may in general be more active with their body while talking than others. Therefore, one can make the hypothesis that stress does not impact how active someone is, but rather how active someone is compared to her usual self. Thus, one can model this comparison by computing the relative difference between the current behaviour and a baseline behaviour. This baseline behaviour needs to be as close as possible to the subject's usual behaviour so that unusual, potentially discriminative behaviour would be extracted.

$$x'_{ps} = \frac{x_{ps} - baseline_p}{baseline_p}$$

with $baseline_p$ the baseline vector of features for person p. This method is used in the following works [25, 45]. In our case, we use the first step of the experiment as the baseline since it is the easiest step. It is noteworthy that although it is the best way we have to implement this normalization method, it also has its flaws. Indeed, making the hypothesis that because the first step is the easiest one it is also the least stressful is risky: subjects may be stressed at the beginning of the experiment because of apprehension.

5.2.5 Box-Cox transformation (BC)

Interindividual differences also impact data distribution. Indeed, there may be some features for which a subject is so different from others that her data acts as an outlier. Therefore, it may prevent the machine learning algorithm from finding the most adequate model. In order to overcome this issue, one can use the Box-Cox transformation [107] to normalize feature distribution in a systematic way. The Box-Cox transformation is defined as:

$$x'_\lambda = \frac{x^\lambda - 1}{\lambda}$$

The aim is to find the value of λ that maximizes the correlation between the transformed feature x'_λ and the normal distribution $\mathcal{N}(\mu(x'), \sigma(x')^2)$. We only compute the correlation for specific

values of λ that can be found in Table 5.1. Unlike the 4 previous normalization methods that we have presented, the Box-Cox is not person-specific: the same transformation is applied to the data of all subjects.

λ	-2	-1	-0.5	0	0.5	1	2
x'	$\frac{1}{x^2}$	$\frac{1}{x}$	$\frac{1}{\sqrt{x}}$	$log(x)$	$\sqrt{x}$	x	x^2

Table 5.1: Tested values of λ for the Box-Cox transformation and their associated transformation function.

5.3 Evaluation

5.3.1 Evaluation process

We use a classification task to evaluate the effectiveness of each normalization method. The objective is to predict the binary stress label - Stress or Non-Stress - of each of the 264 collected examples composing *Dataset-44* for assessments sets EOA and SA. However, we remove the first step of each subject when we evaluate the baseline comparison method since we use these steps as our baseline, lowering the number of examples to 220.

We run 2 experiments in order to evaluate normalization methods. In the first experiment, we use all the features present in *Dataset-44* to compare the classification results obtained with and without normalization. However, the hypothesis we make for each normalization method may be true for some features, but not for others. Therefore, we use a feature selection method in the second experiment. We use a backward elimination wrapper [66] as our feature selection method: starting from the complete set of features, we iteratively remove the worst feature of the remaining set. Once all features have been removed, we keep the subset that gives the best classification performances. We perform this feature selection step on all data, since the objective is to evaluate normalization methods with their most relevant features.

We use SVMs with three different kernel functions - linear, polynomial and radial basis - to compute classification results. We use a 10 fold subject independent cross validation strategy to compute the results: steps from 4 or 5 people are used as the testing set. The steps of the remaining people are used as the training set. This cross validation is also used with the

training set to determine the SVM and kernel function parameters. It is important to note that the parameters for the Box-Cox transformation are computed using only the training set and are then applied to the testing set. Regarding person-specific normalizations, the parameters are already independent between training and testing sets since our cross-validation is subject-independent.

Since our dataset is unbalanced for the 2 assessment sets considered we have chosen the average of the F1 score for both Stress and Non-Stress classes as the performance metric. This metric allows us to consider the recall and the precision of both classes, unlike the usual F1 score that considers the recall and the precision only for the positive class and ignores the true negative rate. We use the Student's t-test to compare the average F1 scores obtained with and without normalization.

5.3.2 Results

In this section, we first present the results obtained with SA and then those obtained with EOA. We discuss the overall results in the next section.

Prediction of SA

The results of the first experiment for the prediction of SA are presented in Figure 5.1. Overall, the best normalization method for this experiment is range normalization (RN): it significantly improves the results compared to raw features for all 3 kernels. Regarding the mean score over the 3 kernels, it obtains the best score with 0.57 and is significantly better ($p < 0.05$) than all the other methods. Then, 2 methods obtain similar results: mean-centering (MC) and baseline comparison (BL). Indeed, MC is significantly better for the polynomial and linear kernels. However, it is not considered significantly better if we look at the mean score over the 3 kernels. On the other hand, BL is only significantly better for the linear kernel but is significantly better on average. Overall, person-specific normalizations seem to be the most efficient ones for this experiment since normalization method which obtains the worst average results is the only one not person-specific: the Box-Cox transformation with a mean score of 0.477.

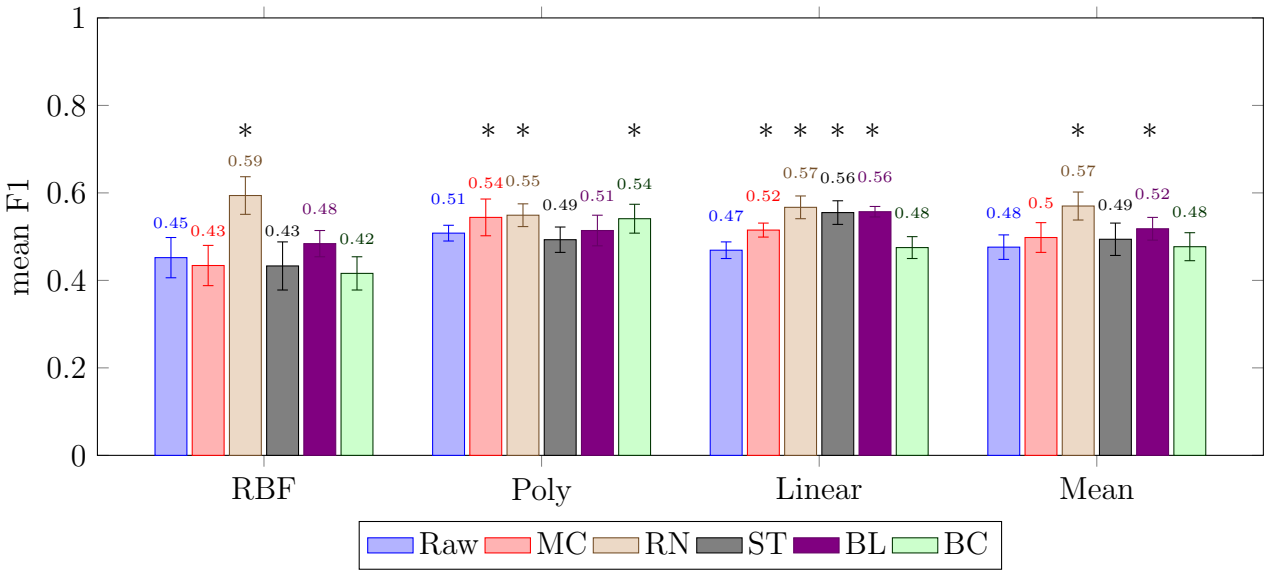

Figure 5.1: Results of the first experiment for the prediction of SA. We present the mean F1 score of each normalization method for each kernel and also their average over the 3 kernels. Raw represents the results when no normalization is applied. For each kernel and for the average, normalization methods which perform significantly better than raw features ($p < 0.05$) are marked by $*$.

The results of the second experiment, for which a feature selection step is added, are presented in Figure 5.2. Once again, the best normalization technique for this experiment is range normalization (RN): as for the first experiment, it improves significantly the results obtained with all 3 kernels. On average, it is also significantly better than raw features and also than other normalization methods. Surprisingly, at least accordingly to the results of the first experiment, the second best normalization technique for this experiment is the Box-Cox transformation (BC). It also improves significantly the results obtained with all 3 kernels and it obtains the second best average F1 score with 0.581. The fact that this normalization is the worst one when there is no feature selection step, and is the second best one when there is tends to show that the Box-Cox transformation may not be relevant for all features.

Overall, it appears clearly that the best normalization method for the prediction of SA from behavioural features is range normalization (RN). It indeed ranked first in both experiments and obtained results significantly better than the other normalization methods we have tested. It is also noteworthy that no normalization methods lowered the classification results on average. Regarding the first experiment, the worst normalization method was the Box-Cox transformation, but it did not significantly impact the results. In the second experiment, the worst

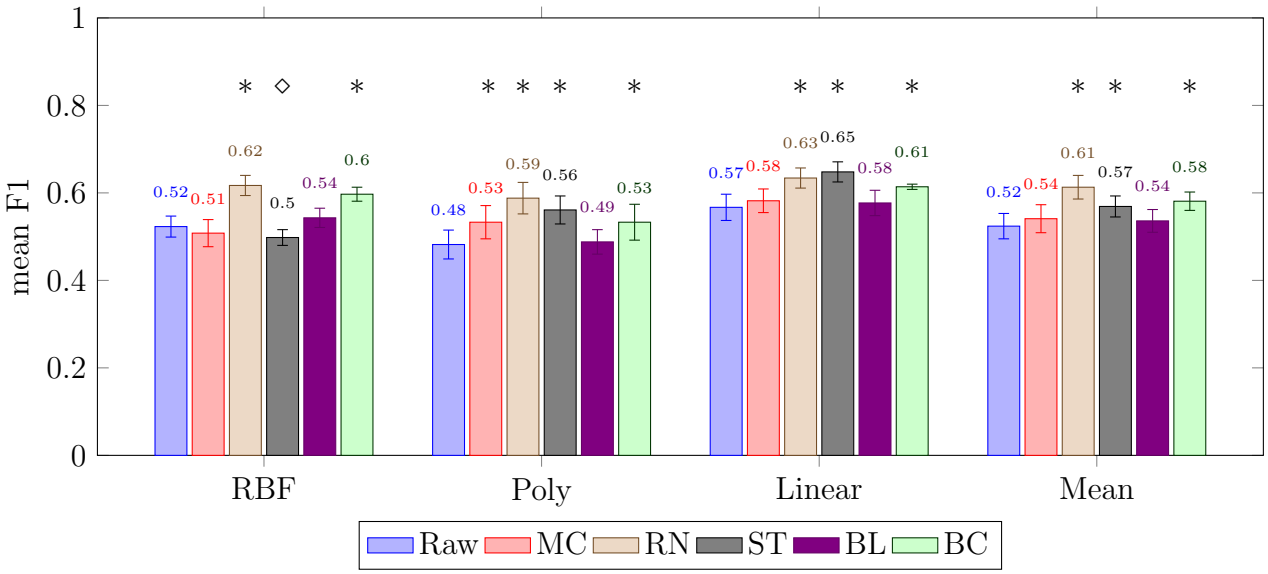

Figure 5.2: Results of the second experiment for the prediction of SA. We present the mean F1 score of each normalization method for each kernel and also their average over the 3 kernels. Raw represents the results when no normalization is applied. For each kernel and for the average, normalization methods which perform significantly better than raw features ($p < 0.05$) are marked by $*$ and those which perform significantly worse are marked by $\diamond$.

normalization method was the baseline comparison (BL), but it still slightly improved the classification results: 0.524 ± 0.029 for raw features and 0.536 ± 0.026 for BL.

Prediction of EOA

The results of the first experiment for the prediction of EOA are presented in Figure 5.3. It appears clearly that the best normalization method is the Box-Cox transformation. It improves extremely significantly ($p < 0.0001$) the results for all 3 kernels and on average. Regarding person-specific normalizations, all of them provide significantly lower performances compared to raw features. Among them, the best one is again range normalization (RN).

As shown in Figure 5.4, the findings of the second experiment are similar to those of the first one: the Box-Cox transformation (BC) is by far the best normalization method. Overall, person-specific normalizations significantly degrade classification results. Only the range normalization (RN) does not obtain significantly lower classification results on average.

Overall, it appears that person-specific normalization methods are not relevant for the prediction of EOA from behavioural features. Almost all of the reviewed person-specific methods

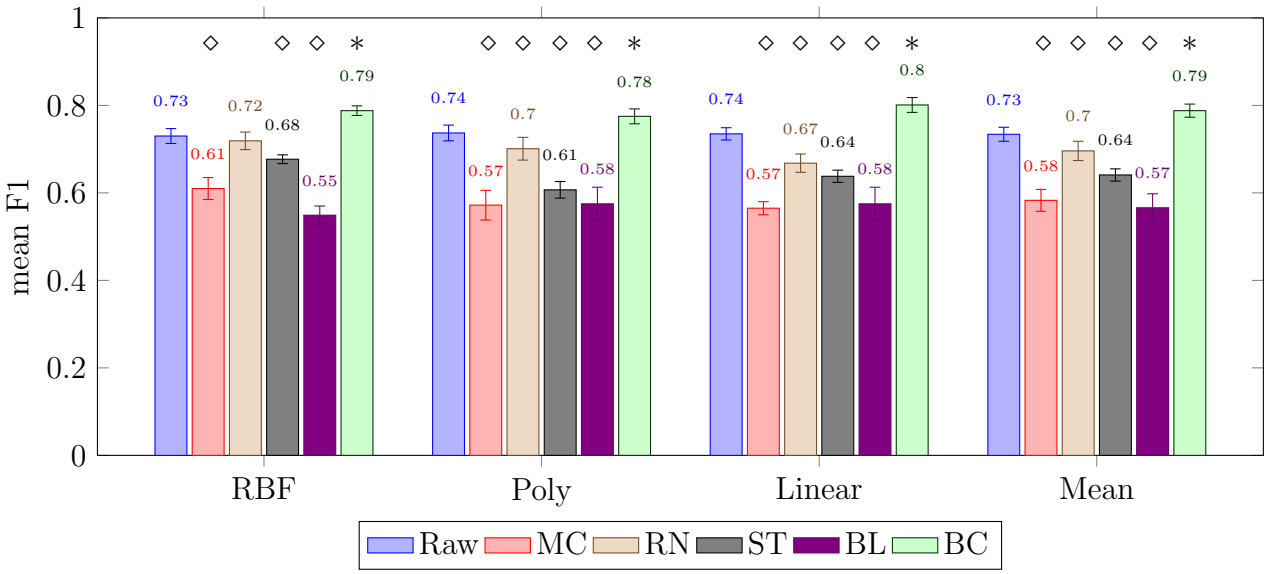

Figure 5.3: Results of the first experiment for the prediction of EOA. We present the mean F1 score of each normalization method for each kernel and also their average over the 3 kernels. Raw represents the results when no normalization is applied. For each kernel and for the average, normalization methods which perform significantly better than raw features ($p < 0.05$) are marked by $*$ and those which perform significantly worse are marked by $\diamond$.

obtained significantly worse classification results compared to raw features in both experiments. On the other hand, the Box-Cox transformation appears especially valuable, as it always provided very significantly better results.

5.4 Discussion

We can see that the effectiveness of normalization methods depends on the assessment considered. On one hand, regarding self-assessments (SA), it appears that most of the evaluated methods improved the results obtained without normalization (Figures 5.1 and 5.2). The method which provided the best classification results is the person-specific range normalization. On the other hand, regarding external observer assessments (EOA), all the person-specific normalization methods obtained significantly lower results than raw features (Figures 5.3 and 5.4). However, the Box-Cox transformation, which is not person-specific, significantly improved classification results. Overall, it seems that person-specific normalizations provide good performances for SA, but not for EOA, while the Box-Cox transformation performs decently for SA, but give excellent results for EOA.

This association between assessment strategy and normalization method may be explained

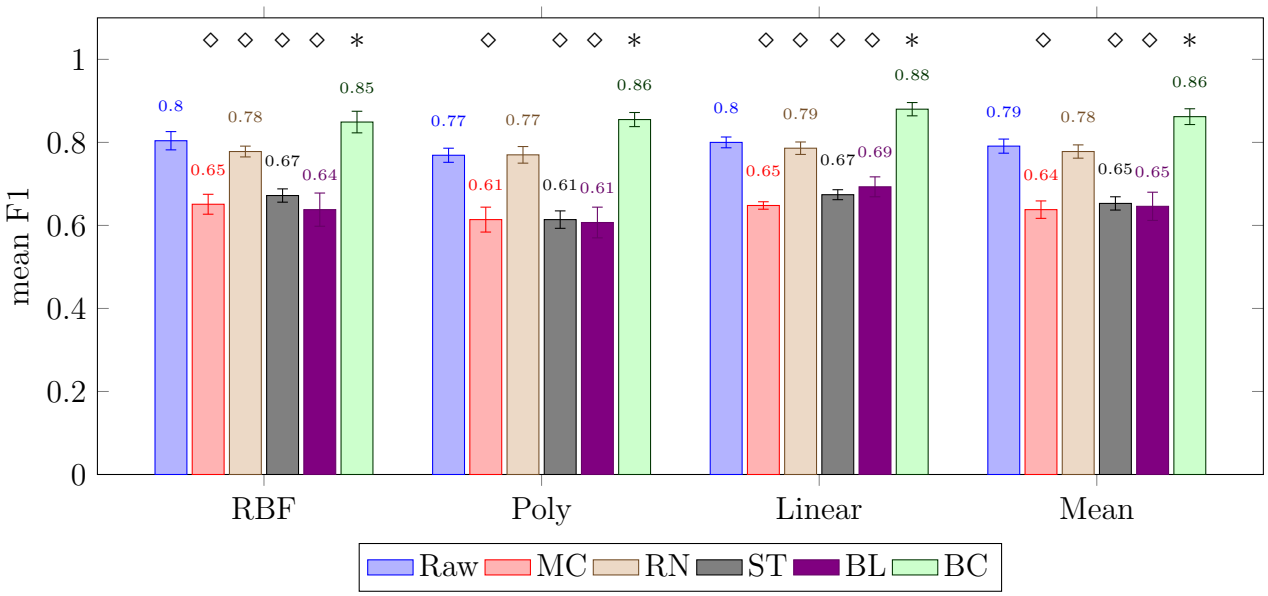

Figure 5.4: Results of the first experiment for the prediction of EOA. We present the mean F1 score of each normalization method for each kernel and also their average over the 3 kernels. Raw represents the results when no normalization is applied. For each kernel and for the average, normalization methods which perform significantly better than raw features ($p < 0.05$) are marked by $*$ and those which perform significantly worse are marked by $\diamond$.

by the annotation processes of SA and EOA. As explained in Section 3.3.2, the experiment participants provide their self-assessment during the debriefing. To do so, for each of the 6 steps composing the experiment, they first watch the video recorded during the given step before providing their self-reported stress level on a 1-5 Likert-scale. Because people have an idea about how they usually behave, they can mentally compare what they see on the video with their normal behaviour. In addition, they also can compare videos between them to see how their behaviour evolved. Therefore, they provide ratings relative to themselves: they behave more or less stressed than usually or than the previous videos. Thus, one can try to extract these relative differences by using person-specific normalizations.

However, external observers may provide very different ratings, as observed in Section 3.3.4. On the crowdsourcing platform we used, 5 random videos are chosen to create a page for which annotators have to provide their ratings. Thus, annotators can compare videos of different people to judge whether someone looks stressed or not. There, they provide ratings relative to other people: a given person behave more or less stressed than others. Thus, by using person-specific normalization methods, we may remove the interindividual differences on which

annotators based their judgements. These differences between the annotation processes of SA and EOA may explain why person-specific normalization methods work well for the prediction of SA, but are irrelevant for the prediction of EOA.

Regardless of the normalization method used, it appears clearly that behavioural features provide much better classification performances for the prediction of EOA than for the prediction of SA. This association between features and assessment strategies is studied in more depth in Chapter 6.

5.5 Conclusion

In this chapter, we have studied how data normalization may help to reduce the impact of interindividual differences. We evaluated 5 normalization methods: mean-centering, range normalization, standardization, baseline comparison and the Box-Cox transformation. To do so, we used the two assessment strategies present in *Dataset-44*: self-assessment and external observer assessment. For each of these assessment strategy, we conducted two experiments. We used all the features included in *Dataset-44* for the first experiment, while we added a feature selection step for the second experiment. Overall, the main findings of these experiments are:

- Person-specific normalization methods are relevant only for the prediction of SA.
- For the prediction of SA, the best normalization method is range normalization.
- For the prediction of EOA, the best normalization method is the Box-Cox transformation.

It appears that selecting the adequate normalization method for affective computing applications is a complex task as it is highly dependent on the features extracted, the way the affective phenomenon is assessed and also on the machine learning algorithms used. Therefore, one has to take a global and comprehensive approach to select the most relevant method.

Chapter 6

Multi-perspective evaluation of the impact of stress

6.1 Introduction

Recent progress in computer vision, affective computing and social signal processing have helped to understand the impact of affective and mental states on human behaviour and body. For instance, frameworks for automated analysis and detection of frustration [62] and depression [29, 59] provide valuable information about the predictive performance of certain features in these specific contexts. In [62], the authors suggest that fidgets and the head velocity are relevant features for frustration detection. However, these results greatly depend on the way frustration is assessed. The authors chose to use self-assessment as their annotations, but the results may have been different if they had used biomarkers or external perception instead.

In this chapter, we propose to study automatic stress prediction in a more comprehensive way by considering the results obtained with the 3 assessments described in Chapter 3. Using these 3 annotations, we evaluate the predictive power of behavioural and physiological cues. We use the *Dataset-21* introduced in Chapter 3 since we need the 3 assessments. Thus, the data is composed of the 101 behavioural and physiological features presented in Chapter 4 (Tables 4.1 and 4.2). We first present how we process the data. We describe feature transformation and feature selection methods. Then, we present the results obtained for each annotation. We describe the prediction power of each modality - behaviour and physiology - and of each feature. Finally, we interpret and discuss the results obtained. Overall, we argue that stress detection

should be tackled with a multiple assessment approach because of the complexity of stress. It allows to better understand associations between features, modalities and assessments, leading to more robust stress detection systems.

6.2 Data preprocessing

6.2.1 Feature transformation

As seen in Chapter 5, the effectiveness of normalization methods is highly dependent on the assessment strategy considered. However, it appeared that the Box-Cox transformation provided good results for both SA and EOA, while the other methods we tested provided good results only for SA. Therefore, we chose to use the Box-Cox transformation to normalize our data.

6.2.2 Feature subset selection

We perform feature subset selection in order to avoid overfitting and better understand the predictive power of each feature. Each result presented in Section 6.3 corresponds to the best one obtained among the 3 following feature subset selection methods.

Forward selection wrapper (FSW)

Wrappers evaluate a subset of features by using the same machine learning algorithm as in the final application [66]. In our case, we use a SVM with a linear kernel function. Since training SVMs is computationally expensive, exploring the space of feature subsets is usually done using greedy methods [49]. With forward selection, starting from using only the feature with the best accuracy, we iteratively add the best feature among the remaining ones. Once all features have been added, we keep the subset that gives the best classification performances.

Backward elimination wrapper (BEW)

This method also uses a SVM to evaluate subsets. Backward elimination is also a greedy search strategy: starting from the complete set of features, we iteratively remove the worst feature of the remaining set. Once all features have been removed, we keep the subset that gives the best classification performances.

Simulated annealing with Hall correlation (SAHC)

For this method, we use the simulated annealing metaheuristic [64] to explore the space of feature subsets. Because of the computational cost of this space search strategy, we use the Hall correlation [50] to evaluate feature subsets. We then get a good approximation of the subset that both maximizes the correlation between features and labels, and minimizes the inter-feature correlation.

6.3 Evaluation

6.3.1 Evaluation process

We use a classification task to evaluate the predictive value of the features introduced in Chapter 4. The objective is to predict the binary stress label - Stress or Non-Stress - of each of the 126 examples composing *Dataset-21*. To do so, we use SVMs with three different kernel functions: linear, polynomial and radial basis. We use a 10 fold subject-independent cross validation strategy to compute the results: steps from 2 or 3 people are used as the testing set. The steps of the remaining people are used as the training set. This cross validation is also used with the training set to determine the SVM and kernel function parameters. Since our dataset is unbalanced for 2 assessment sets - SA and EOA - we have chosen the average of the F1 score for both Stress and Non-Stress classes as the performance metric.

It is important to note that we use all the data for the feature transformation and feature subset selection steps. It has been done in order to facilitate the interpretation of the results and of the relevance of each feature. Consequently, we also present the average F1 score when the parameters for the feature transformation and the feature subset selection are selected using only data from the training set and are then applied on the testing set.

6.3.2 Evaluation of the predictive power of each modality

Figures 6.1, 6.2 and 6.3 show the classification results obtained by the best selected feature subset for all 3 assessment sets. Features are selected from the whole set of features (All*), only behavourial ones (Behaviour*) or only physiological ones (Physio*). Regarding PEA,

we compute the results in 2 different conditions. First, we compute the classification results after having dismissed all the features which are theoretically too correlated to the heart-rate variability: HRV-LF%, HRV-SDNN, HRV-RMSSD and RSP+HR. Including all the functionals applied to each feature (i.e. mean, standard deviation, min and max), we dismissed a total of 10 features that we refer to as Physiology Label Related (PLR) features. In the second condition, we dismiss only the features directly related to the low frequency in the heart-rate variability (HRV-LF%) that we used as our assessment. Including all the functionals, we dismissed 4 features.

In general, we can see that, for most of the subsets, the linear kernel outperforms both RBF and polynomial kernels. It is due to the fact that most of the best subsets were provided by one of the 2 wrappers, which are optimized for the linear kernel. We take this into account in the following discussion. Overall, the results show that depending on the assessment set considered, the effectiveness of each modality and of their combination varies.

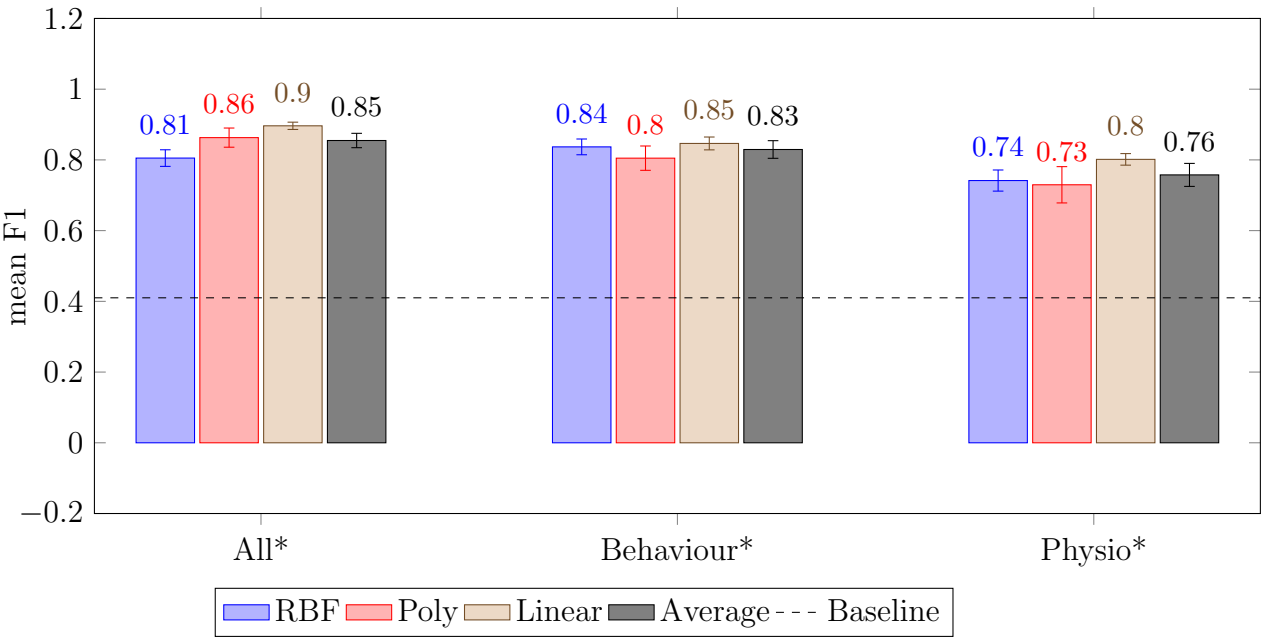

Figure 6.1: Performances of each kernel for each modality for the prediction of EOA. The baseline average F1 score obtained by a random classifier is 0.410 (± 0.083).
Features selected in All*: AU1-std, AU2-mean, AU2-std, AU4-mean, AU6-mean, AU12-std, AU15-mean, AU17-mean, BVP-mean, BVPA-max, HeM, IQoM, FTC, PCC, RSP-var, RSPR-max, RSP+HRC-max, RSP+HRC-mean,RSP+HRC-min, EMG-min, GSR-var

Regarding EOA (Figure 6.1), we can see that both modalities achieve good mean F1 scores: 0.829 (± 0.025) for behavioural features and 0.758 (± 0.033) for physiological ones. It is not

surprising that behavioural features significantly outperform physiological ones ($p < 0.0001$) since annotators based their judgement solely on the behaviour of the person in each video. It is however interesting to see that physiological features can predict how stress is assessed by external observers. It could have been explained by the fact that some of the physiological features selected can be visually perceived: features related to the respiration rate and EMG features, which can reflect the upper body activity. But the results obtained after having dismissed these features are similar with a mean F1 score of 0.751 ($\pm$ 0.028). This feature subset is composed of 15 features: 7 related to HRV, 4 related to HR, 2 related to BVP, one related to GSR and one to skin temperature. The best results are obtained when we combine both modalities with a mean F1 score of 0.855 ($\pm$ 0.020). The subset All* is composed of 24 features: 15 behavioural features and 9 physiological ones. It is however interesting to note that the difference in mean F1 score between the subsets All* and Behaviour* is statistically significant if we consider the 3 kernel functions, but is not if we do not consider the linear kernel. Overall, it seems that using only behavioural features is sufficient for the prediction of EOA. When we use all features and we include the feature transformation and the feature subset selection in the training phase, we obtain a mean F1 score of 0.739 ($\pm$ 0.023).

Regarding SA, Figure 6.2 shows that the combination of physiological and behavioural features outperforms the results obtained when using only one modality. It is understandable since the subjects of the experiment described in Chapter 3 watch their own videos before annotating them. Thus, their answers are the result of both their personal experiences and their behaviour analysis. The subset All* obtains a mean F1 score of 0.795 ($\pm$ 0.028) and is composed of 32 features: 21 physiological features and 11 behavioural ones. When we use all features and we include the feature transformation and the feature subset selection in the training phase, we obtain a mean F1 score of 0.691 ($\pm$ 0.034).

Figure 6.3 displays the results obtained for the prediction of PEA. As expected, physiological features obtain a good classification performance with an average F1 score of 0.777 ($\pm$ 0.021) for the first condition (i.e. no features related to the heart-rate variability) and of 0.810 ($\pm$ 0.026) for the second condition (i.e. no features related only to the low frequency in the heart-rate variability). As expected, using the features related to the heart-rate variability significantly

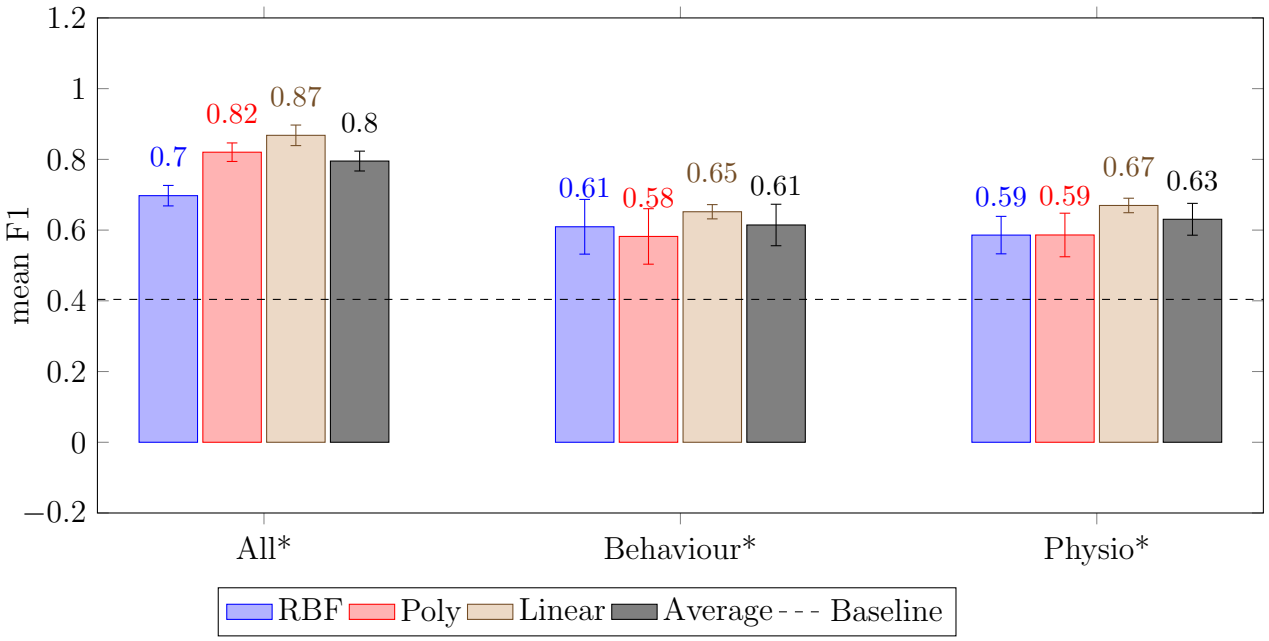

Figure 6.2: Performances of each kernel for each modality for the prediction of SA. The baseline average F1 score obtained by a random classifier is 0.404 (± 0.079).
Features selected in All*: AU4-mean, AU6-mean, AU6-std, AU12-std, AU17-std, BVP-max, BVP-min, BVPA-max, BVPA-min, BVPA-var, EMGMF-max, EMGMF-var, EMG-min, EMG-mean, EMG-var, GSR-var, HAPMV, HR-max, HRVA-var, IQoM, RHM, RSPA-max, RSPA-min, RSPA-var, RSPR-max, RSPR-mean, RSPR-min, RSP+HRC-max, RSP-var, FTMD, FT2HMD, TMP-min

improves the results ($p = 0.0059$) since these features are related to the one we used to compute PEA labels. Overall, both conditions significantly outperform behavioural features ($p < 0.05$). It is however surprising to see that using only behavioural features also provides a good average F1 score of 0.740 (± 0.020). The selected subset Behaviour* is composed of 10 features: 7 features related to Action Units, 2 features related to body movement and the mean duration of face touching (FTMD). Regarding the combination of behavioural and physiological features, there is no significant difference between the results obtained by both conditions: 0.831 (± 0.020) and 0.833 (± 0.021) for the first and second condition respectively. However, it is noteworthy that the best subset selected for the second condition is smaller - 17 features (12 physiological and 5 behavioural ones) against 25 features (16 physiological and 9 behavioural ones) for the first condition - and contains 3 features related to HRV: HRV-RMSSD, HRV-SDNN and RSP+HR-Mean. When we use the features of the first condition and we include the feature transformation and the feature subset selection in the training phase, we obtain a mean F1 score of 0.705 (± 0.020).

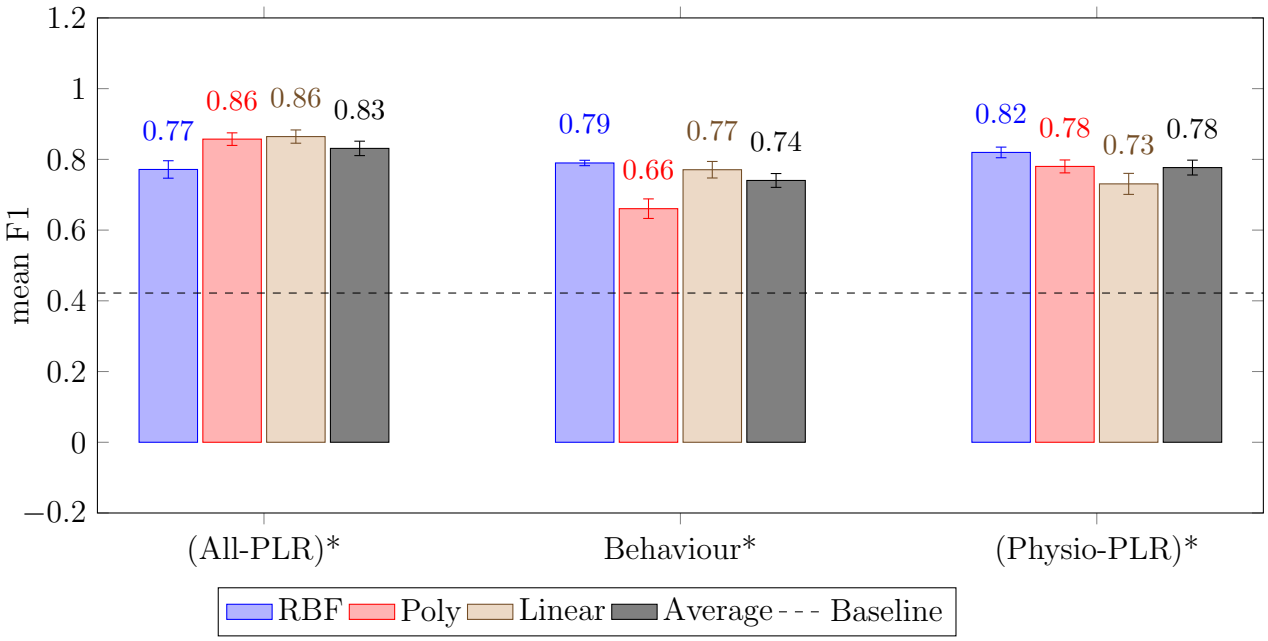

Figure 6.3: Performances of each kernel for each modality for the prediction of PEA without using features related to HRV. The baseline average F1 score obtained by a random classifier is 0.422 (± 0.080).
Features selected in All*: AU1-mean, AU2-std, AU15-mean, AU17-std, AU25-mean, AU25-std, AU26-mean, BVPA-mean, BVPA-min, BVPA-var, EMGA-mean, EMGMF-max, EMGMF-mean, EMG-min, GRS-max, HR-max, HR-mean, HRVA-max, HRVA-var, RSPA-max, RSPA-min, RSPA-var, RSPR-var, FTC, FTMD

6.3.3 Evaluation of the predictive power of each feature

We use the evaluation process described in Section 6.3.1 using only one feature at a time in order to better understand the classification performance of each feature for each assessment set. We compute the F1 score obtained by the three kernel functions. The average F1 score is used to rank features. In order not to overload the charts, we present only the five best features of each assessment set. The average F1 scores for each feature are presented in Tables 6.4 and 6.5.

Regarding EOA, the results obtained by the 5 best features are shown in Table 6.1. We can see that these features achieve good classification performances even when used alone. Among these 5 features, there are 4 behavioural features and one physiological one, which is not surprising since the external observers had only access to the participants' behaviour. The 4 behavioural features are all related to movement: 2 features are related to head movement (HeM and HeMZ), one to hand movement (HM) and one to body movement (IQoM). The best

Feature	Average F1	Stdev
HeM	0.780	0.016
IQoM	0.723	0.025
HeMZ	0.716	0.021
BVP-Min	0.705	0.029
HM	0.696	0.024

Table 6.1: Five best features according to their average F1 score for the prediction of EOA. The Stdev column represents the standard deviation of the average F1 score over 10 runs.

physiological feature is the minimum of the Blood Volume Pulse (BVP - Min). It is noteworthy that the 5 best physiological features for the prediction of EOA are all related to the BVP: 3 are related to the raw BVP signal (BVP-Min, Mean and Var) and 2 are related to the amplitude of the BVP signal (BVPA-Var and Max).

Feature	Average F1	Stdev
IQoM	0.621	0.028
HAPMV	0.617	0.028
SQoM	0.616	0.025
HeM	0.614	0.038
RSP-Min	0.609	0.031

Table 6.2: Five best features according to their average F1 score for the prediction of SA. The Stdev column represents the standard deviation of the average F1 score over 10 runs.

Table 6.2 displays the 5 best features for the prediction of SA. We can notice that these features achieve lower F1 scores than the best features for EOA and PEA. Added to the fact that, as shown in Figure 6.2, only a combination of physiological and behavioural features achieved good classification performances, it tends to show that the prediction of SA is more complex, is based on both behavioural and physiological cues, and requires more information than the prediction of EOA and PEA. Among the 5 best features, 3 are behavioural and 2 are physiological. 2 features are related to body movement (IQoM and HeM), 2 to blood volume pulse (BVP - Min and Var) and one to periods of high activity (HAPC).

The 5 best features for the prediction of PEA are all related to the heart: 3 are related to

Feature	Average F1	Stdev
RSP-Mean	0.621	0.028
RSPR-Max	0.617	0.033
AU4-Mean	0.600	0.031
RSPR-Min	0.590	0.043
AU2-Std	0.590	0.030

Table 6.3: Five best non heart-related features according to their mean F1 score for the prediction of PEA. The Stdev column represents the standard deviation of the average F1 score over 10 runs.

the heart rate (HR- Mean, Max and Var) and 2 are related to the amplitude of the heart rate variablity (HRVA - Max, Mean). Their average F1 scores range from 0.711 $\pm$ 0.047 for HR-Mean to 0.652 $\pm$ 0.024. These results are coherent with what we described in Section 2.2 for the biological perspective: activating the HPA pathway and the ANS leads to an increased heart rate, which impacts the heart rate variability. Thus, it is not surprising to see these features perform well.

However, it is also interesting to see which non cardiac features are relevant for the prediction of a heart-related annotation. Thus, Table 6.3 presents the 5 best non heart-related features for the prediction of PEA. 3 features are related to respiration (RSP-Mean, RSPR-Max and Min) and 2 are related to action units (AU4-Mean, AU2-Std). It is also noteworthy that 4 of the 5 best behavioural features for the prediction of PEA are facial features (AU4-Mean and Std, AU2-Std and AU9-Std).

6.4 Discussion

It was expected to see behavioural features perform better than physiological ones for the prediction of the assessment of external observers (EOA) and to see physiological features achieve better performances than behavioural ones for the prediction of the assessment of a physiology expert (PEA). However, it is interesting to notice that behavioural features still achieved good performances for PEA prediction (Figure 6.3, mean F1 score = 0.74), and that physiological features also provided good performances for EOA prediction (Figure 6.1, mean F1 score =

0.78). The fact that we can predict both the assessment of a physiology expert from behavioural features and the assessment of external observers from physiological features shows that there is some coherence between behavioural and physiological cues when one is experiencing stress despite the lack of agreement between EOA and PEA annotations (Tables 3.5 and 3.6). This interplay between physiology and behaviour that we observe in our results is coherent with several works on facial expressions [32, 36, 80], emotions [109], and stress [48, 131].

However, for all 3 assessment sets considered, the combination of behavioural and physiological features provided the best results. It is especially true for SA, for which multimodal features outperform behavioural features by +31% and physiological features by +27%, as shown in Figure 6.2.

Overall, we think that when the obtrusiveness of physiological sensors is acceptable, it is preferable to use a combination of behavioural and physiological features for automatic stress detection. Nonetheless, when unobtrusiveness is required, using only behavioural features still provides good classification performances. These results are coherent with those presented by Giakoumis *et al.* in [45].

We also investigate whether some features provide relevant information for more than one assessment. The results obtained for SA and EOA are similar in some aspects (Tables 6.1 and 6.2). Indeed, the 5 best features for these 2 assessment sets are mainly related to body or body part movement (HeM, HeMZ, HM, IQoM, SQoM and HAPMV). The 5 best features for the prediction of PEA are all related to the heart and belong to 2 categories: heart rate (HR - max and HR - var) and amplitude of the heart rate variability (HRVA - max, HRVA - mean, HRVA - var).

If we look at both the composition of the best subset for each assessment set and the predictive power of each feature (cf. Tables 6.4 and 6.5), it appears that some features provide relevant information for a multi-perspective stress detection. In particular, IQoM achieves good F1 scores for the 3 assessment sets and is present in the best subsets selected for SA and EOA. In general, features related to body or body part movement (IQoM, HeM, HeMZ, HM) provide good results for both SA and EOA prediction. Then, features related to the amplitude of

blood volume pulse are present in all of the 3 best selected subsets, and features related to the raw signal of blood volume pulse provide good classification performances for EOA prediction. Finally, the maximum of the heart rate achieves reasonably good performances and is present in the subsets selected for SA and PEA. These results and the fact that several works report the effect of stress on BVP [40, 56] and heart-rate [70, 123] lead us to conclude that these 2 categories of features, along with body movement, provide valuable information when designing automatic stress detection systems.

Feature	Description	x'	F1 score			In best subset		
			EOA	SA	PEA	EOA	SA	PEA
AU1	Inner Brow Raiser	mean : log	0.614	0.396	0.479			✓
		std : sqrt	0.579	0.476	0.537	✓		
AU2	Outer Brow Raiser	mean : log	0.516	0.416	0.554	✓		
		std : sqrt	0.516	0.429	0.590	✓		✓
AU4	Brow Lowerer	mean : log	0.463	0.431	0.600	✓	✓	
		std : sqrt	0.483	0.449	0.569			
AU5	Upper Lid Raiser	mean : log	0.549	0.447	0.458			
		std : log	0.553	0.470	0.397			
AU6	Cheek Raiser	mean : log	0.608	0.524	0.473	✓	✓	
		std : log	0.630	0.570	0.439		✓	
AU9	Nose Wrinkler	mean : sqrt	0.509	0.487	0.504			
		std : sqrt	0.531	0.496	0.563			
AU12	Lip Corner Puller	mean : log	0.555	0.468	0.438			
		std : sqrt	0.622	0.481	0.395	✓	✓	
AU15	Lip Corner Depressor	mean : log	0.528	0.509	0.465	✓		✓
		std : sqrt	0.594	0.575	0.506			
AU17	Chin Raiser	mean : log	0.596	0.521	0.404	✓		
		std : sqrt	0.578	0.499	0.405		✓	✓
AU20	Lip Stretcher	mean : log	0.499	0.496	0.478			
		std : sqrt	0.590	0.578	0.476	✓		
AU25	Lips Part	mean : log	0.590	0.498	0.467			✓
		std : none	0.561	0.465	0.423			✓
AU26	Jaw Drop	mean : log	0.552	0.497	0.534	✓		✓
		std : log	0.523	0.503	0.468			
SQoM	QoM computed with the skeleton	log	0.625	**0.616**	0.527			
IQoM	QoM computed with the RGB frames	log	**0.723**	**0.621**	0.548	✓	✓	
HAPC	Number of periods of high activity	log	0.626	0.584	0.557			
HAPMD	Mean duration of periods of high activity	log	0.649	0.565	0.579			
HAPMV	Mean highest value of periods of high activity	log	0.661	**0.617**	0.520		✓	
PCC	Number of posture changes	log	0.602	0.544	0.524	✓		
FTC	Number of times face touching with one hand occured	log	0.577	0.511	0.497		✓	✓
FTMD	Mean duration of face touching with one hand	log	0.571	0.510	0.508	✓		✓
FT2HC	Number of times face touching with two hands occured	log	0.411	0.335	0.406		✓	
FT2HMD	Mean duration of face touching with two hands	log	0.457	0.341	0.464			
LHM	QoM for the left hand	log	0.619	0.517	0.470			
RHM	QoM for the right hand	log	0.674	0.602	0.494	✓	✓	
HM	QoM for both hands	log	**0.696**	0.573	0.515			
HeM	QoM for the head	log	**0.780**	**0.614**	0.493	✓		
HeMZ	QoM for the head only along Z-axis	log	**0.716**	0.589	0.482			

Table 6.4: List of the extracted behavioural features. *x'* represents the transformation given by the Box-Cox transformation for each function applied to the signal. *F1 score* displays the results obtained by the each feature when used alone for each assessment set. The 5 best features of each assessment set are in bold. *In best subset* shows whether the feature is present in the best subset selected for each assessment set.

Feature	Description	x'	F1 score			In best subset		
			EOA	SA	PEA	EOA	SA	PEA
BVP	Blood Volume Pulse	mean : none	0.672	0.534	0.626	✓		
		var : log	0.591	0.510	0.447			
		min : none	**0.705**	0.542	0.551		✓	
		max : none	0.435	0.407	0.403		✓	
BVPA	Blood Volume Pulse	mean : log	0.652	0.492	0.450			✓
		var : log	0.696	0.514	0.570		✓	✓
		min : sqrt	0.567	0.442	0.525		✓	✓
		max : sqrt	0.689	0.513	0.555	✓	✓	
EMG	Electromyographic activity of the sternocleidomastoid and upper trapezius - channel 1	mean : none	0.446	0.412	0.457			
		var : log	0.501	0.425	0.389			
		min : none	0.489	0.429	0.471	✓	✓	
		max : none	0.437	0.398	0.469			
EMG2	Electromyographic activity of the sternocleidomastoid and upper trapezius - channel 2	mean : none	0.468	0.493	0.415		✓	
		var : log	0.560	0.557	0.541		✓	
		min : none	0.507	0.470	0.521			✓
		max : log	0.547	0.523	0.567			
EMGMF	Electromyographic activity of the sternocleidomastoid and upper trapezius Mean Frequency	mean : none	0.472	0.424	0.423			✓
		var : none	0.458	0.478	0.468		✓	
		min : log	0.417	0.458	0.405			
		max : none	0.428	0.349	0.393		✓	✓
EMGA	Electromyographic activity of the sternocleidomastoid and upper trapezius Amplitude	mean : sqrt	0.537	0.508	0.490			✓
		var : log	0.661	0.552	0.518			
		min : sqrt	0.512	0.464	0.516			
		max : sqrt	0.603	0.488	0.522			
GSR	Galvanic Skin Response	mean : log	0.487	0.469	0.500			
		var : log	0.478	0.495	0.471	✓	✓	
		min : log	0.487	0.482	0.527			
		max : log	0.476	0.462	0.512			✓
HR	Heart Rate	mean : none	0.510	0.546	**0.711**			✓
		var : log	0.544	0.519	**0.652**			
		min : sqrt	0.502	0.463	0.424			
		max : log	0.553	0.548	**0.701**		✓	✓
HRVA	Heart Rate Variability Amplitude	mean : log	0.529	0.509	**0.681**			
		var : log	0.547	0.553	0.614		✓	✓
		min : log	0.486	0.472	0.569			
		max : sqrt	0.556	0.537	**0.680**			✓
HRV-LF%	Heart Rate Variability Low Frequency zone	mean : sqrt	0.497	0.510	X			
		var : sqrt	0.547	0.464	X			
		min : log	0.552	0.531	X			
		max : none	0.424	0.451	X			
HRV-RMSSD	Heart Rate Variability square root of the mean squared difference between adjacent N-N intervals	log	0.497	0.532	0.630			
HRV-SDNN	Heart Rate Variability Standard Deviation of Normal to Normal intervals	log	0.482	0.475	0.525			
RSP	Chest and abdominal Respiration	mean : log	0.632	0.595	0.621			
		var : log	0.644	0.503	0.471	✓	✓	
		min : log	0.632	**0.609**	0.590			
		max : log	0.581	0.553	0.567			
RSPA	Chest and abdominal Respiration Amplitude	mean : sqrt	0.647	0.426	0.446			
		var : sqrt	0.466	0.487	0.515		✓	✓
		min : log	0.506	0.417	0.417		✓	✓
		max : none	0.461	0.468	0.443		✓	✓
RSPR	Chest and abdominal Respiration Rate	mean : log	0.448	0.444	0.563		✓	
		var : sqrt	0.606	0.525	0.533			✓
		min : log	0.521	0.521	0.587		✓	
		max : log	0.530	0.511	0.617	✓	✓	
RSP+HR	Level of coherence between the Respiration and the Heart Rate	mean : none	0.559	0.497	0.540	✓		
		var : sqrt	0.526	0.569	0.526			
		min : none	0.513	0.449	0.515	✓		
		max : sqrt	0.564	0.558	0.530	✓	✓	
TMP	Temperature	mean : log	0.498	0.415	0.455			
		var : log	0.467	0.348	0.428			
		min : none	0.426	0.497	0.388		✓	
		max : log	0.497	0.384	0.408			

Table 6.5: List of the extracted physiological features. *x'* represents the transformation given by the Box-Cox transformation for each function applied to the signal. *F1 score* displays the results obtained by the each feature when used alone for each assessment set. The 5 best features of each assessment set are in bold. *In best subset* shows whether the feature is present in the best subset selected for each assessment set.

Chapter 7

On leveraging crowdsourced data for automatic stress detection

7.1 Introduction

Resorting to crowdsourcing platforms is a popular way to obtain annotations. Multiple potentially noisy answers can thus be aggregated to retrieve an underlying ground truth. However, it may be irrelevant to look for a unique ground truth when we ask crowd workers for opinions, notably when dealing with subjective phenomena. In the case of stress, we have seen in Chapter 2 that the definition is still debated [70]. As such, one's behavior can hardly be qualified in terms of stress in an objective fashion, but rather as an interpretation that may be subject to interpersonal biases.

In this chapter, we discuss how we can better use crowdsourced annotations for the prediction of the EOA annotation. As seen in Chapter 3, we use a crowdsourcing platform to obtain labels corresponding to videos in which participants are subject to a stress elicitation procedure. A set of workers each labeled the subjects' behavior as either stressed or non-stressed. We study how we can integrate the information from the multiple workers more efficiently than simply performing binary classification upon the labels aggregated with the majority decision. In particular, we propose to learn consensus-weighted predictors and to formulate the prediction problem as a regression on the proportion of positive (i.e. *Stress*) answers. We propose a thorough evaluation of the adaptation of 4 popular machine learning (machine learning) algorithms for handling crowdsourced data, namely Support Vector Machines (SVM), Neural

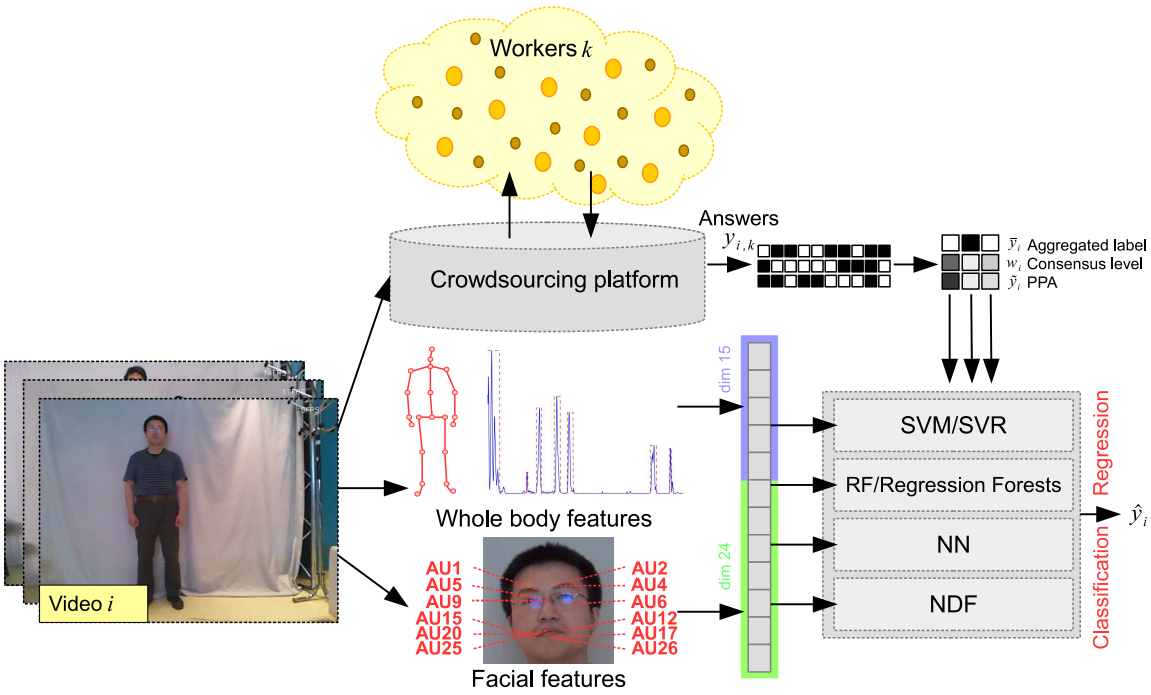

Figure 7.1: Overview of the proposed framework. Videos of recorded subjects are presented to $K = 10$ workers of a crowdsourcing platform who were specifically asked to answer questions regarding the perceived stress level in videos. Those answers are then used to derive annotation labels that are used for automatic perceived stress detection upon a combination of whole body and facial features extracted from the videos. We discuss how the multiple answers from different workers can be integrated for better recognition using a variety of machine learning frameworks.

Networks (NN), Random Forests (RF) and the very recent Neural Decision Forests (NDF). Figure 7.1 summarizes our approach. We show that for the automatic recognition of a subjective phenomenon such as perceived stress, integrating the consensus and proportion of positive answers inside a machine learning framework significantly increases the recognition accuracy. The contributions of this work are the following:

- A case study of modeling crowdsourced data labels to handle the disparity that may exist between the workers' opinions, which consists in measuring the consensus value and proportion of positive answers.
- Propositions on how to train and evaluate machine learning algorithms on crowdsourced data for multiple popular approaches.
- A complete system to perform automatic stress detection from videos, which uses a combination of whole body and facial features and a variety of adaptated machine learning algorithms for classification and regression.

After reviewing some related works, we present the collected labels and the values extracted from them: the binary aggregated label, the consensus level and the proportion of positive answer (PPA). Then, we describe how we adapted the machine learning algorithms for intergrating the consensus and PPA values into the training process. Finally, we present and discuss the results of each experiment.

7.2 Related work

Crowdsourcing has been used to annotate several affective phenomena where no objective ground truth is available, as it allows researchers to collect annotated data at relatively low cost. On platforms such as Amazon Mechanical Turk[1] or CrowdFlower[2], one can design "microtasks" which are performed by people called "crowd workers". Biel *et al.* studied crowdsourced personality impressions using Vlogs from YouTube [10]. They have computed the correlation between personality impressions and social attention measures. Among other results, they have concluded that extraversion is linearly associated with the number of views, the number of times a video is favorited and the number of comments. Soleymani *et al.* used Amazon's Mechanical Turk platform to annotate the boredom response of people to a set of 126 videos [118]. Several works studied subjective quality assessments of images [105, 135] and videos [51].

However, labels obtained through crowdsourcing platforms are often noisy because of the presence of malicious workers and of the different levels of expertise crowd workers have [103]. Thus, the most common way to gather reliable annotations is to collect multiple annotations for each data and then use an aggregation technique to obtain a single label. Aggregation techniques aim at finding the hidden ground truth from a set of possibly noisy answers. As explained in [101], these techniques can be classified into 2 categories: non-iterative and iterative. Majority Decision (MD) is the most common and straightforward non-iterative method. For each data, we count how many times each label has been given as an answer. The label with the highest number of answers is kept as the aggregated label. MD assumes that each worker is equal in skills, which makes it sensitive to spammers and malicious workers. Regarding iterative methods, most of them are extensions of the Expectation Maximization (EM) algorithm. This

[1] www.mturk.com/mturk/welcome
[2] www.crowdflower.com

algorithm performs a series of iterations to update both the aggregated labels and the model parameters. The parameters can include worker expertise [57], self-reported confidence [95], the difficulty of each question [132] or the parameters of the machine learning algorithm which is going to be trained on the aggregated labels [104]. However, finding the hidden ground truth is relevant only when we are trying to label objective phenomena. It may be irrelevant to look for a single ground truth when we ask workers for opinions.

7.3 Collected labels

As explained in Section 3.3.1, we have collected 10 binary labels $y_{i,k}$ for each video. From these answers, we generate 3 values for each video i:

- A binary aggregated label $\bar{y}_i \in \{Stress, Non-Stress\}$, defined as the majority decision (MD): if more than 50% of the workers answered $Non-Stress$ to Q, we assign the $Non-Stress$ label. Otherwise, we assign the $Stress$ label. Note that the data is well balanced w.r.t. aggregated labels, as the proportion of $Non-Stress/Stress$ is 46.2%/53.8%, respectively.
- The consensus level $w_i = 2p - 1 \in [0, 1]$, with p being the proportion of workers who answered $\bar{y}_i$ to Q. Intuitively, this value can be seen as the confidence level we have on a given video. Thus, we consider that there is no consensus ($w_i = 0$) when there are as many answers for the $Stress$ class as there are for the $Non-Stress$ one. Also, by this definition, there is a perfect consensus ($w_i = 1$) when all the workers answered $\bar{y}_i$ to Q.
- The Proportion of Positive Answers (PPA), which is the proportion $\tilde{y}_i$ of workers who answered $Stress$ to question Q. This value is used in the regression experiment as the value to predict for each video, as it encompasses both the stress binary value and the agreement between the workers. Thus, the values $\tilde{y}_i = 0$ and $\tilde{y}_i = 1$ indicate that all the workers answered $Non-Stress$ and $Stress$, respectively. The distribution of this variable is shown in Figure 7.2.

As one can see on Figure 7.2, PPA values $\tilde{y}_i$ are fairly well distributed within the $[0, 1]$ interval. The distribution of the consensus values is also balanced. Thus, it seems relevant to study the

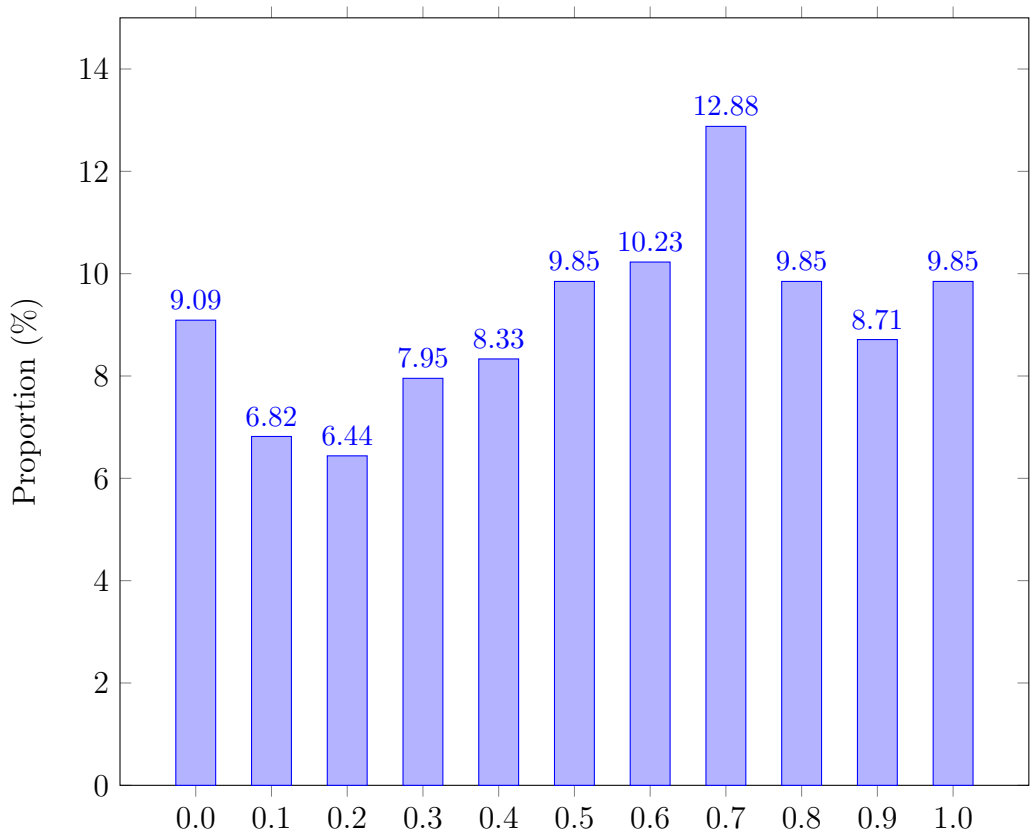

Figure 7.2: Distribution of the regression values $\tilde{y}_i$.

impact of integrating these values into the classification/regression framework to enhance the stress prediction accuracy, as it will be shown in what follows.

7.4 Adaptation of machine learning algorithms for crowdsourced data

In this section, we present how we adapt 4 popular machine learning algorithms to take into account the uncertainty in the crowdsourced annotations, either under the form of a binary classification task or of a regression task.

7.4.1 Motivations

Intuitively, an example that is labelled as *Stress* by 100% of the workers shall be handled differently within the machine learning framework from an example labelled as *Stress* by only 70%. Hence, examples i with a high level of consensus w_i should contribute accordingly to the prediction task, which, *a contrario*, is not the case when we only use the aggregated label for classification. For that matter, we propose a novel way to integrate the level of consensus of a

specific example as a confidence weight in the training process for 4 popular machine learning classification algorithms: Support Vector Machines (SVM), Neural Networks (NN), Random Forests (RF) and the recent Neural Decision Forests (NDF) framework. Specifically, we study the impact of adapting the algorithms on the binary classification performances.

However, it is debatable to consider the aggregated labels as the ground truth for testing, particularly when studying a subjective phenomenon such as perceived stress. Thus, it also makes sense to directly predict the proportion $\tilde{y}_i$ of workers who labelled a given video as *Stress*. To do so, we use the regression counterparts of the SVM(SVR), NN, RF and NDF algorithms.

7.4.2 Machine Learning adaptation

In this section, we describe how we adapt machine learning frameworks to handle crowdsourced data $\boldsymbol{X} = \{x_{i,j}\}$ with $i = 1, ..., N$ and $j = 1, ..., M$ with $N = 264$ the number of examples and $M = 39$ the input feature vector dimension, respectively. We denote the full set of labels $\boldsymbol{Y} = \{y_{i,k}\}$, $i = 1, ..., N$, $k = 1, ..., K$ with K being the (total) number of annotators. Also we denote $\bar{\mathbf{y}} = \{\bar{y}_i\}$ the (binary) aggregation label obtained by MD and $\tilde{\mathbf{y}} = \{\tilde{y}_i\}$ the (continuous) PPA value.

Support Vector Machine

Support Vector Machine (SVM) is a traditional machine learning framework for binary classification [12]. It aims at finding an optimal hyperplane parametrized by a normal vector $\mathbf{w}$, and a bias b, that separates by the widest margin points from 2 classes. For binary classification, a prediction $\hat{y}_i$ is provided by the sign of the scalar product $\mathbf{w}^t\phi(\mathbf{x}_i) + b$. Points can be projected into a transformed feature space by kernel ϕ in order to perform nonlinear classification.

We propose to use an adaptation of the fuzzy SVM extension proposed by Wu *et al.* in [134] to handle crowdsourced data. It introduces fuzzy membership values $\{\mu_i\}_{i=1}^n$ corresponding to each training sample. As stated in [134], the membership value μ_i reflects the fidelity of the data; in other words, how confident we are about the actual class information of the data. We

propose to use the consensus value w_i as the membership value μ_i. The optimization problem is formulated as follows:

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n} w_i\xi_i \\ \text{Subject to} \quad & \bar{y_i}(\mathbf{w}^t\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \quad i = 1, ..., n \end{aligned}$$

Where C is the regularization parameter and ξ_i are the slack variables. For regression, we use the Support Vector Regression (SVR) [117] algorithm to predict $\tilde{y_i}$. In our experiments, we use 3 different kernel functions for both classification and regression, which are the linear, polynomial and Radial Basis Function (RBF) kernels.

Random Forests

Random Forest (RF) is a popular machine learning framework introduced in the seminal work of Breiman [15]. Specifically, given an input $\mathbf{x}_i$, a RF provides a prediction probability $p(\hat{y_i}|\mathbf{x}_i)$ that can be written as the average prediction of T trees $\frac{1}{T}\sum_{t=1}^{T} p_t(\hat{y_i}|\mathbf{x}_i)$. In order to generate accurate and decorrelated individual tree classifiers, Breiman suggests to combine bagging (each tree is grown using only a subset of the N examples) and random subspace (each split node n is set by looking only at a restricted number $k' = \{j_1, ..., j_{k'}\}$ of the input dimension, with $k' < k$). The binary split candidates are thus equal to either $\delta^n(\mathbf{x}_i) = 1$ if a selected dimension $x_{i,j}$ is superior to a threshold θ^n and 0 otherwise (axis-aligned splits). Alternatively, the split candidates are generated under the form of a linear combination of the input dimensions (oblique splits [55]) parameterized by vector β^n: $\delta^n(\mathbf{x}_i) = 1$ if $\sum_j \beta_j^n x_{i,j} - \theta^n > 0$, 0 otherwise.

There exists a number of RF variants [44] that may differ from each other w.r.t. when to set a leaf or a split node, how the split candidates are selected or how the leaf predictions are computed. Our implementation is close to Breiman's original RF [15], in which trees are grown upon bootstraps that each contains approximately 66% of the examples. Then, in the case of a classification problem, for each node, we choose the combination of feature ϕ and threshold θ

that provides the maximal information gain over examples $\{\mathbf{x}_i\}_{i=1...N'}$ falling in current node:

$$H(\phi, \theta) = \sum_{y \in 0,1} -r_y^l log(r_y^l) + \sum_{y \in 0,1} -r_y^r log(r_y^r)$$

Where r_0^l, r_1^l, r_0^r, r_1^r denotes the repartition of label $non - stress$ and $stress$ for left and right subtree, respectively. For instance, $r_0^l = \frac{1}{N'} \sum_i \mathbf{1}(\phi(\mathbf{x}_i) < \theta)$. As for regression we select the candidate that minimizes the average subtree variance. Trees are grown unpruned. For each node, 10 candidate features with 50 candidate thresholds are examined with replacement.

Quite similarly to what is done in [24] for class weighting to deal with unbalanced data, we propose to weight each individual element $\mathbf{x}_i$ by its consensus value w_i for classification. This weight is thus used to weight Shannon's entropy for each element in the computation of the information gain. For instance, we now have $r_0^l = \frac{\sum_i w_i \mathbf{1}(\phi(\mathbf{x}_i) < \theta)}{\sum_i w_i}$. Finally, the consensus value w_i is also used to weight the leaf predictions.

For regression, we simply grow regression trees to predict the PPA value $\tilde{y}_i$.

Neural Networks

Neural Networks (NNs) are perhaps the most famous machine learning framework. Generally speaking, a NN consists of a stack of multiple layers of non-linear neuron units. The output y_i^l of a neuron i of layer l consists in (a) a scalar product between this layer's input $\{x_{i,j}^l\}$ and it's weights $w_{i,j}^l$ and (b) a non-linear activation function σ: $y_i^l = \sigma(\sum_j w_{i,j}^l x_{i,j}^l + b)$. Training is usually performed via Stochastic Gradient Descent (SGD), by first computing the networks' activations in a feed-forward manner given a specific example $\mathbf{x}_i$, then by backpropagating the error between the top layer prediction and the ground truth label (for classification) or value (for regression).

In our experiments, we use a 2-layer NN with a single 100-units sigmoid hidden layer and an output 2-units softmax layer for classification, or a single-unit sigmoid layer to perform

regression on the PPA $\tilde{y}_i$. Furthermore, for classification, we propose to weight the learning rate with the consensus value w_i for each example.

Neural Decision Forests

Neural Decision Forests (NDFs [69]) are a recent NN/RF hybrid algorithm that consists in a collection of differential decision trees with oblique probabilistic split nodes. More specifically, for a split node n example $\mathbf{x}_i$ goes to the right subtree with a probability $d^n(\mathbf{x}_i) = \sigma(\sum_j \beta_j^n x_{i,j} - \theta^n)$ and to the left subtree with a probability $1 - d^n(\mathbf{x}_i)$. Consequently, each leaf node l of a tree t is reached with probability μ^l that can be estimated as the product of the neurons' activations throughout the tree, from the root node to leaf l. Each leaf node l of tree t contains either a two-dimensional probability distribution $p_t^l(\hat{y}_i)$ in the case of a classification task, or directly an estimation $\hat{y}$ for regression. As in RFs, the final probability to predict $\hat{y}$ is provided by the average among the trees in the forest, *i.e.* $p(\hat{y}_i|\mathbf{x}_i) = \frac{1}{T}\sum_{t=1}^{T}\sum_l \mu^l(\mathbf{x}_i)p_t^l(\hat{y}_i)$.

Note that our NDF implementation differs from the one proposed in [69], as we obtained satisfying results without having to periodically update the leaf nodes after a number of epochs, as suggested in the paper. Instead, we initialize the predictions with pure distribution (either $(0,1)$ or $(1,0)$ respectively for the $(Non-Stress, Stress)$ classes) for binary classification or with randomly sampled prediction values in the interval $[0,1]$ for regression. Then, since NDFs are differential models, they can be trained similarly to NNs using SGD and error backpropagation. We thus sequentially apply SGD updates to the split nodes' parameters for a number of epochs without altering the leaf predictions (see [69] for a more thorough description of the optimization). Furthermore, as NDF training is performed similarly to NNs, we use the same confidence-weighting of the learning rate for classification task. For regression, as explained above, we directly learn to predict the PPA $\tilde{y}_i$ using a $\mathcal{L}_2$-loss function.

7.5 Experiments

In this section, we present the results obtained for both the classification and regression experiments. We also present the evaluation process for both experiments.

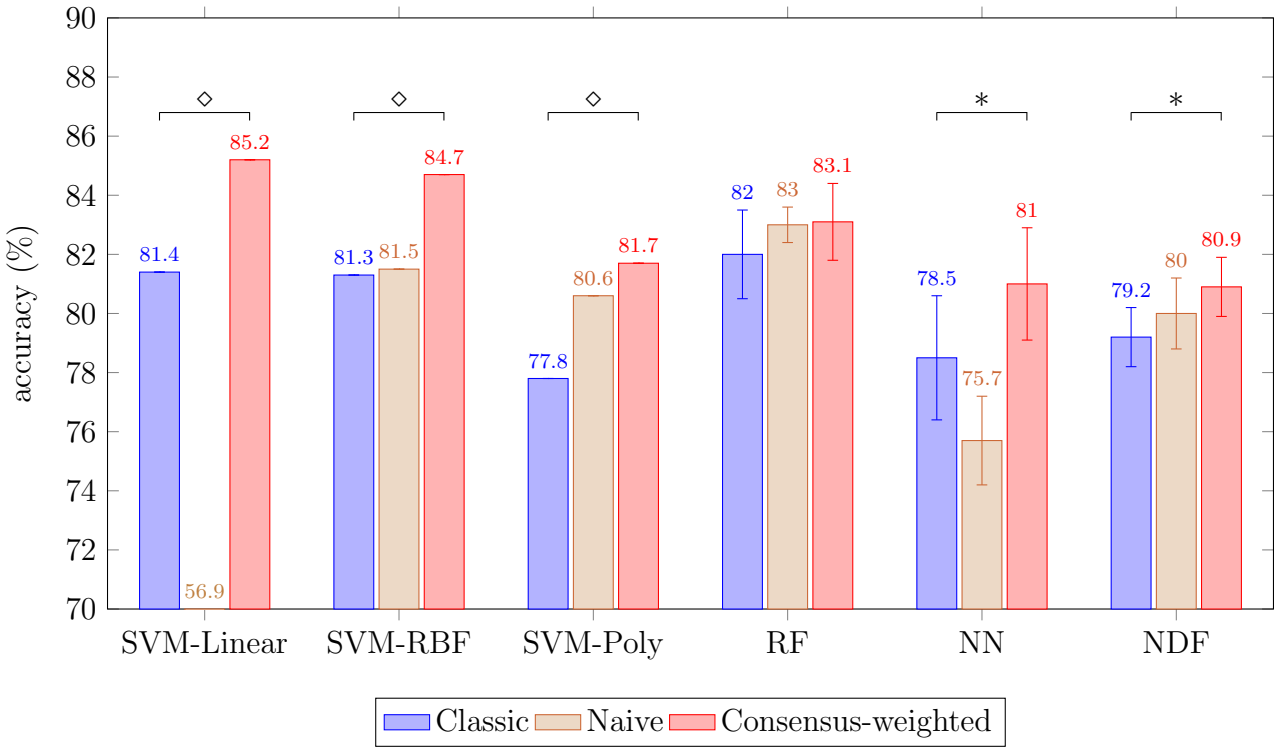

Figure 7.3: Classification Results ($* \ p < 0.05$ for Student's t-test, $\diamond$ for deterministic results)

7.5.1 Evaluation process

For both experiments, we perform 10-fold subject-independent cross-validation: the data of 10% of the subjects are used as testing samples while the data of the remaining 90% are used as training samples. In order to be able to faithfully reproduce the evaluation conditions, the composition of each fold and the algorithm parameters are fixed. For classification, we use the binary aggregated labels $\bar{\mathbf{y}}$ as the ground truth. We use the overall accuracy as the evaluation metric since the distribution of binary labels is balanced (see Section 7.3). For each algorithm, we evaluate 3 versions:

- A **Classic** version trained on the labels aggregated with majority decision.
- A straightforward, **Naive** adaptation trained on all the labels. More specifically, training samples are duplicated $K = 10$ times each and associated with the answers $y_{i,k}$. Consequently, a training sample can have contradictory labels if its level of consensus is not perfect, which can be a hindrance for the stability of certain algorithms, as it will be shown in what follows.
- The **Consensus-weighted** version that was introduced in Section 7.4.

Since NN, RF and NDF training involve random initialization and/or data sampling, we present the average accuracy and its standard deviation over 10 runs for those algorithms. For regression, we use the Mean Square Error (MSE) and the Correlation Coefficient (CC) as the evaluation metrics, as both are complimentary. We also present the average results over 10 runs for NNs, RFs and NDFs.

7.5.2 Classification

Figure 7.3 presents the average accuracy obtained by each version of each algorithm. One can see that the naive version slightly improves the classification accuracy for 4 algorithms: SVM-RBF, SVM-Poly, RF and NDF, with significant improvement only for SVM-Poly. However, NN and *a fortiori* SVM-Linear seem to face some issues with contradictory labels. Indeed, the accuracy obtained by the naive version of NN is significantly lower than the accuracy obtained by the classic version (from 74.3% down to 72.6%, $p < 0.05$). This is likely due to the training procedure at stake, where applying SGD updates with multiple versions of the same examples with contradictory labels may cause instability in the learning procedure. This can be observed in Figure 7.4, where it appears that the training error of the naive version is much less stable than for the classic and consensus-weighted versions. Moreover the Consensus-weighted version does not fit the examples that are the most uncertain w.r.t their consensus values, hence a training error that lies slightly above that of the Classic version. As for SVM, the average accuracy remains approximately the same when using either the naive or the classic version of the SVM-RBF (from 81.3% to 81.5%). The naive version of the SVM-Poly actually outperforms the classic one with an improvement of 3.6% (from 77.8% to 80.6%). However, the classification accuracy for the SVM-linear is much lower with the naive version than with the classic one (from 81.4% to 56.9%). This performance gap is most likely due to the instability caused by the contradictory labels on the margin optimization procedure, similarly to what happens for NN training (see Figure 7.4).

Moreover, using the consensus level to weight the individual training examples seems to benefit all the machine learning algorithms, compared to the classic version. This improvement seems significant for the SVM with the 3 kernel functions: +4.5% for the linear kernel (from 81.5% to

85.2%), +4.2% for the RBF kernel (from 81.3% to 84.7%) and +5.0% for the polynomial kernel (from 77.8% to 81.7%). According to a Student's t-test, the improvement is considered to be statistically significant for the NN ($p = 0.0121$) and the NDF ($p = 0.0013$), but is not considered significant for the RF ($p = 0.0967$). Overall, the best classification accuracy is provided by the consensus-weighted version of the SVM-linear. Tables 7.5 and 7.6 respectively display the confusion matrix of the SVM-linear for the classic version and the consensus-weighted version, empathizing the fact that integrating the consensus level allows to increase both the true negative and true positive rates, while diminishing the false positive and false negative rates.

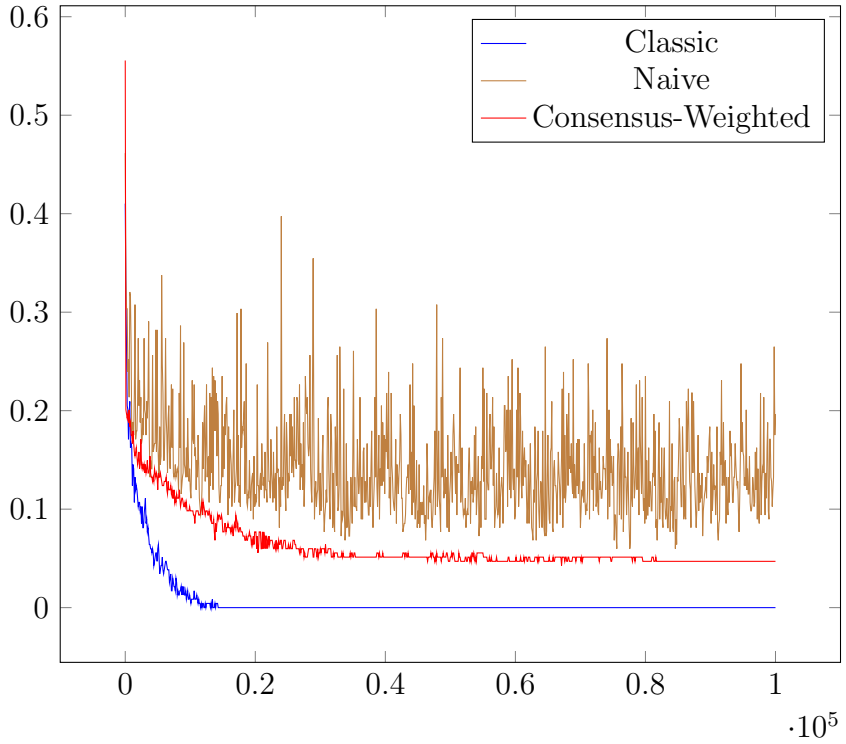

Figure 7.4: Evolution of the training error through the updates for each version of NN. The training error is computed every 100 updates on the first fold.

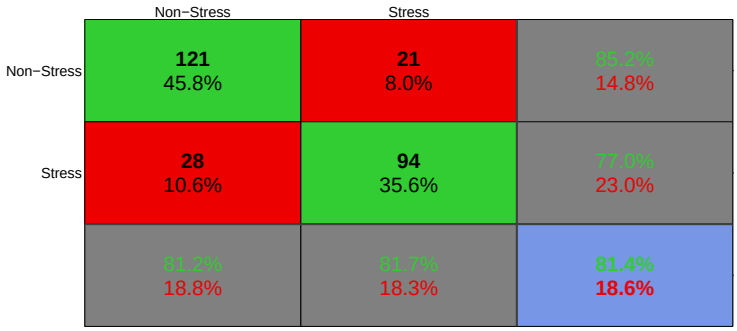

Figure 7.5: Confusion matrix for classic SVM-linear. Best viewed in color.

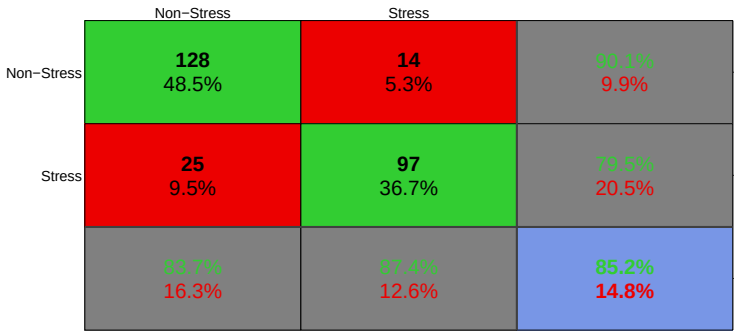

Figure 7.6: Confusion matrix for consensus-weighted SVM-linear. Best viewed in color.

7.5.3 Regression of the PPA

Table 7.1 presents the results for the regression experiment. We can see that all the algorithms provide a low MSE and a high CC. The average squared prediction error is about 5%, except for the NN where it is about 7%. Note that the predicted value integrates the information of perceived stress level as well as the consensus level among the workers. Thus, the fact that we can obtain satisfying results with several machine learning algorithms suggests that the PPA is a valuable information that can be efficiently modeled and generalized across the subjects. Furthermore, the PPA values are fairly well distributed within the $[0, 1]$ interval, ensuring a low prediction bias. As such, this quantity appears as a valuable information to train and evaluate machine learning algorithms with more precise information than a single label aggregated using MD.

Algorithm	**MSE**	**CC**
SVR-Linear	0.047	0.699
SVR-RBF	0.046	0.688
SVR-Poly	0.051	0.663
RF	0.050 ± 0.001	0.680 ± 0.009
NN	0.071 ± 0.006	0.699 ± 0.046
NDF	0.049 ± 0.001	0.699 ± 0.007

Table 7.1: MSE and CC for regression using the different methods.

7.6 Conclusion

In this chapter, we have presented a framework to perform automatic perceived stress detection from crowdsourced data. To do so, we used the *Dataset-44* and the EOA annotation, both introduced in Chapter 3.

We studied how information such as the level of consensus among workers and the proportion of positive answers (PPA) can be used to train and evaluate machine learning algorithms for the prediction of perceived stress. To do so, we conducted a classification and a regression experiment.

Regarding the classification experiment, we proposed a way to integrate the level of consensus in the training process for 4 classification algorithms: SVM, RF, NN and NDF. Then, we compared the classification accuracy between 3 different versions of each algorithm. Among those versions, the consensus-weighted version significantly outperforms the two other ones for almost every classification algorithm.

Moreover, in the regression experiment, we tried to directly predict the proportion of votes for the *Stress* class (PPA) for each video, using the regression counterparts of each machine learning algorithm: SVR, RF, NN and NDF. The high accuracies obtained indicate that the level of consensus and the proportion of votes are valuable information that are generalisable enough to either enhance classification performances or be efficiently predicted.

Chapter 8

Conclusion

In this Chapter, we summarize the thesis achievements and the publications obtained during this period. We also discuss the possible applications of this work as well as its evolutions and perspectives.

8.1 Summary of thesis achievements

In this thesis, we developed an automatic stress detection framework, with a focus on stress assessment strategies. We saw in Chapter 2 that there are several ways to assess stress. According to the biological perspective, stress can be assessed using biomarkers such as the cortisol level, the skin conductance of the heart-rate variability. According to the phenomenological perspective, stress can be self-assessed by subjects. Finally, according to the behavioural perspective, stress can be assessed on the basis of behaviour modifications. In the same chapter, we also observed that previous works used a wide variety of assessment strategies, making it difficult to provide a fair comparison of the performances of these frameworks.

In Chapter 3, we described how we collected stress data with multiple assessments. Our stress elicitation procedure is a socially evaluated mental arithmetic test composed of 6 steps of increasing difficulty. We acquired Kinect and HD video data from 44 subjects. We also collected physiological data for 21 of the 44 subjects. For each subject and each step of the test, stress is assessed in 3 ways, based on the 3 perspectives presented in Chapter 2:

- Crowdworkers provide External Observers Assessment (EOA).

- Subjects of the experiment provide Self-Assessment (SA).
- A physiology expert provides Physiology Expert Assessment (PEA)

We described the features extracted from the acquired data in Chapter 4. We presented original body features for stress detection such as periods of high activity, posture changes, face touching and fingers rubbing. We also extracted the level of activation of 12 Action Units and physiological features which are traditionally used in stress detection frameworks.

In Chapter 5, we studied how 5 data normalization methods may help reduce the impact of interindividual differences. We also evaluated the impact of assessment strategies on the performance of normalization methods. On one hand, we concluded that person-specific normalization are efficient for the prediction of SA, but are actually deleterious for prediction of EOA. On the other hand, we concluded that the Box-Cox transformation - the only non person-specific normalization method that we evaluated - provide pretty good results for the prediction of SA, but is especially relevant for the prediction of EOA.

In Chapter 6, we evaluated the classification performance of 101 features and 2 modalities (behaviour and physiology) for the prediction of the 3 assessments presented in Chapter 3. We showed that assessment strategies greatly impact the performance of modalities. On one hand, behavioural features performed significantly better than physiological ones for the prediction of EOA, which is based on subjects' behaviour. On the other hand, physiological features performed significantly better for the prediction of PEA, which is based on subjects' physiology. Moreover, we made two noteworthy observations from the results of our experiments. First, it seems necessary to use multimodal features to predict SA. Second, there seems to be an interplay between physiology and behaviour, as it turned out to be possible to predict with good accuracy how stressed one appears using only physiological features and how stressed one's body is using only behavioural features. Regarding the evaluation of feature performance, we observed that feature related to blood volume pulse, heart-rate and body movement provide valuable information for several assessment strategies.

Finally, we presented a framework for handling crowdsourced labels. We showed that one can extract valuable information from crowdworkers' answers, such as the level of consensus or

the proportion of positive answers. These information can then be used to train and evaluate machine learning algorithms. We proposed a way to integrate the level of consensus in the training phase of 4 machine learning algorithms: SVM, random forest, neural networks and neural decision forest. We showed that using a consensus-weighted version provide significantly better classification results for almost every classification algorithms that we evaluated. We also showed that the proportion of positive answers can be accurately predicted using regression algorithms.

Overall, our main conclusion from the findings of this thesis is that one has to take a global and comprehensive approach to design affect recognition solutions. It is especially true when choosing an assessment strategy, as it impacts the performance of data normalization methods (Chapter 5), the classification performance of features and modalities (Chapter 6) and the design of machine learning algorithms (Chapter 7).

8.2 Applications

There are two main fields in which automatic stress detection frameworks are usually applied: human-computer interaction and healthcare. Regarding human-computer interactions, several works already applied their solutions to specific problematics. For instance, the Tardis project [4] aims at building a serious-game which simulates job interviews. In this context, applicants are likely to experience stress, which may have a negative effect on their performances. Thus, the framework detects stress from behaviour in order to provide feedback to the user. This feedback can then be used by the user to improve her body language and her awareness about how much stress she experienced. In her PhD thesis [76], Lefter stated that automatic stress detection would improve the performance of video surveillance systems, as aggressive behaviour is sometimes linked with stress. Therefore, as said by Lefter, *"Detection of stress and negative emotions in an early stage is very valuable since it can help prevent aggression and other unwanted situations."*. Regarding healthcare, several works proposed mobile architectures to monitor stress in everyday life [20, 124, 133]. These works aims at providing support to clinician for decision making and diagnosis.

Regarding the application of our framework, it is going to be used in a clinical project on borderline personality disorder (BPD) in teenagers, in association with the department of Child and Adoslescent Psychiatry at La Salpétrière hospital. BPD is associated with intense emotional and behavioural responses to stressful events, characterized by highly negative emotions, impulsivity, and risk-taking behaviors. These responses are often associated with high morbidity, including substance use problems, self-harm and frequent and severe social conflicts. BPD has been shown to begin in adolescence. Despite the severity of this disorder, very few studies have addressed the physiopathology of BPD in adolescents. Thus, the presented framework will be used along structural and functional imaging to study the dimensional aspects of the disorder. The objective of this project is to better understand the psychophysiopathology of BPD.

8.3 Perspectives

As explained in the previous section, one of the perspectives for this work is to use the framework we presented to study how specific populations handle stress. Post-traumatic stress disorder and anxiety disorder are also examples of pathologies characterized by an acute response to certain or all stressful situations. Studying how stress impact the behaviour and the physiology of people affected by these disorders can provide valuable information about the functioning of these disorders.

Another perspective would be to study the feasibility of a stress assessment strategy that would take into account the multidimensional nature of the expression of stress. This would improve the theoretical soundness and the robustness of stress detection solutions. It would also allow comparison between these solutions, which would accelerate research on this problematic.

Finally, we believe that it would be interesting to explore in more depth the interplay between physiology and behaviour that we observed in our results. In particular, we think that it is important to study the temporal aspect of this interplay. As we discussed in Chapter 2, physiological, emotional and behavioural responses are triggered with different timings. Therefore, it is necessary to take this fact into account to study whether some pattern of beha-

viour/thought/physiological changes are caused by specific events/stimulus. Thus, it would be necessary to design a new experiment for which sensors are accurately synchronized.

8.4 Publications

Journal paper

J. Aigrain, M. Spodenkievicz, S. Dubuisson, M. Detyniecki, D. Cohen & M. Chetouani (2016). Multimodal stress detection from multiple assessments. *Submitted in Transactions on Affective Computing. Minor Revision.*

Conference papers

J. Aigrain, A. Dapogny, K. Bailly, S. Dubuisson, M. Detyniecki & M. Chetouani (2016). On leveraging crowdsourced data for automatic perceived stress detection. *Proceedings of the International Conference on Multimodal Interaction. Tokyo, 2016.*

J. Aigrain, S. Dubuisson, M. Detyniecki & M. Chetouani (2015). Person-specific behavioural features for automatic stress detection. *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition: Context Based Affect Recognition. Ljubljana, 2015.*

Bibliography

[1] J. G. Adair. The hawthorne effect: A reconsideration of the methodological artifact. *Journal of applied psychology*, 69(2):334, 1984.

[2] T. C. Adam and E. S. Epel. Stress, eating and the reward system. *Physiology & behavior*, 91(4):449–458, 2007.

[3] B. Ahmed, H. M. Khan, J. Choi, and R. Gutierrez-Osuna. ReBreathe: A Calibration Protocol that Improves Stress/Relax Classification by Relabeling Deep Breathing Relaxation Exercises. *IEEE Transactions on Affective Computing*, 7(2):150–161, 2016.

[4] K. Anderson, E. André, T. Baur, S. Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, et al. The Tardis framework: intelligent virtual agents for social coaching in job interviews. In *Advances in computer entertainment*, pages 476–491. Springer, 2013.

[5] G. Andrews, M. J. Hobbs, T. D. Borkovec, K. Beesdo, M. G. Craske, R. G. Heimberg, R. M. Rapee, A. M. Ruscio, and M. a. Stanley. Generalized worry disorder: a review of DSM-IV generalized anxiety disorder and options for DSM-V. *Depression and anxiety*, 27(2):134–47, Feb. 2010.

[6] A. Barreto, J. Zhai, and M. Adjouadi. Non-intrusive Physiological Monitoring for Automated Stress Detection in Human-Computer Interaction. In *Human-Computer Interaction*, pages 29–38, 2007.

[7] C. Becker, H. Prendinger, M. Ishizuka, and I. Wachsmuth. Evaluating affective feedback of the 3D agent Max in a competitive cards game. *Lecture Notes in Computer*

Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3784 LNCS:466–473, 2005.

[8] A. Ben-Hur and J. Weston. A user's guide to support vector machines. *Data mining techniques for the life sciences*, pages 223–239, 2010.

[9] D. Bernhardt and P. Robinson. Detecting affect from non-stylised body motions. In *Affective Computing and Intelligent Interaction*, pages 59–70, 2007.

[10] J. I. Biel and D. Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55, 2013.

[11] S. Boonnithi and S. Phongsuphap. Comparison of heart rate variability measures for mental stress detection. In *Computing in Cardiology*, volume 38, pages 85–88, 2011.

[12] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[13] D. Bozovic, M. Racic, and N. Ivkovic. Salivary cortisol levels as a biological marker of stress reaction. *Medical archives*, 67(5):374, 2013.

[14] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[15] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[16] E. Brunswik. *Perception and the representative design of psychological experiments*. Univ of California Press, 1956.

[17] L. a. Bugnon, R. a. Calvo, and D. H. Milone. A Method for Daily Normalization in Emotion Recognition. *15th Argentine Symposium on Technology*, pages 48–59, 2014.

[18] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan. Iterative feature normalization scheme for automatic emotion detection from speech. *IEEE Transactions on Affective Computing*, 4(4):386–397, 2013.

[19] W. B. Cannon. The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory. *The American journal of psychology*, 39(1/4):106–124, 1927.

[20] N. Carbonaro, P. Cipresso, A. Tognetti, G. Anania, D. De Rossi, A. Gaggioli, and G. Riva. A mobile biosensor to detect cardiorespiratory activity for stress tracking. In *7th International Conference on Pervasive Computing Technologies for Healthcare*, number January 2016, pages 440–445, 2013.

[21] G. Caridakis, A. Raouzaiou, K. Karpouzis, and S. Kollias. Synthesizing Gesture Expressivity Based on Real Sequences. In *Workshop Multimodal Corpora. From Multimodal Behaviour Theories to Usable Models. In: International Conference on Language Resources and Evaluation*, pages 19–23, 2006.

[22] D. Carroll, G. D. Smith, M. J. Shipley, A. Steptoe, E. J. Brunner, and M. G. Marmot. Blood pressure reactions to acute psychological stress and future blood pressure status: a 10-year follow-up of men in the whitehall ii study. *Psychosomatic Medicine*, 63(5):737–743, 2001.

[23] M. Chance. *An interpretation of some agonistic postures; the role of "cut-off" acts and postures.* Symposia of the Zoological Society of London, 1962.

[24] C. Chen, A. Liaw, and L. Breiman. Using random forest to learn imbalanced data. *University of California, Berkeley*, 2004.

[25] T. Chen, P. Yuen, M. Richardson, G. Liu, Z. She, and S. Member. Detection of Psychological Stress Using a Hyperspectral Imaging Technique. *IEEE Transactions on Affective Computing*, 5(4):391–405, 2014.

[26] H. Cohen, J. Benjamin, A. B. Geva, M. a. Matar, Z. Kaplan, and M. Kotler. Autonomic dysregulation in panic disorder and in post-traumatic stress disorder: Application of power spectrum analysis of heart rate variability at rest and in response to recollection of trauma or panic attacks. *Psychiatry Research*, 96(1):1–13, 2000.

[27] S. Cohen, D. Janicki-deverts, and G. E. Miller. Psychological Stress and Disease. *Journal of American Medical Association*, 298(14):1685–1687, 2007.

[28] S. Cohen, T. Kamarck, and R. Mermelstein. A Global Measure of Perceived Stress. *Journal of health and social behavior*, 24(4):385, Dec. 1983.

[29] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7, Sept. 2009.

[30] C. L. Cooper and S. Cartwright. An intervention strategy for workplace stress. *journal of psychosomatic research*, 43(1):7–16, 1997.

[31] T. Dalgleish. The emotional brain. *Nature Reviews Neuroscience*, 5(7):583–589, July 2004.

[32] R. J. Davidson, P. Ekman, C. D. Saron, J. a. Senulis, and W. V. Friesen. Approach-withdrawal and cerebral asymmetry: emotional expression and brain physiology. I., 1990.

[33] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012.

[34] D. R. Denney and M. B. Frisch. The role of neuroticism in relation to life stress and illness. *Journal of Psychosomatic Research*, 25(4):303–307, 1981.

[35] S. S. Dickerson and M. E. Kemeny. Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. *Psychological bulletin*, 130(3):355–91, May 2004.

[36] P. Ekman, R. J. Davidson, and W. V. Friesen. The Duchenne smile: emotional expression and brain physiology. II. *Journal of personality and social psychology*, 58(2):342–353, 1990.

[37] P. Ekman and W. V. Friesen. Facial action coding system, 1977.

[38] E. Elenko, L. Underwood, and D. Zohar. Defining digital medicine. *Nature biotechnology*, 2015.

[39] R. Fernandez and R. W. Picard. Modeling drivers' speech under stress. *Speech communication*, 40(1):145–159, 2003.

[40] L. Finsen, K. Søgaard, C. Jensen, V. Borg, and H. Christensen. Muscle activity and cardiovascular response during computer-mouse work with and without memory demands. *Ergonomics*, 44(14):1312–1329, 2001.

[41] G. L. Flett, P. L. Hewitt, and D. G. Dyck. Self-oriented perfectionism, neuroticism and anxiety. *Personality and Individual Differences*, 10(7):731–735, 1989.

[42] H. Gao, A. Yüce, and J.-P. Thiran. Detecting emotional stress from facial expressions for driving safety. In *IEEE International Conference on Image Processing (ICIP)*, volume 1, pages 5961–5965, 2014.

[43] A. Garg, M.-M. Chren, L. P. Sands, M. S. Matsui, K. D. Marenus, K. R. Feingold, and P. M. Elias. Psychological stress perturbs epidermal permeability barrier homeostasis: implications for the pathogenesis of stress-associated skin disorders. *Archives of dermatology*, 137(1):53–59, 2001.

[44] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[45] D. Giakoumis, A. Drosou, P. Cipresso, D. Tzovaras, G. Hassapis, A. Gaggioli, and G. Riva. Using activity-related behavioural features towards more effective automatic stress detection. *PloS one*, 7(9):e43571, 2012.

[46] T. Giraud, M. Soury, J. Hua, A. Delaborde, M. Tahon, D. A. G. Jauregui, V. Eyharabide, E. Filaire, C. Le Scanff, L. Devillers, B. Isableu, and J. C. Martin. Multimodal Expressions of Stress during a Public Speaking Task: Collection, Annotation and Global Analyses. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 417–422, Sept. 2013.

[47] D. Glowinski, N. Dael, a. Camurri, G. Volpe, M. Mortillaro, and K. Scherer. Toward a Minimal Representation of Affective Gestures. *IEEE Transactions on Affective Computing*, 2(2):106–118, 2011.

[48] M. R. Gunnar. Psychobiological studies of stress and coping: An introduction. *Child Development*, pages 1403–1407, 1987.

[49] I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[50] M. a. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, 1999.

[51] P. Hanhart, P. Korshunov, and T. Ebrahimi. Crowd-based quality assessment of multiview video plus depth coding. *IEEE International Conference on Image Processing*, pages 743–747, 2014.

[52] J. H. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom. Getting started with susas: a speech under simulated and actual stress database. In *Eurospeech*, volume 97, pages 1743–46, 1997.

[53] J. Harrigan. Self-touching as an indicator of underlying affect and language processes. *Social Science and medicine*, 20(11):1161–1168, 1985.

[54] J. Healey and R. Picard. Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166, June 2005.

[55] D. Heath, S. Kasif, and S. Salzberg. Induction of oblique decision trees. In *Journal of Artificial Intelligence Research*, pages 1002–1007, 1993.

[56] N. Hjortskov, D. Rissén, A. K. Blangsted, N. Fallentin, U. Lundberg, and K. Sogaard. The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology*, 92(1-2):84–89, 2004.

[57] P. G. Ipeirotis, F. Provost, and J. Wang. Quality Management on Amazon Mechanical Turk. *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 64–67, 2010.

[58] S. C. Jacobs, R. Friedman, J. D. Parker, G. H. Tofler, A. H. Jimenez, J. E. Muller, H. Benson, and P. H. Stone. Use of skin conductance changes during mental stress

testing as an index of autonomic arousal in cardiovascular research. *American Heart Journal*, 128(6):1170–1177, Dec. 1994.

[59] J. Joshi, R. Goecke, G. Parker, and M. Breakspear. Can body expressions contribute to automatic depression analysis? *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013.

[60] J. F. Kaiser. On a simple algorithm to calculate the energy of a signal. *International Conference on Acoustics, Speech, and Signal Processing*, 2(10):381–384, 1990.

[61] M. Kalia. Assessing the economic impact of stress [mdash] the modern day hidden epidemic. *Metabolism*, 51(6):49–53, 2002.

[62] A. Kapoor, W. Burleson, and R. W. Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736, Aug. 2007.

[63] G. Keinan. Decision making under stress: Scanning of alternatives under controllable and uncontrollable threats. *Journal of personality and social psychology*, 52(3):639, 1987.

[64] S. Kirkpatrick, M. Vecchi, et al. Optimization by simmulated annealing. *Science*, 220(4598):671–680, 1983.

[65] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer. The trier social stress test.pdf. *Neuropsychobiology*, 28:76–81, 1993.

[66] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

[67] S. Koldijk, M. Sappelli, S. Verberne, M. a. Neerincx, and W. Kraaij. The SWELL Knowledge Work Dataset for Stress and User Modeling Research. *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 291–298, 2014.

[68] M. Kompier and C. L. Cooper. *Preventing stress, improving productivity: European case studies in the workplace*. Psychology Press, 1999.

[69] P. Kontschieder, M. Fiterau, A. Criminisi, and S. Rota Bulo. Deep neural decision forests. In *International Conference on Computer Vision*, pages 1467–1475, 2015.

[70] J. M. Koolhaas, a. Bartolomucci, B. Buwalda, S. F. de Boer, G. Flügge, S. M. Korte, P. Meerlo, R. Murison, B. Olivier, P. Palanza, G. Richter-Levin, a. Sgoifo, T. Steimer, O. Stiedl, G. van Dijk, M. Wöhr, and E. Fuchs. Stress revisited: a critical evaluation of the stress concept. *Neuroscience and biobehavioral reviews*, 35(5):1291–301, Apr. 2011.

[71] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[72] R. S. Lazarus. Psychological stress and the coping process. 1966.

[73] R. S. Lazarus. From Psychological Stress to the Emotions: A History of Changing Outlooks. *Annual Review of Psychology*, 44(1):1–22, Jan. 1993.

[74] R. S. Lazarus. Psychological stress in the workplace. *Occupational stress: A handbook*, 1:3–14, 1995.

[75] J. LeDoux. Rethinking the Emotional Brain. *Neuron*, 73(4):653–676, Feb. 2012.

[76] I. Lefter. *Multimodal Surveillance Behavior analysis for recognizing stress and aggression*. PhD thesis, 2014.

[77] I. Lefter, G. Burghouts, and L. Rothkrantz. Recognizing stress using semantics and modulation of speech and gestures. *IEEE Transactions on Affective Computing*, 3045(c):1–1, 2015.

[78] I. Lefter, L. J. Rothkrantz, D. A. van Leeuwen, and P. Wiggers. automatic stress dectection in emergency calls. *International Journal of Intelligent Defence Support Systems*, 4(2):148–168, 2011.

[79] F. X. Lesage, S. Berjot, and F. Deschamps. Clinical stress assessment using a visual analogue scale. *Occupational medicine*, 62(8):kqs140–605, Sept. 2012.

[80] R. W. Levenson, L. L. Carstensen, W. V. Friesen, and P. Ekman. Emotion, physiology, and expression in old age. *Psychology and aging*, 6(1):28–35, 1991.

[81] C. L. Lisetti and F. Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP Journal on Advances in Signal Processing*, 2004(11):1–16, 2004.

[82] A. D. Lopez and C. J. Murray. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. Harvard School of Public Health, 1996.

[83] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[84] S. J. Lupien, F. Maheu, M. Tu, a. Fiocco, and T. E. Schramek. The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. *Brain and cognition*, 65(3):209–37, Dec. 2007.

[85] Y. Lutchyn, G. Mark, A. Sano, P. Johns, M. Czerwinski, and S. Iqbal. Stress is in the eye of the beholder. In *Affective Computing and Intelligent Interaction (ACII)*, 2015.

[86] K. B. Matheny, D. W. Aycock, W. L. Curlette, and G. N. Junker. The coping resources inventory for stress: A measure of perceived resourcefulness. *Journal of Clinical Psychology*, 49(6):815–830, Nov. 1993.

[87] I. McDowell. *Measuring health: a guide to rating scales and questionnaires*. Oxford University Press, 2006.

[88] A. McVicar. Workplace stress in nursing: a literature review. *Journal of advanced nursing*, 44(6):633–642, 2003.

[89] C. Mohiyeddini, S. Bauer, and S. Semple. Displacement Behaviour Is Associated with Reduced Stress Levels among Men but Not Women. *PloS one*, 8(2):e56355–9, Feb. 2013.

[90] C. Mohiyeddini and S. Semple. Displacement behaviour regulates the experience of stress in men. *Stress*, 16(2):163–171, Sept. 2012.

[91] H. Mönnikes, J. Tebbe, M. Hildebrandt, P. Arck, E. Osmanoglou, M. Rose, B. Klapp, B. Wiedenmann, and I. Heymann-Mönnikes. Role of stress in functional gastrointestinal disorders. *Digestive Diseases*, 19(3):201–211, 2001.

[92] A. Moriguchi, A. Otsuka, K. Kohara, H. Mikami, K. Katahira, T. Tsunetoshi, K. Higashimori, M. Ohishi, Y. Yo, and T. Ogihara. Spectral change in heart rate variability in response to mental arithmetic before and after the beta-adrenoceptor blocker, carteolol. *Clinical Autonomic Research*, 2(4):267–270, 1992.

[93] D. K. Mroczek and D. M. Almeida. The effect of daily stress, personality, and age on daily negative affect. *Journal of personality*, 72(2):355–378, 2004.

[94] J. Nicolle, K. Bailly, and M. Chetouani. Facial Action Unit Intensity Prediction via Hard Multi-Task Metric Learning for Kernel Regression. *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015.

[95] S. Oyama, Y. Baba, Y. Sakurai, and H. Kashima. Accurate integration of crowdsourced labels using workers' self-reported confidence scores. *International Joint Conference on Artificial Intelligence*, pages 2554–2560, 2013.

[96] F. Ozel. Time pressure and stress as a factor during emergency egress. *Safety Science*, 38(2):95–107, 2001.

[97] L. Pessoa and R. Adolphs. Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nature Reviews Neuroscience*, 11(11):773–783, Nov. 2010.

[98] R. W. Picard. Affective computing: Challenges. *International Journal of Human Computer Studies*, 59(1-2):55–64, 2003.

[99] R. W. Picard and R. Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[100] K. Plarre, A. Raij, S. M. Hossain, A. A. Ali, M. Nakajima, M. Absi, E. Ertin, T. Kamarck, S. Kumar, M. Scott, D. Siewiorek, A. Smailagic, and L. E. Wittmers. Continuous Infer-

ence of Psychological Stress from Sensory Measurements Collected in the Natural Environment. *Information Processing in Sensor Networks (IPSN)*, pages 97–108, 2011.

[101] N. Quoc Viet Hung, N. T. Tam, L. N. Tran, and K. Aberer. An evaluation of aggregation techniques in crowdsourcing. In *Web Information Systems Engineering*, pages 1–15. Springer Berlin Heidelberg, 2013.

[102] B. Rammstedt and O. P. John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1):203–212, 2007.

[103] V. Raykar and S. Yu. Ranking annotators for crowdsourced labeling tasks. *Advances in neural information processing systems*, pages 1809–1817, 2011.

[104] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[105] F. Ribeiro, D. Florencio, and V. Nascimento. Crowdsourcing subjective image quality evaluation. *IEEE International Conference on Image Processing*, pages 3097–3100, 2011.

[106] J. M. Robbins and J. I. Krueger. Social Projection to Ingroups and Outgroups: A Review and Meta-Analysis. *Personality & Social Psychology Review (Lawrence Erlbaum Associates)*, 9(1):32–47, 2005.

[107] R. M. Sakia. The Box-Cox Transformation Technique: A Review. *Journal of the Royal Statistical Society*, 41(2):169–178, 1992.

[108] H. Salam and M. Chetouani. A multi-level context-based modeling of engagement in human-robot interaction. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 3, pages 1–6. IEEE, 2015.

[109] K. R. Scherer, K. R. Scherer, and P. Ekman. On the nature and function of emotion: A component process approach. *Approaches to emotion*, 2293:317, 1984.

[110] A. Schulz and C. Vögele. Interoception and stress. *Frontiers in psychology*, 6:993, 2015.

[111] H. Selye. A syndrome produced by diverse nocuous agents. *Nature*, 1936.

[112] H. Selye. *The stress of life*. McGraw-Hill, New York, NY, US, 1956.

[113] H. Selye. Stress without Distress. In *Psychopathology of Human Adaptation*, pages 137–146. Springer US, Boston, MA, 1976.

[114] F. Shaffer, R. McCraty, and C. L. Zerr. A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability. *Frontiers in psychology*, 5, 2014.

[115] N. Sharma, A. Dhall, T. Gedeon, and R. Goecke. Modeling stress using thermal facial patterns: A spatio-temporal approach. *Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 387–392, 2013.

[116] Y. Shi, M. H. Nguyen, P. Blitz, B. French, S. Fisk, F. D. Torre, A. Smailagic, D. P. Siewiorek, M. Absi, E. Ertin, T. Kamarck, and S. Kumar. Personalized Stress Detection from Physiological Measurements. In *Internation Symposium on Quality of Life Technology*, 2010.

[117] A. Smola and V. Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997.

[118] M. Soleymani and M. Larson. Crowdsourcing for Affective Annotation of Video : Development of a Viewer-reported Boredom Corpus. *Proceedings of the ACM SIGIR workshop on crowdsourcing for search evaluation*, pages 4–8, 2010.

[119] M. Soury. *Détection multimodale du stress pour la conception de logiciels de remédiation*. PhD thesis, Université Paris-sud, 2014.

[120] C. D. Spielberger. Manual for the state-trait anxiety inventory stai (form y)(" self-evaluation questionnaire"). 1983.

[121] O. Svenson and A. J. Maule. Time pressure and stress in human judgment and decision making. 1993.

[122] S. Szabo, Y. Tache, and A. Somogyi. The legacy of Hans Selye and the origins of stress research: A retrospective 75 years after his landmark brief "Letter" to the Editor # of Nature. *Stress*, 15(5):472–478, Sept. 2012.

[123] J. Taelman, S. Vandeput, A. Spaepen, and S. Van Huffel. Influence of Mental Stress on Heart Rate and Heart Rate Variability. In *4th European Conference of the International Federation for Medical and Biological Engineering*, pages 1366–1369. 2009.

[124] G. Tartarisco, G. Baldus, D. Corda, R. Raso, A. Arnao, M. Ferro, A. Gaggioli, and G. Pioggia. Personal Health System architecture for stress monitoring and support to clinical decisions. *Computer Communications*, 35(11):1296–1305, 2012.

[125] P. A. Thoits. Stress, coping, and social support processes: where are we? What next? *Journal of health and social behavior*, 35(1995):53–79, Jan. 1995.

[126] M. Traina, A. Cataldo, F. Gallulo, and G. Russo. Effects. of anxiety due to mental stress on heart rate variability in healthy subjects. *Minerva psichiatrica*, 2011.

[127] A. Troisi. Displacement Activities as a Behavioral Measure of Stress in Nonhuman Primates and Human Subjects. *Stress*, 5(1):47–54, July 2009.

[128] A. Troisi. Displacement Activities as a Behavioral Measure of Stress in Nonhuman Primates and Human Subjects. *Stress*, 5(1):47–54, July 2009.

[129] A. Vinciarelli and G. Mohammadi. A Survey of Personality Computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.

[130] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[131] O. Weisman, M. Chetouani, C. Saint-Georges, N. Bourvis, E. Delaherche, O. Zagoory-Sharon, D. Cohen, and R. Feldman. Dynamics of non-verbal vocalizations and hormones during father-infant interaction. *IEEE Transactions on Affective Computing*, 2016. to appear.

[132] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems*, 22(1):1–9, 2009.

[133] J. Wijsman, B. Grundlehner, H. Liu, H. Hermens, and J. Penders. Towards mental stress detection using wearable physiological sensors. In *IEEE Engineering in Medicine and Biology Society*, volume 2011, pages 1798–801, 2011.

[134] K. Wu and K.-H. Yap. Fuzzy SVM for Content-Based Image Retrieval: a pseudo-label support vector machine framework. *IEEE Computational Intelligence Magazine*, pages 10–16, 2006.

[135] Q. Xu, Q. Huang, and Y. Yao. Online crowdsourcing subjective image quality assessment. *ACM International Conference on Multimedia*, pages 359–368, 2012.

[136] P. Zachar and R. D. Ellis. *Categorical versus dimensional models of affect: a seminar on the theories of Panksepp and Russell*, volume 7. John Benjamins Publishing, 2012.

[137] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang. Audio-visual affective expression recognition through multistream fused HMM. *IEEE Transactions on Multimedia*, 10(4):570–577, 2008.

[138] G. Zhou, J. H. L. Hansen, S. Member, and J. F. Kaiser. Nonlinear Feature Based Classification of Speech Under Stress. *IEEE Transactions on speech and audio processing*, 9(3):201–216, 2001.