PhD Dissertation

International Doctorate School in Information and Communication Technologies

DISI - University of Trento

Managing the Scarcity of Monitoring Data through Machine Learning in Healthcare Domain

Alban Maxhuni

Advisors:

Dr. Oscar Mayora

Dr. Venet Osmani

Prof. Imrich Chlamatac

Universitá degli Studi di Trento

January 2017

Abstract

Nowadays, the advances in information and communication technology have brought a revolution in many disciplines, including medicine and public health. Due to these advances, there is an enormous amount of data generated daily from individuals. Extracting knowledge from large amounts of data involves different challenges, such as processing and extracting valuable information from collected data in real-life activities. In the past decades, data collected at the different sources were neglected, due to the lack of efficient machine learning algorithms and many opportunities to improve patients' knowledge of their chronic diseases were missed. One of the advanced technologies in healthcare is wearable sensing that is integrated into various accessories such as wristwatches, headphones, and smartphones. Significant advances in sensor manufacturing and data analysis methods have opened up new possibilities for using wearable technology for continuous vital signs monitoring in order to prevent, treat and control users' diseases. However, despite their potential use of remotely-sensed data, for some healthcare applications often we need to deal with scarce data. Dealing with scarce data is a significant problem, especially in predicting wellbeing of the individuals from data acquired in real-life activities.

In the field of Ubiquitous Computing, a significant problem of building accurate machine learning models is the effort and time consuming process to gather labeled data for the learning algorithm. Moreover, efficient data use demands are constantly growing. These demands for efficient data use are growing constantly. Researchers are therefore exploring the use of machine learning techniques to overcome the problem of data scarcity. In healthcare, classification tasks require a ground truth normally provided by an expert physician, ending up with a small set of labeled data with a larger set of unlabeled data. It is also common to rely on self-reported data through questionnaires, however, this introduce an extra burden to the user who is not always able or willing to fill in. Finally, in some healthcare domains it is important to be able to provide immediate response (feedback), even if the user is not familiarized with the use of an application. In all of these cases the amount of available data may be insufficient to produce reliable models.

This thesis proposes a new approach specifically designed for the challenges in producing better predictive models. We propose using our novel Intermediate Models to predict the mood variables associated with the questionnaire using data acquired from smartphones. Then, we use the predicted mood variables with the rest of the data to predict the class, in our empirical assessment, the state mood of a bipolar disorder patient or stress levels of employees have been used. The motivation behind this new approach is that there are relevant proposed methods such as latent variables used as intermediate information *helping to create better predictive models. These methods are used in literature to complete the missing data using the most common value, the most probable value given the class, or induce a model for predicting missing values using all the information from features and the class. However, these variables are artificially created and used as intermediate information to build better model. In our Intermediate Models, we know in advance how many mood variables to use and we have the information from these variables, which allow us to produce better models.*

To address scarce data, we propose applying a semi-supervised learning setting while taking advantage of the presence of all unlabeled datasets. In addition, we propose using transfer learning methods that is used to improve the learning performance with the aim at avoiding expensive data labeling efforts. To the best of our knowledge, there are few works that have used transfer learning for healthcare applications to address the problem of limited labeled data. The proposed methods have been applied in two different healthcare fields: mental-health and human behaviour field. This thesis addresses two classification problems, a) classification of episodic state of bipolar disorder patients, and b) detecting work-related stress using data acquired from smartphone sensing modalities.

The proposed approaches improve classification performance in terms of accuracy: a) classification of bipolar disorder episodes yielded overall accuracy from ≈73% to ≈90%, and b) the results in predicting work-related stress yielded the accuracy from ≈71.68% and ≈78%. Results obtained overcome previously proposed approaches that use traditional supervised learning techniques. Finally, results shown that the proposed approaches are capable to successfully deal with scarce data.

Keywords: [Intermediate models, Semi-supervised learning, and Transfer-learning.]

Acknowledgements

First, I thank my advisors's Dr. Oscar Mayora and Dr. Venet Osmani. They have influenced my view of the research process, and instilled in me the importance of aiming to produce quality research with the potential for impact. Most of all, I value their honest opinions, critics, their calmness and clarity of advice during difficult times, and their patience and understanding over the past several years. I am indebted to have had advisor's that gave me all of the resources, guidance and support I could ever need during the period that lead up to this dissertation.

I feel fortunate to have had the opportunity to work closely with Dr. Angélica Muñoz-Meléndez, Dr. Eduardo F. Morales, Prof. Enrique L. Sucar, and Dr. Pablo Hernandes-Leal (who became a great friend) during my internship that included time at Instituto Nacional de Astrofísica, Óptica y Electrónica. - INAOE, Puebla, Mexico. Their valuable comments are of great help in my research work. I thoroughly enjoyed the chance to work with many other researchers and interns at INAOE.

During the time of being Ph.D. candidate I was researcher in the Ubiquitous Technologies for Health - UbiHealth group at CREATE-NET, Trento, Italy. I was fortunate to be part of the EU MONARCA, Turnout BurnOut, Virtual SocialGym and UbiHealth projects and to have collaborated with Dr. Agnes Grúnerbl and Prof. Paul Lukowicz (DFKI-Kaiserslautern), EIT ICT Labs in Trento and University of Trento. I would also like to thank many of my former colleagues there for all the activities which meant a welcoming distraction from the hard and stressful work of a scientist, such as, hiking or just enjoying a nice espresso coffee from the first floor vending machine.

Finally, I am deeply indebted to my dear parents for their love and encouragement all the years. Now it's almost done, this means no more jokes about it anymore! I'm infinitely grateful for the values they have passed down to me, and for their continuous support throughout all my studies. Therefore, I would like to dedicate my thesis to my parents.

Thanks you all!

Trento, January 2017
Alban Maxhuni

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

> *"Arguably the greatest technological triumph of the century has been the public-health system, which is sophisticated preventive and investigative medicine organized around mostly low and medium-tech equipment; ... fully half of us are alive today because of the improvements."*
>
> — **Richard Rhodes**

This chapter opens with a brief overview of machine learning approaches and its role in healthcare. We explain several motivations for addressing the issues of scarce data and learning the model with a small amount of labeled data. We discuss the most common methods used in this thesis and offer intuitive explanations and key insights why the proposed methods work under scarcity settings. The motivation, importance of the topic, problem statement, methodology, as well as the research contributions of these studies are then provided. Finally, the structure of the dissertation is discussed.

1.1 Machine learning and the advances in healthcare applications

Prevalence of chronic diseases is increasing all over the world and management of them represents one of the greatest healthcare challenges. Due to this reason, healthcare systems are struggling to find better solutions that improve quality, efficiency and reduce care costs. In addition, using wearable technologies for supporting individuals with chronic conditions has been associated with significant improvements in quality of life (Park and Jayaraman, 2003). Mobile computing and sensing technologies have shown potential to improve healthcare quality, and efficiency. Compared with standard clinical practice for monitoring patients, measurements rely on observation data collected in laboratory settings or in person (Pentland, 2004). For example, in (Vancampfort et al., 2013a) authors argue that increased knowledge about motor activity and repetitive movements of bipolar disorder patients during the manic episodes offers deeper insight towards new therapies. As a result, accurate and continuous monitoring has become increasingly important in healthcare, where employing technology for patient monitoring can help assess the impact of mental illness on patients daily activities (Hansen and Christensen, 2011), to increase effectiveness in treating mental disorders (Vancampfort et al., 2013b), and to monitor perceived stress at working environments (Quer et al., 2013). Due to its embedded sensors, smartphones have become capable of monitoring multiple dimensions of human behavior, including physical, mental, and social interaction dimensions (Grunerbl et al., 2015; Osmani et al., 2013a).

In general, healthcare institutions are becoming more and more dependent on advances in technology, and the use of Machine Learning (ML) techniques can provide useful support to assist physicians in many ways. In the last decade, ML received attention from many domains, including healthcare with an aim of improving service quality and care. To-date, advanced ML techniques that are used in healthcare aimed at solving prognostic problems, including mental-health (Grunerbl et al., 2014; Osmani et al., 2013a) and human behaviour fields (Ertin et al., 2011; Muaremi et al., 2013). It is often argued that using ML tools in medicine will lead to improvements in patients care. ML has already demonstrated global impact in clinical medicine due to the latest advances in sensor technology that can use data to assess a patient's wellbeing in real-time settings (Matthews et al., 2014). It offers the opportunities to enhance physicians work, including efficiency and quality of healthcare (Clifton et al., 2015).

Current technologies are generating and collecting large volume of data, that makes them too complex to be analysed using traditional methods, such as supervised-learning methods. There are various demands faced in building such systems, however, of particular interest is that of building robust learning systems that function in real-life environments, where data are acquired continuously by multiple sensing modalities. However, acquiring patient data poses several difficulties, due to incompleteness (missing label values), noise in the dataset, irrelevant features selection, and scarceness of data as result of low number of patient records available. This research addresses the problems of data scarcity and incompleteness using ML methods that are able to handle datasets aiming at improving the classification performance.

This chapter lays the groundwork for this thesis by providing an overview of the challenges posed for acquiring qualitative data from users real-life activities with respect to learning a classification models and other known learning strategies used for tackling the scarcity and incomplete data. This thesis tackles some of the facing challenges by building efficient and accurate classification system. Those challenges mainly arise from the difficulty of acquiring large-scale labeled training data, since they are very costly in terms of human time and effort. The availability of training data is also a common problem in machine learning, as they are a critical resource to build classifiers. As result, the scarcity of training data is also the most common problem in monitoring human behaviour which leads to a poor classification accuracy using the standard supervised-learning techniques (Brodley and Friedl, 1999).

1.2 Research challenges

In the field of machine learning, classification is a common problem in most applications. Machine learning approaches have been widely applied in many domains. Supervised learning is the most popular category of machine learning algorithms. However, the performance of supervised algorithms strongly depend on sufficient labeled training data to build an accurate model and make prediction on the future data (Yang and Wu, 2006). Nonetheless, in real-life settings, labeled training data can be obtained with expensive cost which has been a major bottleneck of making effective predictive models that can be applicable in practice. For example, the monitoring systems applied in healthcare settings have a common problem, in particular when the size of available training set is small, supervised methods may fail to correctly classify individuals behaviour (Longstaff et al., 2010).

On the contrary, unlabeled training can be easily achieved which can positively impact classification performance. However, selecting unlabeled data for labeling to achieve high classification performance with minimal labeling efforts is a challenging problem. It is generally assumed that labeled data is available, however, in practice, there is often associated with high costs for obtaining this labeled data. As a result, it is often the case that only a small amount of labeled data is available. Thus, it may be possible that a large amount of unlabeled data is available, which is usually ignored, that can be exploited with the appropriate algorithms like the ones proposed in this thesis.

For this thesis, we identify key challenges and core issues surrounding the detection of behavioral changes from smartphones sensing, which collects data from the different sensing modalities. The main challenge is to address issues of having scarce information. Further, we identify also challenges for extracting information from datasets and results interpretation. Finally, we explore the challenges of reducing annotation cost and improve prediction accuracy.

1.3 Research objectives

In this research work main research questions are:

RQ.1 *How can we manage scarce data sets from long-term monitoring studies using machine learning techniques in the healthcare domain? To what extent can we improve the knowledge about individuals wellbeing using scarce data?*

RQ.2 *Is it possible to build a system that reduces users' burden in terms of providing annotated data in monitoring systems?*

1.4 Overview of the proposed approaches

Despite significant advances made towards a better understanding of human behaviour through smartphone sensing capabilities, there are still many open challenges associated with monitoring individuals in real-life setting. In this thesis, we focus on analysing data from participants monitored in real-life settings collecting information from smartphone sensor and self-assessment questionnaires. The problems addressed in this thesis are divided into two classes of problems, classification of episodic state of bipolar disorder patients and detection of work-related stress. In both domains, we deal with large number of unlabeled data. At first view it might seem that nothing is to be gained from unlabeled data, however, we demonstrate how unlabeled data can be used to address issues of having scarce information.

In order to overcome aforementioned obstacles, in this research we propose using approaches that aim at improving classifier accuracy using unlabeled data and investigate algorithms that require only minimal feedback from users. We start by introducing *Semi-supervised learning* (SSL), which concerns the issue on how to improve classification performance and to reduce the need of expensive labeled data via unlabeled data (Zhu, 2006). Data acquired in our studies contains missing data and a large number of instances available are unlabeled. Thus, the key challenge of this thesis is to exploit the amount of unlabeled data to enhance the overall performance of the proposed models.

In this thesis, we propose three methods based on: semi-supervised learning, transfer learning and the novel use of intermediate models. However, it is important to notice that transfer learning has not been included in mental-health areas, because we are dealing with only five patients with relatively different symptoms.

First, we propose a SSL approach to increase the accuracy of classification models to address missing labeled instances. The proposed semi-supervised learning demonstrated their efficacy in situations where participants have a small amount of labeled data available, and then used a trained model to predict the rest of unlabeled data. We assume that only a small amount of the dataset collected from participants are labeled and we try to use the unlabeled instances to learn about data structure. In this thesis we apply a well-known semi-supervised learning algorithm to improve the classification performance of models induced for two health-care applications. However, despite the capability to handle scarce information, there are still several limitations using semi-supervised learning approaches. For example, data acquired from few participants in our studies contain very few amounts of labeled or incomplete number of classes. Thus, semi-supervised learning techniques may fail in these settings, if existing classes are not covered in the training phase.

For this reason, the proposed approach is based on *Transfer Learning* (TL) (Pan and Yang, 2010) and consists in using information from other known models to tackle the problem of users having large amount or complete unlabeled data. In this thesis we use TL approaches to build suitable models even with scarce data. We propose several approaches for transferring data from other users under different conditions and by combining different users models. We transfer information from previously built models (*i.e.*, subjects with sufficient labeled instances) to the target model which contains insufficient data to produce an accurate one. Using this approach assumes to have a set of previously learned models along with their respective data (used to learn the model). Furthermore, we investigate *Ensemble Learning* (EL) (Turner and Oza, 1999) to improve classification performance by combining multiple learning algorithms. The goal of ensemble learning is the same as in transfer learning.

The above-mentioned approaches attempt to reduce the annotation cost by resorting to semi-supervised learning and transfer learning. In this thesis, we aim at studying the information collected by users through questionnaires as useful information. However, it is a tedious task for each user. In this research, we propose to predict mood variables associated with questionnaires using data from smartphones to alleviate the user from this burden. Then, the predicted mood variables are used with the rest of data from smartphones for class prediction. We call the models that predict mood variables from the questionnaires: *intermediate models* as they are used as input for the final predictive model. In terms of machine learning techniques, although we can relate this technique with other existing methods, we are not aware of any research that uses the same approach. For instance, some techniques use *latent variables* (Muthén and Muthén, 2007), which can be exploited to create better predictive models. However, latent variables are artificially created and used as intermediate information to build better models.

To the best of our knowledge, this is the first approach to demonstrate the ability for increasing classification accuracy by using intermediate models in healthcare domain. The proposed methods have shown to overcome several obstacles surrounding smartphone classification which can be used to reduce the burden during long-term users monitoring in terms of providing labeled training data.

This thesis findings open new research perspectives to improve future monitoring systems based on 'autonomous' data classification reducing costs from human annotations or self-assessment.

1.5 Contribution of this research

In this thesis, data scarcity problem is tackled considering real-life applications. The usefulness of proposed methods are evaluated in two health related problems: mental-health and stress detection. The proposed approaches may contribute for future monitoring systems to infer human behaviour with small amount of labeled data and to reduce the bottleneck in self-monitoring systems with high burden placed on users. To the best of our knowledge, no research studies to date in bipolar disorder and work-related stress have reported the challenges of having all annotations or labeled classes or other issues for including unlabeled dataset within classification processes and yet to improve their performance.

This thesis improves the state of the art in several directions:

- We propose a novel method, namely *intermediate models* to predict psychological and wellbeing conditions using measures derived from self-assessment questionnaires of individuals wellbeing as they are used as input for the final predictive model. Thus, enhancing the accuracy in classification of episodic state of patients with bipolar disorder, as well as the perceived stress of employees.
- We propose a semi-supervised learning method to address the challenge of smartphone monitoring in healthcare domains, in order to improve a supervised learning algorithm with unlabeled data.
- We propose a transfer learning algorithm based on models comparison to select the closest subject for knowledge transfer. The aim of using this approach is to identify a similar subject and to improve the model of a new user with scarce data, thus, improving overall classification accuracy through transfer learning.

Through the investigation carried out in this thesis, foundations are laid for managing data scarcity in monitoring data while opening new research guidelines for future methods development.

1.6 Structure of the research

This thesis is organized as follows, see Figure 1.1:
Chapter 2. Background: It presents relevant concepts that are necessary for understanding our research. It describes the stages of data processing chain, describing the challenges and sensing capabilities in both fields, mental-health and human behaviour. It also introduces the steps in pre-processing, such as feature extraction, feature selection and classification methods in detail. Finally, it also faces the challenges when dealing with scarce data and it aims at improving classification performance using the *Semi-supervised learning and Transfer Learning methods.*

Chapter 3. Related Work: It provides an overview of human behaviour monitoring research in the field of mental-health and work-related stress in general. We address the most important topics related to smartphone sensing systems, including: type of monitored behaviour, type of sensing modalities required in mental-health and other application areas in healthcare. Finally, we provide machine learning methodologies used to date for classification of patients health state in bipolar disorder and individuals stress at working places.

Chapter 4. Description Of The Monitoring Systems: It introduces formally the key challenges addressed in this thesis for classification of state of bipolar disorder patients and classification of employees perceived stress at working environment. It also shows the problems of existing solutions and their practical limitation of monitoring systems in mental-health and field of behaviour, concerning for instance, classification of scarce data and unlabeled data from annotators (*i.e.*, patients, employees).

Chapter 5. Classification Of Episodic States Of Bipolar Disorder Patients Using Smartphones:
Provides the methodology used in our research to detect behaviour patterns in bipolar disorder patients. First section shows the association of physical activity level with psychiatric evaluations and estimates differences of physical activity level (morning, afternoon, night context) into different stages of patients episodic state. Second section, focuses on classification of bipolar disorder episodes based on analysis of voice and motor activity of patients during phone conversations. This research evaluates the performance of several classifiers, different sets of features and the role of questionnaires for classifying bipolar disorder episodes.

The research in physical activity monitoring is first described in the conference paper (Osmani et al., 2013a):

◆ Osmani, V., Maxhuni, A., Grunerbl, A., Lukowicz, P., Haring, C., and Mayora, O. Monitoring activity of patients with bipolar disorder using smart phones. In

ACM Proceedings of International Conference on Advances in Mobile Computing and Multimedia (MoMM2013), Vienna, Austria, December 2013.

The research work on classification of bipolar disorder episodes based on analysis of voice and motor activity is described in the journal (Maxhuni et al., 2016a):

- Maxhuni, A., Munoz-Melendez, A., Osmani, V., Perez, H., Mayora, O., and Morales, E. F. (2016). Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. Pervasive and Mobile Computing.

Chapter 6. Monitoring Stress@Work Using Smartphones: The contents of this chapter consist of three parts. First section propose a new approach based only on smartphone data to predict subjects' daily stress. Comprehensive analysis of the association between objectively measured data (*e.g.*, physical activity, location, social-interaction and social-activity) with subjective self-assessment of work-related stress based on demographic information. In the second section we propose an approach based on transfer learning for building a subject model with scarce data. Finally, in the third section, we focus on classification of employees stress based on analysis of motor activity during phone conversations.

The research in smartphone assessment of stress and modeling stress had as a result the following papers:

- Maxhuni, A., Hernandez-Leal, P., Osmani, V., Sucar, E., Mayora, O., and Morales, E. Stress assessment using Smartphones. Transactions on Intelligent Systems and Technology, Submission January 2016 (in review).
- Maxhuni, A., Hernandez-Leal, P., Sucar, E., Osmani, V., Morales, E., and Mayora, O. Stress Modeling and Prediction in Presence of Scarce Data. Journal of Biomedical Informatics , January 2016.
- De Santa, A., Gabrielli, S., Mayora, O., and Maxhuni, A., (2015). Strumenti Innovativi per la Misura dello Stress Correlato al Lavoro. Medico competente Journal, 2015, Journal Article.

The work in Transfer Learning and Intermediate Models in Stress Prediction are introduced in conference papers (Hernandez-Leal et al., 2015)

- Hernandez-Leal, P., Maxhuni, A., Sucar, L. E., Osmani, V., Morales, E. F., and Mayora, O. (2015). Stress Modeling Using Transfer Learning in Presence of Scarce Data. In Ambient Intelligence for Health (pp. 224-236). Springer International Publishing.

- Maxhuni, A., Hernandez-Leal, P., Morales, E. F., Sucar, L. E., Osmani, V., Munoz-Melendez, A., and Mayora, O. (2016). Using Intermediate Models and Knowledge Learning to Improve Stress Prediction. In AFI360 Conference Track on Future Internet e-Health.

Relevant research in monitoring wellbeing at work are described in the conference papers Maxhuni et al. (2011), Matic et al. (2012), Matic et al. (2013), and Garcia-Ceja et al. (2014) :

- Maxhuni, A., Matic, A., Osmani, V., and Ibarra, O. M. (2011, May). Correlation between self-reported mood states and objectively measured social interactions at work: A pilot study. In Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011 5th International Conference on (pp. 308-311). IEEE.
- Matic, A., Osmani, V., Maxhuni, A., and Mayora, O. (2012, May). Multi-modal mobile sensing of social interactions. In Pervasive computing technologies for healthcare (PervasiveHealth), 2012 6th international conference on (pp. 105-114). IEEE.
- Matic, A., Maxhuni, A., Osmani, V., and Mayora, O. (2013, September). Virtual uniforms: using sound frequencies for grouping individuals. In Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication (pp. 159-162). ACM.
- Garcia-Ceja, E., Osmani, V., Maxhuni, A., and Mayora, O. (2014). Detecting walking in synchrony through smartphone accelerometer and wi-fi traces. In Ambient Intelligence (pp. 33-46). Springer International Publishing.

Altogether, these chapters provide the research work in investigating development and evaluation of robust methods in everyday life scenarios in both mental-health and human behaviour fields, with a focus on the behaviour patterns recognition and intensity of motor activity estimation. They introduce the design and the outcome of a number of experiments that were conducted. These experiments confirm that the semi-supervised learning methods outperforms existing traditional-methods due to missing labeled instances. We proposed four different transfer learning approaches to cope with scarce data. Ensemble weighted approach obtained the best scores increasing accuracy almost by 10% in average. Finally, the Intermediate Models approach proposed have demonstrated to improve the accuracy performance in stress prediction and classification of bipolar disorder episodes.

Chapter 7. Conclusion: Summarizes the thesis, draws conclusions and gives ideas for possible future extension of the presented research work.

Appendix A Supplementary Material.

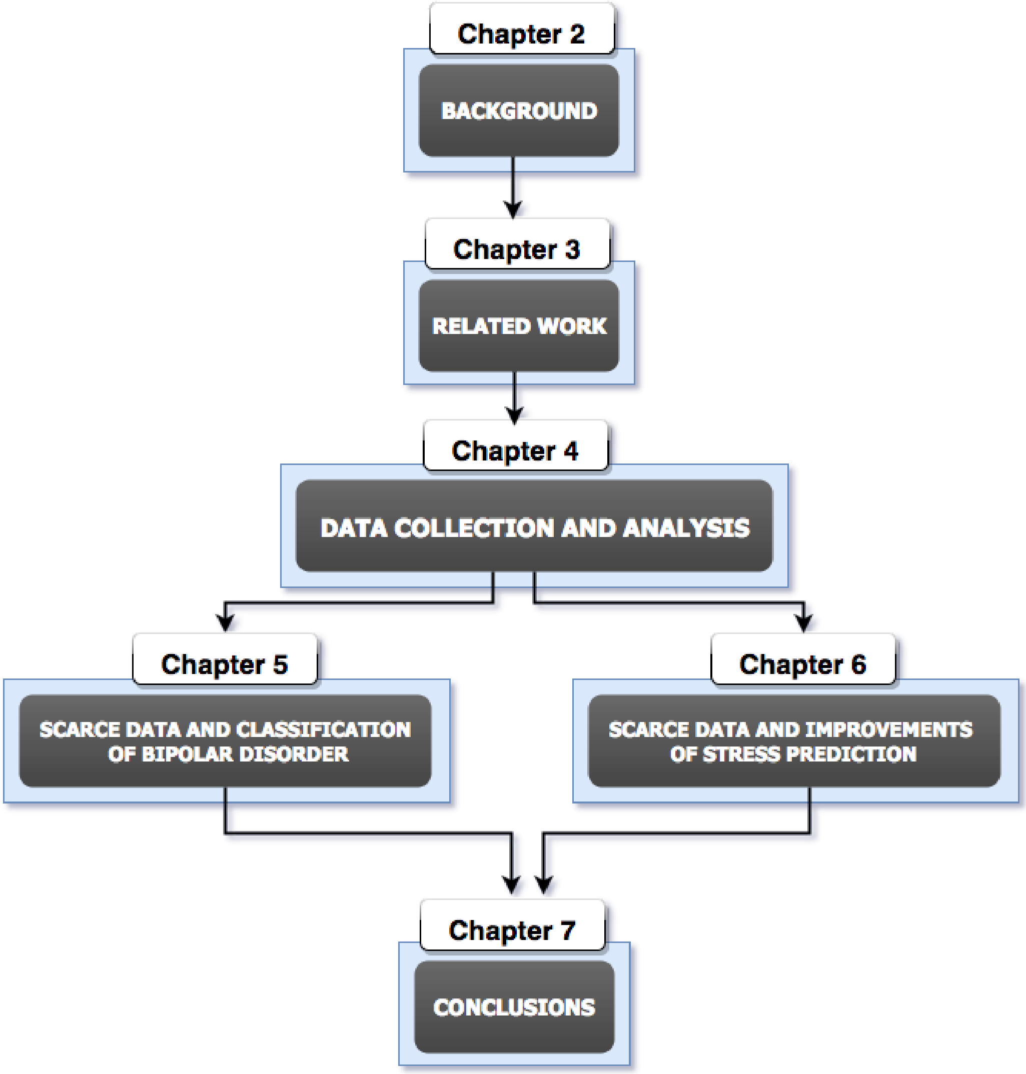

Chapter 2
BACKGROUND
Chapter 3
RELATED WORK
Chapter 4
DATA COLLECTION AND ANALYSIS
Chapter 5
SCARCE DATA AND CLASSIFICATION OF BIPOLAR DISORDER
Chapter 6
SCARCE DATA AND IMPROVEMENTS OF STRESS PREDICTION
Chapter 7
CONCLUSIONS

Figure 1.1: The organisation of thesis.

Chapter 2

BACKGROUND

> *"When you apply computer science and machine learning to areas that haven't had any innovation in 50 years, you can make rapid advances that seem really incredible."*
>
> – **Bill Maris**

This chapter presents an overview of the methods used to extract and interpret data. We begin by providing a review of pervasive health computing and data that can be acquired from the sensing modalities to infer human behaviour. We present our main focus of this thesis: models, concepts and algorithms in machine learning that are relevant for this research work. Finally, in this chapter we highlight the novel intermediate models proposed in this research.

2.1 Pervasive health computing

By far the most dominant concepts in research literature in healthcare context includes pervasive computing, ubiquitous computing, and ambient intelligence (AI) that evolve toward the development and deployment of pervasive health application (Borriello et al., 2007). The main goal of pervasive health is to support patients, and clinicians through the use of mobile, pervasive, and ubiquitous computing technologies (Mihailidis and Bardram, 2006).

In 1991, Mark Weiser (Xerox PARC) envisioned the future of smart-sensing environment as:

"...the most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it" (Weiser, 1991).

His visions toward advances of technology have predicted that computing units will become so ubiquitous like an invisible servant without noticing their presence within environments. Current trends are rapidly moving toward Weiser's vision. Further, authors such as (Mihailidis and Bardram, 2006) argue that nowadays healthcare models need to be transformed into a more distributed and highly responsive healthcare processing model, where patients can take control of their own health to manage their wellness, preventive care and proactive intervention. Despite their complexity, these concepts have demonstrated that integration of interventions into hospitals and home-care are moving toward patient-centered healthcare delivery system.

Additionally, the information used for personal healthcare today largely derives from self-report questionnaires and infrequent short visits to the cabinet of physicians or clinicians. These methods are often prone to certain biases in self-rating of individual state, such as recall and social desirability bias (Mortel et al., 2008). Therefore, research studies in healthcare advocate moving toward patient-centered healthcare (Mihailidis and Bardram, 2006), that would improve current healthcare services by understanding individuals' context through non-intrusive and wearable sensing devices.

Nowadays, smartphones are viewed as an essential part of life and considered as a personal accessory (Ventä et al., 2008). Taking all these advantages into account, including their functionality in ubiquitous computing and communication, smartphones are considered as an important accessory in users social behaviour (Srivastava, 2005). As such, smartphones are becoming widely accepted among all population ages and suit their needs and lifestyles. These features have turned attention also to healthcare systems, allowing them access to pervasive healthcare applications (Korhonen, 2004).

In the field of pervasive computing, research work initiatives are looking for better alternatives for continuous and regular measurements of participants through behaviour and lifestyle (*e.g.*, bad habits, sedentary behaviour, inactivity). This information may contribute to prevention of chronic diseases and to reduce the risk of premature death. The ongoing progress in technology has led healthcare institutions to increase efficiency and to improve service quality throughout the use of pervasive healthcare applications while providing support to hospitals and promoting preventive healthcare, where data processing is integrated into everyday objects and activities (Korhonen, 2004).

Sensor-enabled in wearable devices (*e.g.*, smartphones) have the potential for collecting real-day activities from continuous sensor data that have a huge impact in changing the way health and wellbeing are assessed, and how care and treatment are delivered. For example, the Ubitfit Garden project (Consolvo et al., 2008), demonstrate the importance of capturing levels of physical activities and relate this information to personal health goals when presenting feedback to the users. These types of systems have been suggested to decrease the risk of sedentary behaviour through their physical exercises.

These systems have been gaining attention also in mental-healthcare services. Marcu et al. (2011) suggest that using sensing capabilities of smartphones in pervasive healthcare would help both, patients and physicians, to control diseases by providing continuous feedback based on objective measurements. The authors emphasize the importance of using smartphone sensing technologies that could also serve to manage mental illnesses by monitoring behaviour patterns, daily activities and self-reported mood. Data measured from these systems would allow clinicians to help patients reacting accordingly and to prevent moving toward extreme severe states of the disease.

Additionally, smartphone are also capable of monitoring physiological reaction (*i.e*, expressing various emotions during phone conversations) that can derive from speech (Scherer, 1986). Several quantitative studies demonstrated the importance in assessing individuals emotional state by observing various forms of non-verbal communication, such as non-verbal elements of speech or body postures (Hansen and Christensen, 2011).

2.2 Monitoring in healthcare

According to recent reports presented from World Population Prospect of the United Nations (Nations, 2013) the average age of the population is expected to grow rapidly in developed countries within the next decades. This increase will automatically raise the cost of healthcare and result in significant effects of a government's budget. However, the latest advances in different fields of technology has enabled the healthcare institutions to decrease the problem with integration help of these technologies to accelerate and to improve services in healthcare.

Healthcare is an essential part of humans in everyday life and periodic monitoring of vital parameters, such as treatments are the basic function of healthcare. These processes become even more crucial when it comes to the treatment of mental disorders which require trained medical personnel to monitor their state. Current practice for monitoring these basic health parameters in healthcare are measured from physicians or medical personnel only at discrete intervals. In addition, self-reported subjective health measurements are essential to assess the effectiveness of treatments (*i.e.*, mental-health). These approaches often leads to the loss of crucial information about individuals treatments during entire day or night periods.

Therefore, a particular interest is focused on continuous monitoring techniques capable of monitoring long-term information about individuals to understand significant changes of their health conditions in real-time. These systems could be used to assess certain health parameters during a long period and to provide a complete information about patients health and perhaps find new unknown markers of specific diseases.

In the past two decades, physicians in healthcare have used new advanced technology to improve healthcare services and at the same time to increase the quality of care and health outcomes. In particular, there has been a focus on using computers, smartphones, and other wearable sensors as means to support medical and healthcare education and provide more cost-effective care.

2.3 Advantages of using smartphone computing to monitor human behaviour

In 2014 approximately 1.8 billion people worldwide owned smartphones. This number is increasingly growing and in just a few years smartphone users are expected to exceed one third of the world's population (Portio, 2011). Due to their mobility and power afforded with the embedded sensors in these devices, smartphones have been identified as an important component that opens up smartphones to new advances across a wide spectrum of applications in healthcare domain.

A number of research works have demonstrated the potential of monitoring human behaviour using mobile computing and sensing technologies. In order to infer relation dynamics of people and behaviour changes in real-life activities, smartphones have been suggested as a promising candidate (Raento et al., 2005). Research using smartphones for long-term monitoring (Maurer et al., 2006; Raento et al., 2005) have reported several advantages of using smartphone sensing to collect many types of contextual data continuously, such as locations, physical activities, body postures, emotion from speech and social networking.

They are often reported as deeply personal devices, regarding them as personal accessory (Ventä et al., 2008). Alongside these technological advances, there has been also increasing interest from researchers and clinicians in harnessing smartphones as a means of delivering behavioural interventions for health. In the last decades there has been an increasingly wide range of research on using smartphone applications and various features to support general physical and mental wellbeing. They have shown the effect of using smartphones in non-clinical settings for supporting regular physical activity and behavioural health, which is of critical importance for reducing the risks of several chronic diseases. Findings have reported that using smartphones for managing physical and mental-health in supporting changes in health-related behaviour have been widely accepted from individuals participating to those studies (Dantzig et al., 2013; Lane et al., 2011; Mukhtar and Belaid, 2013).

Regarding behaviour monitoring, several research initiatives have shown that human activity recognition involving the use of smartphone technology provides a great potential for personal health systems by monitoring daily activities, wellness, and health status of individuals. As discussed above, smartphones are easily accepted by the mass for adoption of personal activity recognition's systems on wearable platforms due to their integrated rich set of sensors together with their ubiquity. Besides their rich set of new sensors and their ubiquity, using smartphones have been presented in Table 2.1 as unobtrusive device, less costly in installing, and ease of use.

Table 2.1: Summary of methods for monitoring individual in healthcare.

Methods for healthcare performance monitoring	**Accuracy**	**Intrusiveness**	**Privacy**
Medical Personal	High	Medium	High
EEG - Brain Sensing	High	High	Medium-High
Image-based Sensing	High	Medium	High
Audio-based Sensing	High	Medium	High
Physiological Sensing	Low-Medium	Low-High	Low-High
Smartphone Sensing	Low-Medium	**Low**	Low-High

2.4 Instruments

A smartphone has often been suggested as a computing platform which functionality and performance has always been introduced with embedding new sensors. For instance, embedding accelerometer sensors in smartphone has been introduced to enhance the user interface of the smartphone in order to determine the orientation of display, while the user is holding or interacting with the phone.

Sensors embedded on smartphones include a gyroscope, magnetometer, barometer, accelerometer, proximity sensor, ambient light sensor as well as other more conventional devices that can be used to sense including front and back facing cameras, a microphone, GPS and WiFi, Bluetooth radios, Near field communication (NFC) and recently embedded sensors, such as SpO2 sensor for blood oxygen saturation levels, Ultra-Violete (UV) radiation, and Heart Rate Monitor (HRV) sensor. All above mentioned sensors have been used or combined to improve smartphones functionality, such as support to user's interface (*e.g.*, the accelerometer), augment location base services (*e.g.*, magnetometer and GPS) or measuring health-related aspects, such as heart-rate variability using HRM sensor.

Moreover, accelerometer data is capable of characterizing physical movements of user's while carrying the phone (Constandache et al., 2010). Analysing and measurement of accelerometer data derived from smartphone have been exploited to recognize different activities when the smartphone is carried (*e.g.*, running, walking, standing). Fusion of accelerometer data measurement with location, distances, and speed estimated from the

GPS can be used to recognize the mode of transportation of a user, such as using a bus, bike, or car (Mun et al., 2009).

The most powerful and ubiquitous sensors in smartphones are camera and microphone. With the audio recording from the smartphone's microphone it is possible to classify a diverse set of distinctive sounds associated with a particular context or activity in a person's life, such as social interaction, listening to music, or driving (Lu et al., 2009). In addition, camera embedded in smartphones have not only been used for traditional task of capturing images, but also tracking user's eye and movement across phone's display which help understand users' interaction by activating applications using cameras that are embedded in the front of the phone (Miluzzo et al., 2010). Finally, this trend of advances and the incorporation of new sensors will also improve healthcare services.

2.5 Continuous sensing

When it comes to smartphone sensing, applications are particularly developed and designed for a single individual (namely human-centered sensing) and their main focus is on how data are collected, analysed and represented. However, utility of these systems for inferring users behaviour, requires their active involvement in the sensing system (Lane et al., 2008).

Despite advances described in previous section, such as computation, memory, storage, sensing and communications capabilities, smartphones resources are limited if complex signal processing and inference are required (Postolache et al., 2007). For instance, signal processing and machine learning algorithms involves a large volume of sensed data (*e.g.*, classification of audio data (Lu et al., 2009)) and different sensing applications place further distinct requirements in the execution of these algorithms.

Continuous sensing applications that require real-time interferences or frequent sampling rate from energy expensive sensor (*e.g.*, GPS), are vulnerable to quickly drain smartphone's battery and shorten usability time and therefore sensing capabilities. Furthermore, application used in healthcare have specific demands when it comes to continuous sensing of the user since they require actual time classification in response to the incoming data stream measurements (Lu et al., 2010). Thus, for continuous sensing to be feasible there needs to be new breakthroughs or a boost in low energy algorithms that organize duty cycles of smartphone devices while maintaining necessary applications running (Lane et al., 2008)

In this line, earlier research work that used smartphone continuous-sensing systems tended to trade off performance accuracy and decrease battery usage by implementing algorithms that require less computation of sensor data. Further strategies that are often used to synchronize and transfer different sensor data collected to the cloud infrastructures (Cuervo et al., 2010) which are responsible for further processing and analysis steps as well as improving the battery life. In addition, similar techniques have been using duty-cycling methods in order to synchronize sensing using the user context (*i.e.*, during night time hours) which tend to trade off the battery consumption against the sensing performance and latency (Wang et al., 2009). These challenges are actively being studied and are currently hot-topic in the field of continuous sensing in pervasive health.

We believe that the future sensing applications will be successful if they adapt to users context in a smarter way, which will decrease the energy costs and offer sufficient accuracy.

2.6 Methods and techniques

In the next sections, we will present briefly methods and data processing techniques to describe data that are collected from smartphone sensors and infer behaviour changes of individuals in real-life activities. These methods can be described as a chain of processing steps, which starts from raw sensory data and resulting in a prediction of the users well-being while inferring their behaviour changes. In Figure 2.1, we have demonstrated the simplest example of data flow from left to right presenting a human behaviour recognition system.

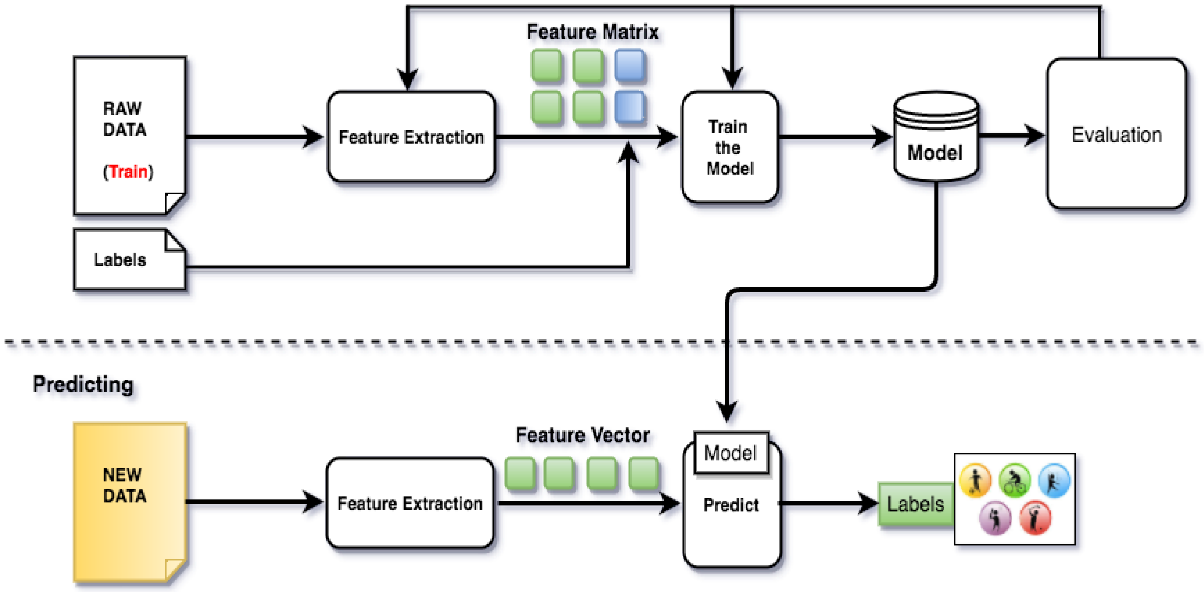


Figure 2.1: Supervised Learning: data collection, feature extraction and prediction for classifying human context and activities.

In the following subsections, we describe each stage of the process starting with: interpreting raw sensor data and feature extraction collected from smartphone sensors that are relevant to our studies. Further, we discuss about classification methods (*e.g.*, standard supervised learning methods, semi-supervised learning methods, and transfer learning usage in healthcare applications.

2.6.1 Interpreting raw sensor data

As presented in Figure 2.1, the first stage to feed a recognition system is data collection. At this stage data measured requires being converted to numerical form before the behaviour pattern vector is set for further training stages. Moreover, raw sensor data that are collected or captured from embedded sensors (*e.g.*, accelerometer, gyroscope, magnetometer) are meaningless without their interpretation, such as interpreting human behaviour or other related aspects. As such, a variety of data mining techniques and statistical measurement tools are often used to interpret information from the data collected by smartphones, *e.g.*, total activity level, daily steps, the total distance run by a user and also cluster with their group of friends of individuals performing in nearby area (*e.g.*, Runtastic (2015)).

In following subsections, we discuss the challenges of interpreting sensor data with more focus on human behaviour monitoring and context modeling.

2.6.2 Pre-processing

Noise in the raw-data collected from sensor modalities is a common problem. Thus, in a field of machine learning, mining raw data begins with pre-processing methods that aim at improving the efficiency of the mining process. This process is one of the most critical step in data mining process since it deals with transformation of the raw dataset.

Data collected from non-invasive sensors (*e.g.*, accelerometers) embedded on smartphones have low signal noise due to environment noise and other artifacts. The term noise in the field of human pattern recognition is used in the broad sense, for instance, all the properties that limits the performance of a recognition system tasks is regarded as noise. The goal of pre-processing is to improve signal noise with the use of filtering methods, normalisation, and other artifact removal. In order to enhance representation of behavioural patterns, filtering methods are techniques that reduce the amount of noise from the data collected.

- ◆ **Segmentation:**
  In machine learning, preprocessed input data are often suggested being split to provide useful entities for classification, namely segmentation. In Figure 2.2 we illustrate segmentation processes, a) each segment add the information regarding

the segment (*i.e.*, window size), and b) the process is interwoven with the previous or following processes (*i.e.*, window overlap). Some recognition systems require segmentation of each individual pattern, *i.e*, segmenting dataset into hours or days to create meaningful entities for the feature extraction and for final classification step. For instance, in activity recognition segmentation is used for partitioning data into fixed-sized windows and at fixed temporal intervals to describe better the activities.

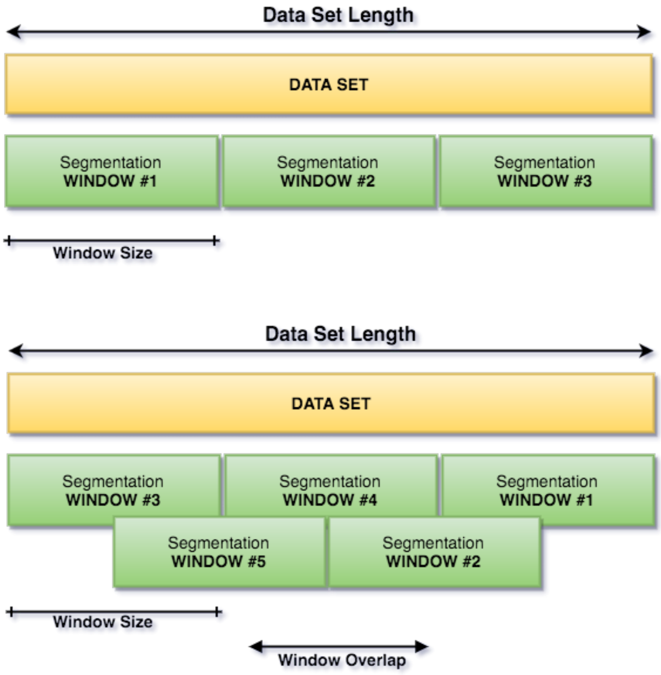


Figure 2.2: Feature segmentation process.

- **Normalisation:**
  In order to prevent singular features from dominating other and to obtain comparable values ranges, feature normalisation is performed to decrease variation within ranges. Normalisation is used to scale features of data in order to fall within a specific range. The goal of these process is to make it easier for the learning algorithm to learn and to make comparison more straightforward across dataset. An instance to normalisation method is the Neural-Network Models (Cochocki and Unbehauen, 1993; Duda et al., 2012), which require dataset normalisation to be within the range of -1 to 1 or 0 to 1.

In order to increase effectiveness of human behaviour recognition and reduce the computation, feature selection is performed. In cases where dimensionality of features set is too high, feature selection detects and discards features that are irrelevant and useless information to train the classifier. In addition, different methods are used to perform artifacts removal. Principal Components Analysis (PCA) (Johnson, Wichern, et al., 1992) and Independent Components Analysis (ICA) (Lee, 1998) are often used to separate artifacts from dataset using higher order statistics of data. All operations used in the pre-processing step, contribute to define a compact representation of behaviour patterns and improve the classification performance (Duda et al., 2012).

2.6.3 Feature extraction

The meaning of the feature extracting step is defined in Devijver and Kittler (1982):

> *"Feature extraction problem ....is that of extracting from the raw data the information which is most relevant for classification purposes, in the sense of minimizing the within-class pattern variability while enhancing the between-class pattern variability."*

In the area of human behaviour recognition, selecting the best set of features that help to reduce the dimension from dataset is considered as the most important issue. Feature extraction aims at reducing the number of features extracted from dataset and chooses the features which are similar in the same class and very different from other different classes. After data dimensionality has been reduced during features extraction process, classification step will yield saving in memory and also alleviate the worst effects of the curse of dimensionality (CD) (Bellman, 1957). At first, CD increases dimensionality of the feature vector space which enhances the classification accuracy but rapidly leads to sparseness of the training data, poor representation of the vector densities, which decreases classification accuracy. Thus, many experts in the field, such as the work in (Duda et al., 2012) emphasized that to properly carry out behaviour recognition it is necessary to use the right features.

In Table 2.2 we show an example of features that can be extracted from smartphone sensors and infer the behaviour pattern of the smartphone users. Collectively, GPS, microphone, and accelerometers have proven to be effective at inferring more complex human behaviour. The microphone is one of the most ubiquitous smartphone sensor. They are capable of detecting individuals being involved at social interactions (*e.g.*, verbal interaction), or their surrounding ambient noise (Lu et al., 2009). Researchers in the field have demonstrated that variety of human activities can be inferred from multi-modal sensors, *e.g.*, significant places and activity level from GPS and accelerometers (Jeong et al., 2007; Proper et al., 2003).

Table 2.2: Example of feature extraction from smartphones.

Data Type	Sensor Type	Description
Physical Activity	Accelerometer Accelerometer Accelerometer	# Number of level of Activities # Number of Steps # Intensity of motor activity
Gestures Recognition	Magnetometer Gyroscope	# Direction of a movement # Gestures
Locations	GPS Cell Tower Google Maps Wi-Fi	# Location Clusters (Outdoor locations) # Most frequented places (Indoor, Outdoor locations) # Most frequented places (Indoor, Outdoor locations) # Most frequented places (Indoor, Outdoor locations)
Email contacts	Phone sensor Phone sensor	# Number of Messages # Number of Characters
Phone call contacts	Phone sensor Phone sensor	# Number of Calls # Duration of Calls
Calendars	Phone sensor Phone sensor	# Number of Events # Location of Events
Applications	Phone sensor Phone sensor	# Count Application launches # Duration of Application launches
Categories of Applications	Phone sensor Phone sensor	# Count Application launches # Duration of Application launches
Web-browsing	Phone sensor	# Count Visits
Voice	Microphone Microphone	# Speech activity # Ambient Noise
Social Interactions / Proximity	Bluetooth Wi-Fi Microphone	# Count number address of Bluetooth Id Tags # Count similar AP address and location changes # Count verbal proximity

2.6.4 Classification

Finally, selected features obtained from complete datasets are used as input for the next processing step, namely the classification. Results obtained from the classifier are typically a discrete selection of one of the per-defined classes. The degree of classification difficulty may depend directly from the similarity relations between pattern belonging to a different classes. Thus, its performance accuracy is significantly affected by the feature extraction stage.

Next, a general framework of supervised learning, semi-supervised and transfer learning methods is presented.

2.7 Learning from data

Machine learning is the field of study that is concerned with the question of how to construct computer applications that automatically improve with experience (Mitchell, 1997). In Figure 2.3, main types of techniques in ML are presented, such as supervised learning, and unsupervised learning.

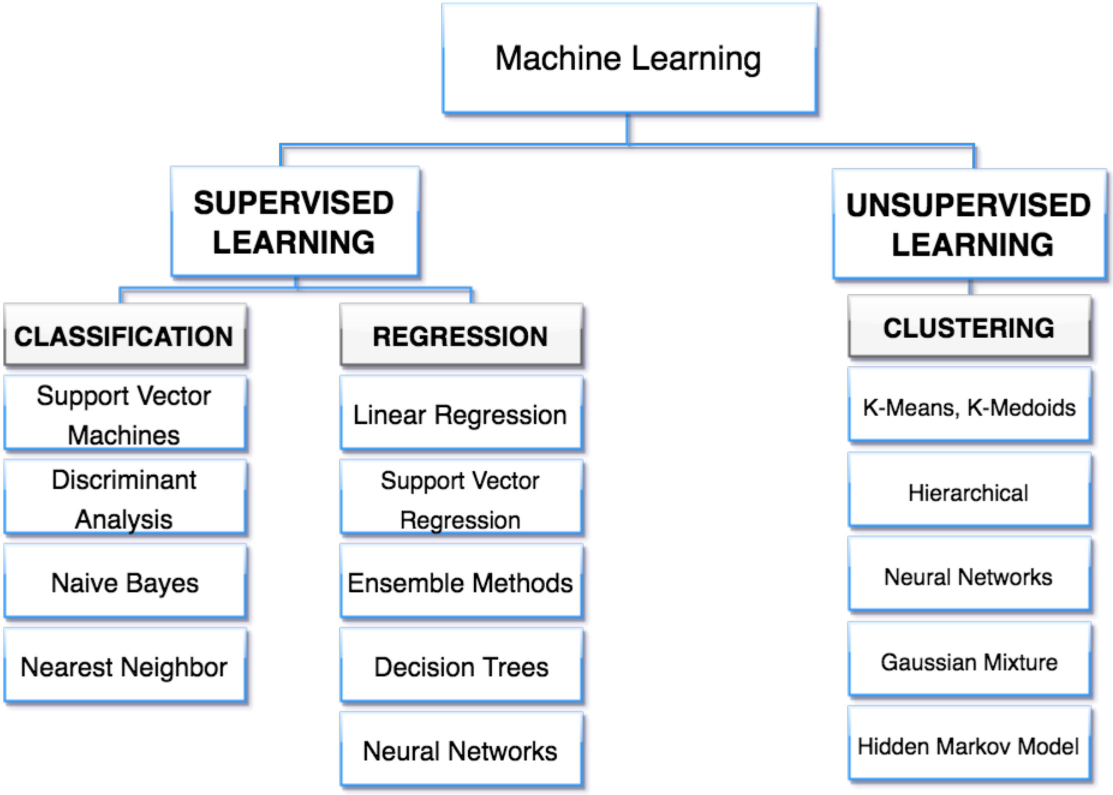


Figure 2.3: Learning paradigms.

2.7.1 Supervised learning

One important task of supervised learning is classification, where usually data is known before the learning task starts, which is called *offline* learning. Data consists of a set of examples containing a feature vector X_i and a label (class) Y_i. A supervised learning algorithm produces a function $g : X \to Y$, with X and Y input and output spaces, respectively. In order to satisfy classification performance requirements, the following conditions are required: (a) all data instances should be assigned to a class, and (b) all data instances are assigned to only one class. There exists different techniques for performing classification, such as Bayesian Networks (BN) (Pearl, 2014), Support Vector Machines (SVM) (Vapnik et al., 1997) and Decision Trees (DT) (Quinlan, 1993) (as shown in Figure 2.3).

There are several methods that have been developed for supervised classification methods in human behaviour recognition and are listed in the Table 2.3.

Decision Tree (DT)

Decision trees are the most commonly used decision modeling techniques. As a powerful

Table 2.3: Most common supervised classification methods.

Classification Methods	**Description**
Naive Bayes (NB)	Naive Bayes is one of the most efficient and effective inductive learning algorithm. It is known as probabilistic classifier which uses Bayes' theorem with naive independence assumptions to simplify the estimation of $P(X\|C) = \prod_{i=1}^{n} P(X_i\|C)$, where $X = (X_1,...,X_n)$ is a feature vector and C is a class (Rish, 2001).
Bayesian Network (BN)	Bayesian Network is a probabilistic graphical model. It represent a probabilistic dependencies among the corresponding variables of interest by using training dataset. It is often used in healthcare studies to learn relationships between the symptoms and the disease outcomes (Friedman et al., 1997).
k-NN	k-NN classifier is based on the closest training instances in the feature space. Euclidean distance k is used to measure similarity between instances by finding the closest instance (Altman, 1992). k denotes the number of classes.
Support Vector Machine (SVM)	SVMs are binary classifiers, derived from statistical learning theory and kernel-based methods (Cortes and Vapnik, 1995). SVM classifier separates the classes with decision surface that maximizes the margin between the classes (data points closest to decision surface *support vectors*). While SVM is a binary classifier, it is often used as a multi-class classifier by combining several binary SVM classifiers.
Decision Tree (DT)	Decision tree algorithms are used extensively for data mining in many domains. DT is a tree data structure consisting of decision nodes and leaves and the leaf specifies a class value (Witten and Frank, 2005). Decision Tree algorithms predicts the labeled instances based on features values. Decision nodes of the tree denote the different features whereas the branches between nodes provide possible values that selected feature can have. Leaf nodes provide the final classification accuracy. The algorithm used to generate a decision tree is information entropy (Witten and Frank, 2005).

classification algorithm, DT are becoming increasingly popular in the field of information systems applications in healthcare and medicine, including in mental-health (Batterham et al., 2009). The most popular DT algorithms include Quinlan's ID3, C4.5, C5 (Quinlan,1993) and Breiman's Classification and Regression Tree (Breiman et al., 1984). In clinical research studies, decision tress were widely used in disease models and are often used to represent the progress of patients wellbeing through different degree of their states over time (Batterham et al., 2009).

Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. Learned trees can also be transformed to sets of if-then rules to improve human readability (Mitchell, 1997). The objective of a decision tree is to specify a model that predicts the value of a certain variable, called *class*, given that some input information is provided.

Definition: (Decision tree). *A decision tree D is composed of nodes which represent tests to be carried out on variables known as* attributes. *Each test has different outcomes, which are branches of the node. These outcomes can be of two types: a leaf in which a*

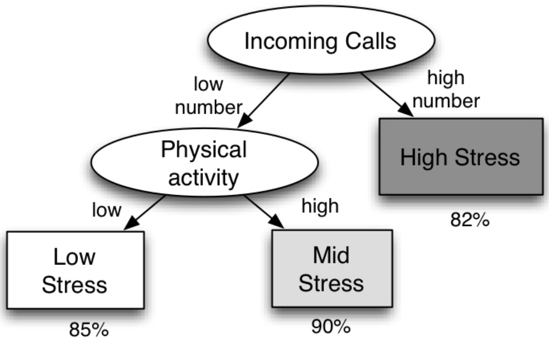


Figure 2.4: An example of a decision tree that classifies the level of Stress of a subjects. Ovals represent decision nodes. Rectangles are leaves (terminal nodes) that give the classification value, in this case they represent low, mid or high level of stress. Below each leaf accuracy is presented as a percentage.

value for the class *(predicted variable) is provided and represents a final node for the tree. Or it can be another test.*

One of the most well-known algorithms for learning decision trees from a batch of information is C4.5 Quinlan, 1993. In our domains, trees are useful to represent individuals wellbeing. For example, in Figure 2.4 a decision tree to predict the stress level is depicted. Each oval represents a decision node and rectangles correspond to a stress level (*low*, *mid*, *high*) of a person.

There are different performance measures to evaluate the prediction quality. Let TP, FP, TN and FN be the number of true positives, false positives, true negatives and false negatives, respectively:

- **Accuracy**: $\frac{TP+TN}{TP+TN+FP+FN}$
- **Precision**: $\frac{TP}{TP+FP}$
- **Recall**: $\frac{TP}{TP+FN}$
- **F-score**: $2 \cdot \frac{(precision)\ (recall)}{precision+recall}$

When using decision trees, a sensible measure to compare them is needed. There are two common approaches to compare decision trees, measures based on comparing the structure (Shannon and Banks, 1999) and measures based on comparing the prediction results (Miglio, 1996). Miglio, 1996 presented a dissimilarity measure that can combine the structure (the nodes attributes) and predictive (the predicted classes) similarities in a single value (Miglio and Soffritti, 2004). Let D_i and D_j be two trees with H and K leaves respectively used to classify n observations. We label $1, \ldots, H$ D_i leaves, and $1, \ldots, K$ D_j leaves to form the matrix:

$$M = [m_{hk}] \; h = 1, \ldots, H \text{ and } k = 1, \ldots, K$$

where m_{hk} is the number of instances which belong to both hth D_i leaf and to kth D_j leaf and $m_{h0} = \sum_{k=1}^{K} m_{hk}$, $m_{0k} = \sum_{h=1}^{H} m_{hk}$.

The dissimilarity measure is defined as:

$$d(D_i, D_j) = \sum_{h=1}^{H} \alpha_h (1 - s_h) \frac{m_{h0}}{n} + \sum_{k=1}^{K} \alpha_k (1 - s_k) \frac{m_{0k}}{n} \tag{2.1}$$

where m values measure the predictive similarity and α and s values measure the structural similarity. In detail, s_h coefficient is a similarity coefficient whose value synthesizes similarities s_{hk} between hth leaf of D_i and K D_j leaves. The value s_{hk} measures similarities of two leaves taking into account their classes and objects they classify:

$$s_{hk} = \frac{m_{hk} c_{hk}}{\sqrt{m_{h0} m_{0k}}} \; k = 1, \ldots, K$$

where $c_{hk} = 1$ if the hth leaf of D_i has the same class label as the kth lead of D_j, and $c_{hk} = 0$ otherwise. Choosing the maximum s_{hk} is a way to synthesize them as:

$$s_h = \max\{s_{hk} \; k = 1, \ldots, K\}. \tag{2.2}$$

Coefficient $\alpha_h = q - p + 1$ is a dissimilarity measure computed between a leaf of D_i and with respect to the leaf identified by Equation 2.2 of D_j. When paths associated to those leaves are not discrepant, then the value is set equal to 0. If, on the contrary, those paths are discrepant, the value is > 0 depending on the length of the longest path, p, and the level where two paths differ from each other, q. The maximum value of $d(D_i, D_j)$ can be reached when the difference between the structures of D_i and D_j is maximum and the similarity between their predictive powers is zero. The normalizing factor for $d(\mathrm{T}_i,\mathrm{T}_j)$ is thus equal to:

$$\max d(D_i, D_j) = \sum_{h=1}^{H} \alpha_h \frac{m_{h0}}{n} + \sum_{k=1}^{K} \alpha_k \frac{m_{0k}}{n}$$

where α_h is the length of the path from the root node to the h^{th} leaf. Thus, the normalized version of the dissimilarity is:

$$d_n = \frac{d(D_i, D_j)}{\max d(D_i, D_j)} \tag{2.3}$$

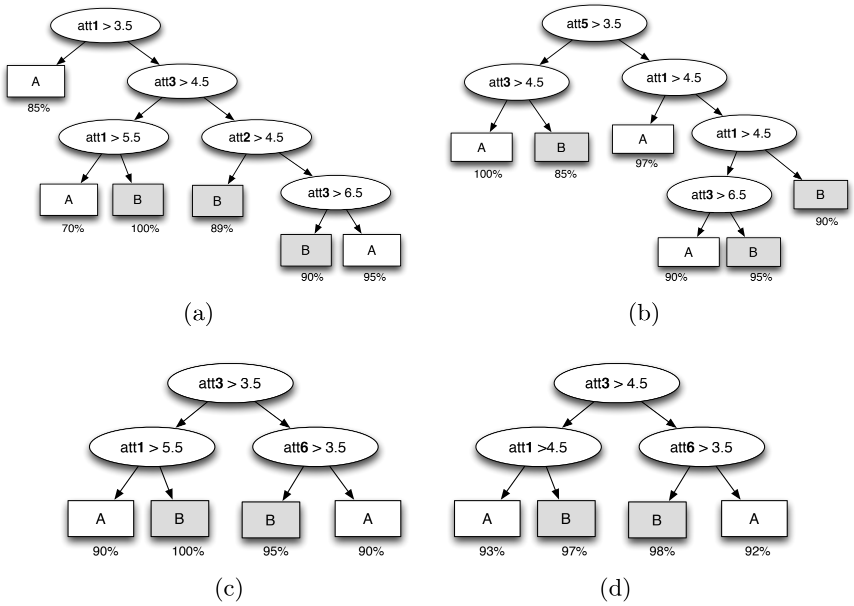


Figure 2.5: Example of highly dissimilar decision trees (a) and (b) using measure in Equation 2.3 (since their paths and predictions differ); in contrast (c) and (d) depict highly similar trees since the attributes in the nodes are the same and the predictions are similar.

where a $d_n = 0$ represents that the trees are very similar[1] and $d_n = 1$ that they are totally dissimilar. The normalization factor defined in Equation 2.2 can be interpreted as the weighted sum of paths lengths from the root node to all leaves of both trees. The length of each path is weighted with the proportion of observations classified in the related leaf.

Now, we present some trees with results using the dissimilarity measure presented in Equation 2.3. We refer to the reader to (Miglio and Soffritti, 2004) for a more detailed example. Figures 2.5 (a) and (b) depict trees with a high dissimilarity value, ($d = 0.38$). The reason is that paths are discrepant (structural similarity) and their predictive classification is different. In contrast, Figures 2.5 (c) and (d) depict highly similar trees, ($d = 0.0$), note that attributes in the nodes are the same (even when the split value is different they are considered the same).

C4.5 Classifier

Among decision tree algorithms, the C4.5 tree-induction algorithm deserves a special mention for several reasons, including their good classification accuracy and is the fastest (*i.e.*, for large amount of datasets) compared with main-memory algorithms for machine

[1] Nodes with numeric attributes with the same variables but with different splitting values are seen as totally similar.

learning and data-mining (Quinlan, 1993). The C4.5 is an extension of the ID3 algorithm used to improve its disadvantages:

- Dealing with training data that have missing values of attributes.
- Handling different cost in the tree.
- Pruning the decision tree after its construction (namely post-pruning).
- Handling attributes with discrete and continuous values.

C4.5 algorithm constructs a big trees with a *divide and conquer* strategy (Quinlan, 1993). The trees are constructed by considering amount of attribute values and finally it applies the decision rule by pruning. In C4.5 pruning trees after creation, it prevents the tree from over-fitting and attempts to remove branches in the tree by replacing them with leaf nodes (as shown in Algorithm 1). Similarly, as shown in the Figure 2.6, decision trees are constructed as following:

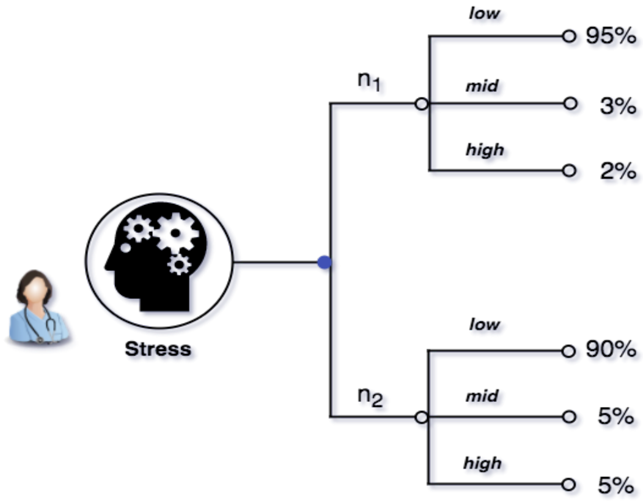


Figure 2.6: Example of C4.5 decision tree nodes.

- On top of the node of the tree are root nodes that select the attributes that are most significant.
- The measured information is passed to branch of nodes (*e.g.*, branch n_1 and n_2) which terminate in leaf nodes that give decisions.
- Finally, rules are generated by highlighting the path from the root node to leaf node.

The construction of DT classifiers are relatively fast and the accuracy of decision trees is often superior if we compare with other models. DT algorithms present several advantages over other learning algorithms, due to their robustness and lower computational cost for generating of the model. The models created from DT are capable to predict the class based on several input variables, *e.g.*, each node correspond to one of the input attributes and edges to children for each of the possible values of that input attribute (as shown in Figure 2.6). Every leaf in the tree represents a value of the target variable given the values of the input attributes defined by the path from the root to the leaf (Witten and Frank, 2005).

Algorithm 1: C4.5 Algorithm

```
Input: an attribute-valued dataset $D$
Tree = [ ]
if $D$ is "pure" then
    terminate
end if
for all attribute $a \in D$ do
    Compute information-theoretic criteria if we split on $a$
end for
$a_{best}$ = Best attribute according to above computed criteria
Tree = Create a decision node that test $a_{best}$ in the root
$D_v$ = Induced sub-datasets from $D$ based on $a_{best}$
for all $D_v$ do
    $Tree_v$ = C4.5($D_v$)
    Attach $Tree_v$ to the corresponding branche of Tree
end for
return Tree
```

C4.5 can be built by splitting the dataset into subsets based on an attribute value test and can be repeated on each subset in a recursive manner (namely recursive partitioning). The recursion process finalizes when splitting no longer adds value to the predictions or when the subset at a node has achieved same value of the target variable. The structure of DT algorithms are based on a greedy top-down recursive partitioning for tree growth and uses various impurity measures, *information gain (IG)*, *gain Ration*, *Gini Index* and *distances based measures* as an input attribute to be associated with an internal node.

To form DT, the following steps are required:

1. **Step 1: Define *x* entropy,**

$$H(X) = \sum_{j} pj\, log2(p_j) \tag{2.4}$$

where x is a random attribute with k discrete values which are distributed according to probability value $P = (p_1,\ p_2,...,\ p_n)$.

2. **Step 2: Calculate the *weighted sum* the entropies for each subsets,**

$$H_T = \sum_{i=1}^{k} P_i H_S(T_i) \tag{2.5}$$

where P$_i$ is the proportion of attributes in subset i.

3. **Step 3: Measurement of information gain,**

$$Information\ Gain\ IG\,(S) = H(T) - H_S(T) \tag{2.6}$$

The information gain (IG) is the criterion needed for selecting the most effective attribute in order to make decision. The selection of the attribute at each decision node would be the one with the highest IG.

Moreover, one of the unique feature of C4.5 algorithm is handling with missing attributes in the dataset. The C4.5 uses probability values for missing attributes rather then assigning existing most common values of that attribute. Handling missing attribute values is an important issue for classifier learning, since it can affect the prediction accuracy of learned classifiers. Thus, C4.5 has gained increased attention in semi-supervised learning methods to address the missing instances for improving the classification performance.

2.7.2 Ensemble learning techniques

One technique used by machine learning to increase the accuracy of different classifiers is to use several of them and then join their collective decisions into one. These are called ensemble methods which use multiple models to obtain better predictive performance than could be obtained from any single model. By joining multiple classifiers decisions into one final classifier, ensemble methods aim at leveraging the wisdom of the crowds (Rokach, 2010). Their task can be described as a group of individuals trying to solve one particular problem, but within the group might be an individual very skilled to lead the group toward a correct solution, however, there is still an advantage to have the rest of the group around.

Two most popular methods are Bagging (Breiman, 1996) and Boosting (Freund, Schapire, et al., 1996). Bagging methods train multiple instances of a classifier on different subsamples (bootstrap samples) of the training data (Breiman, 1996). Decision are made by a majority vote among the base classifiers. On the other hand, Boosting methods (Freund, Schapire, et al., 1996), training data is more logically by sampling instances that are difficult for the existing ensemble to classify with higher preference.

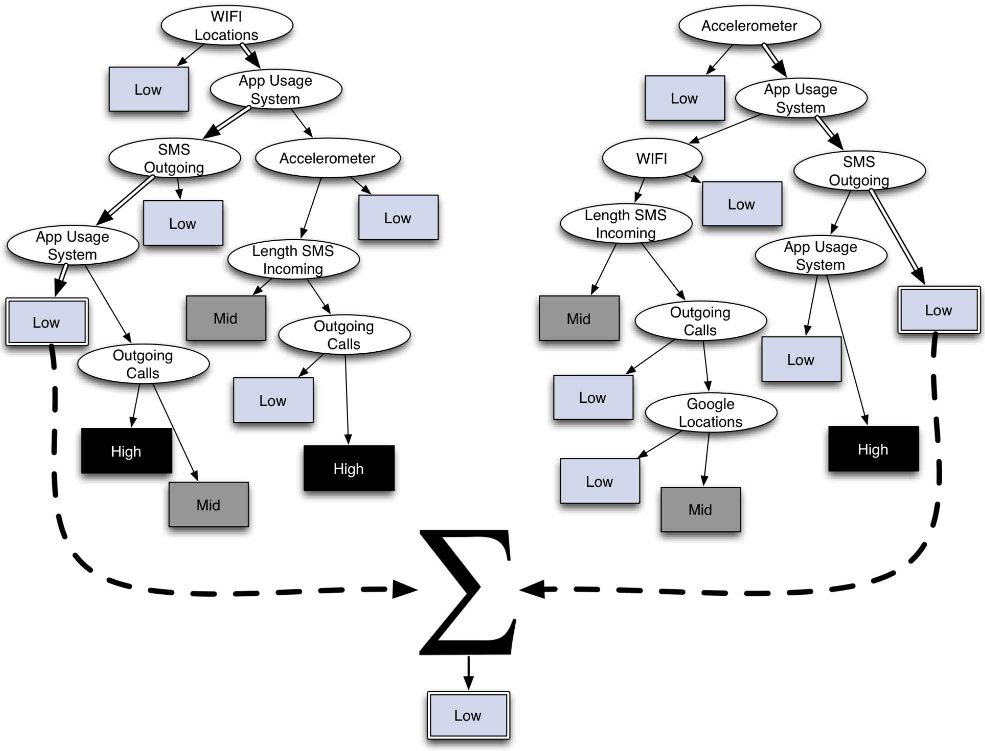


Figure 2.7: Example of the ensemble learning.

In particular, one ensemble method commonly used is called *random forests* (Dietterich, 2000) and it is based on decision trees. The method constructs a multitude of decision trees at training time and the predicted class is the mode of the classes of the individual trees. In our research work, we have used weighted ensemble of models that is used after transfer learning is applied (see the Algorithm 3) and discussed in Chapter 6. The example of ensemble learning used in prediction of stress at work (in Chapter 6 is presented in Figure 2.7).

When dealing with real-world data it is likely to have missing data, some techniques from machine learning that deal with this problem are called semi-supervised learning techniques.

2.7.3 Semi-supervised learning (SSL)

Semi-supervised learning (SSL), is in fact a missing link between the supervised learning and clustering methods. Having a limited training set, using the SSL aims to accurately predict correct classes for unseen data. Semi-supervised learning has got various applications in real life. It became particularly popular in the 1990s when it proved to be useful technique in text classification and natural language processing (Zhu, 2006).

According to (Chapelle et al., 2006b),
"SSL is halfway between supervised and unsupervised learning. In addition to unlabeled data, the algorithm is provided with some supervision information – but not necessarily for all examples."

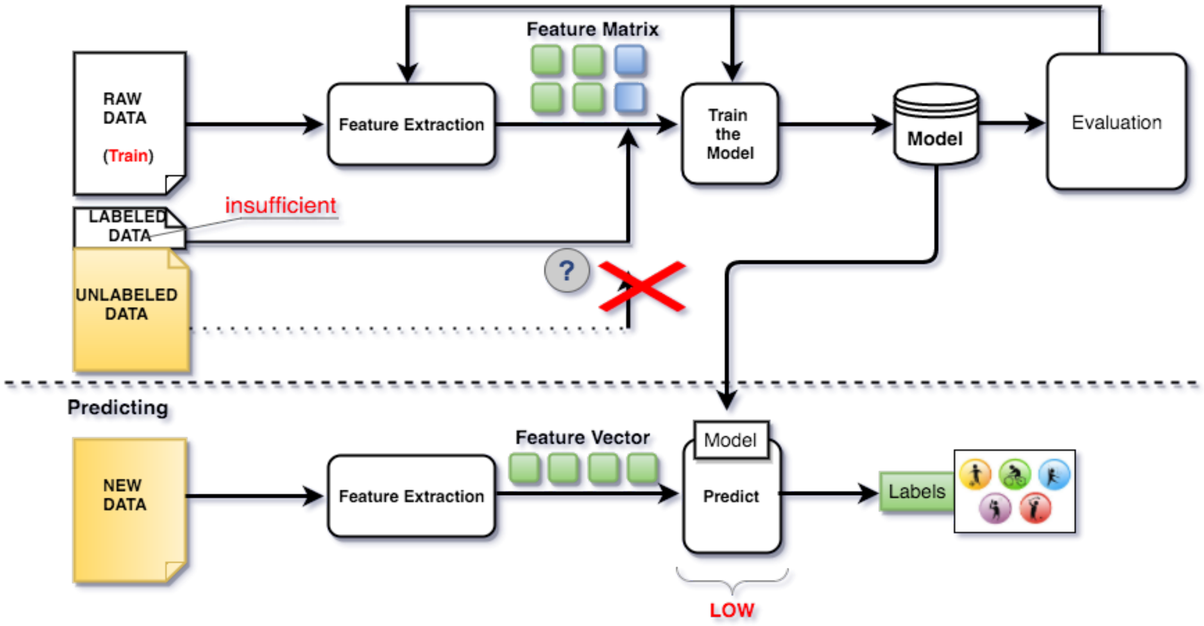


Figure 2.8: Issues in model learning and usage process using supervised learning methods.

Definition of semi-supervised learning

Semi-supervised learning methods have been suggested in machine learning field as the right choice aiming to exploit unlabeled samples to improve learning performance (Longstaff et al., 2010; Zhu, 2006). The main objective of semi-supervised learning in machine learning is to combine the advantages of supervised and unsupervised approaches by learning from both labeled and unlabeled data. In Figure 2.9 the advantage to utilize and to exploit the costless unlabeled data during the training process makes semi-supervised learning algorithms to be one of the hottest research topics in machine learning. There are a number of different algorithms for semi-supervised learning, some are designed specifically for a classifier such as semi-supervised SVMs (S3VM) (Zhu, 2006). Others offer a general approach for any classifier period.

Self-training

In this section, we briefly describe the method used in our research for semi-supervised learning, namely, *Self-training* (Nigam and Ghani, 2000). Self-training approach allows a classifier to start with a small amount of labeled instances to build an initial classifier and later to incorporate both labeled and unlabeled data with the aim at improving the accuracy performance. As discussed in previous chapter, having small amount of labeled instances is a common problem in machine learning. Let us assume that we have a set L (usually small amount) of labeled instances, and a set U (usually large) of unlabeled data. As shown in Figure 2.9 supervised methods will ignore unlabeled instances to build a classifier.

Algorithm 2: Self-Training Algorithm

```
Input: L = (x_i, y_i); set of labeled instances
       U = (x_i, ?); set of unlabeled instances
       T; threshold for confidence
while U ≠ ∅ or U' ≠ ∅ do
    Train a classifier C with training data L
    Classify data in U with C
    Find a subset of U' of U with the most confident scores (confidence > T)
    L + U' ⟹ L
    U - U' ⟹ U
```

Using self-training algorithm only one classifier is need, thus, only one feature set is required. This classifier is trained on existing labeled data and then applied on a set of unlabeled data. For several iterations, the classifier labels the unlabeled data and includes the most confidently predicted instances of each class into a labeled training set Nigam and Ghani (2000). Algorithm 2 shows the pseudo-code for a typical self-training algorithm. Self-training begins with a set of labeled data L, and builds a classifier C, which is then applied to the set of unlabeled data U. T which is the set of most confidently predicted instances are added to the labeled set. The classifier is then retrained on the new set of labeled instances, and the process continues for several iterations (see Figure 2.9).

In this thesis, we focus on the self-training algorithm (Zhu, 2006) that uses its own predictions to assign values to unlabeled data that achieved higher confidence in predictions (in our studies we use confidence $\geq 80\%$). The unlabeled data with high confidence in its predicted class is added, with its class, to the labeled data. This new augmented labeled data is used to induce a new model from which new predictions over the reduced unlabeled data are produced (see Algorithm 2). The procedure is repeated until there are no more instances above the threshold value or until the unlabeled data becomes empty. Adding new labeled instances acquired from unlabeled data, is often shown to achieve a

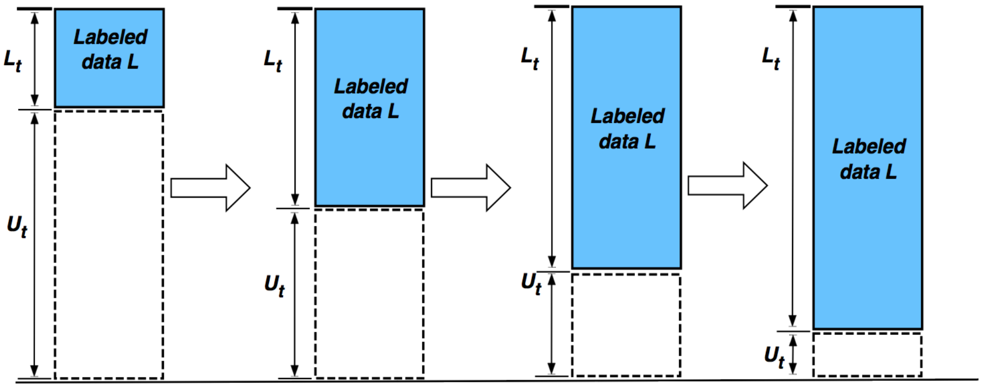


Figure 2.9: Semi-supervised learning method (SSL), where L represents labeled instance, U unlabeled instances, and t number of iterations, $L = \mathrm{L}_t \cup \mathrm{U}_t$.

better accuracy than supervised learning that uses only the labeled data.

2.7.4 Transfer Learning (TL)

Being capable to learn an accurate model for predicting subjects outcomes from a specific behaviour typically depends on the amount of available training data. Acquiring sufficient labeled data is often very difficult and expensive to obtain in many domains. A system with the capability to use not only labeled but also unlabeled data holds a great promise in terms of broadening the applicability of learning methods. In this regard, the area of machine learning has proposed semi-supervised methods to overcome these problems. However, these methods assume that both labeled and unlabeled data are generated from the same distribution. In contrast, a more general approach will allow these distributions to be different, this is the case of Transfer Learning (Rashidi and Cook, 2010). In this way, we can benefit from previous acquired knowledge from other related domain, task or model to improve our learning process.

TL methods have been successfully applied to establish more accurate models using scarce data (Luis et al., 2010) in different domains such as social networking (Roy et al., 2012), text classification (Roy et al., 2012), image classification (Raina et al., 2007) and indoor and outdoor localization problems (Pan et al., 2008). While these are only a handful of examples, TL has been used in many other applications as shown in the surveys in (Pan and Yang, 2010; Weiss et al., 2016). However, in the healthcare domain, the use of TL is still in its infancy. For our work *related model* refers to information from other subjects, that is when a new subject is added into the system, it is expected to have scarce data.

In this thesis, we used the following approach to address scarcity of data:

◆ Initially, we learn a model T_i for a new subject i using the available data.

- We compare the model with the rest of the T models generated for the other subjects.
- Finally, we apply transfer learning to infer a better model.

Our proposed approach is described in more detail in Algorithm 3 where decision trees have been used in to induce subjects models.

A categorization of Transfer Learning techniques

In transfer learning, we have the following three main research issues:

- *What to transfer*
- *How to transfer*
- *When to transfer*

"What to transfer": focuses in understanding knowledge that can be transferred across tasks. This knowledge can be similar between the individuals tasks that may help improve performance for the targeted task. When similarity between individuals is determined, this knowledge can be transferred which corresponds to *"How to transfer"*. At this step learning algorithms need to be developed to transfer knowledge.

"When to transfer": focuses in transferring intelligence that should be used. We are interested in knowing in which cases knowledge transfer can be applied. For instance, in situation where the source domain and target domain are not related, transfer may result unsuccessfully. In our dataset collected from bipolar disorder patients, transfer learning could not be applied due to small number of participants and due to different degree of their state and would result to negative transfer.

Figure 2.10 presents our approach proposed combining TL and SSL which has been applied in data collected from 30 employees at working environments.

2.7.5 Intermediate models

The information provided by the users through questionnaires is useful, however, it is a tedious task for each user. In this research, we propose to predict the mood variables associated with questionnaires using data from smartphone to alleviate the user from this burden. Then, the predicted mood variables are used with the rest of data from the smartphones to predict the class, in our experiments, the mood state of a bipolar disorder patient or stress levels at working environments (see Algorithm 4). We call the models that predict the mood variables from the questionnaire: *intermediate models* as they are used as input for the final predictive model.

In terms of machine learning techniques, although we can relate this technique with other existing methods, we are not aware of any research that uses the same approach. For instance, some techniques use *latent variables* to help to create better predictive models.

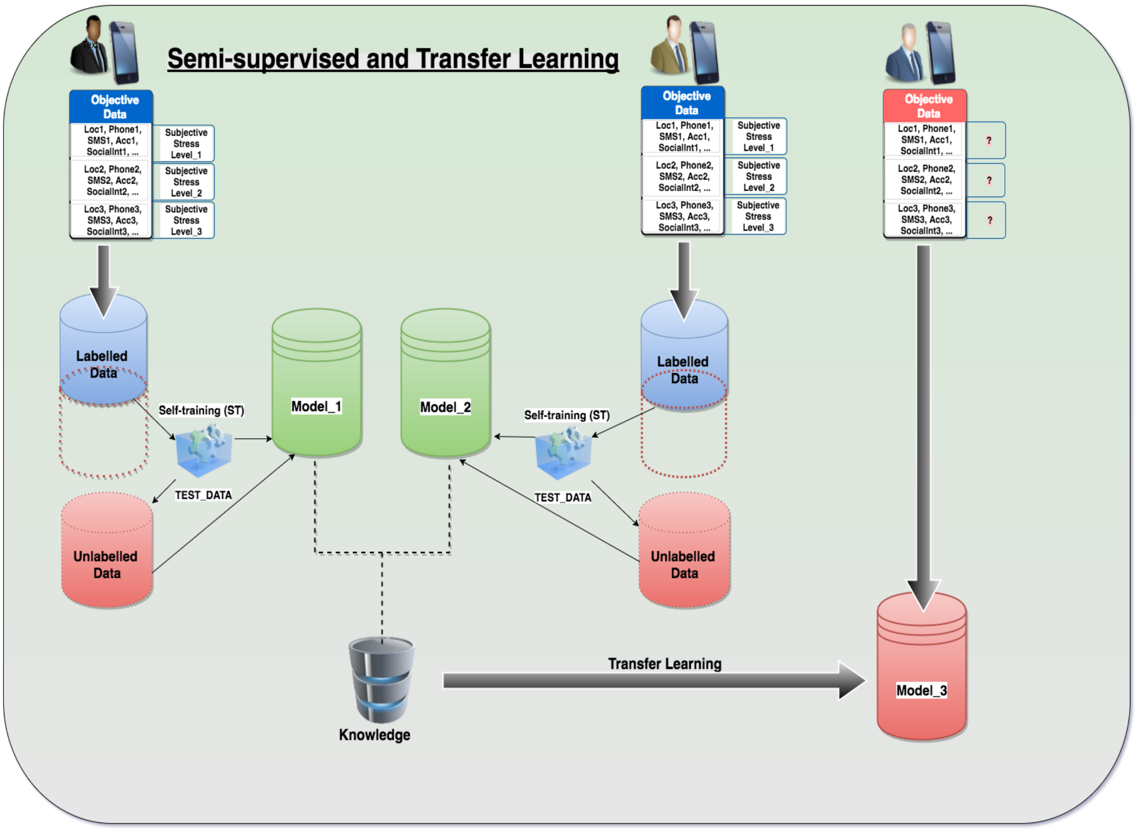

Semi-supervised and Transfer Learning
Objective Data
Loc1, Phone1, SMS1, Acc1, SocialInt1, ...
Subjective Stress Level_1
Loc2, Phone2, SMS2, Acc2, SocialInt2, ...
Subjective Stress Level_2
Loc3, Phone3, SMS3, Acc3, SocialInt3, ...
Subjective Stress Level_3
Labelled Data
Self-training (ST)
TEST_DATA
Unlabelled Data
Model_1
Model_2
Objective Data
Loc1, Phone1, SMS1, Acc1, SocialInt1, ...
Subjective Stress Level_1
Loc2, Phone2, SMS2, Acc2, SocialInt2, ...
Subjective Stress Level_2
Loc3, Phone3, SMS3, Acc3, SocialInt3, ...
Subjective Stress Level_3
Labelled Data
Self-training (ST)
TEST_DATA
Unlabelled Data
Objective Data
Loc1, Phone1, SMS1, Acc1, SocialInt1, ...
?
Loc2, Phone2, SMS2, Acc2, SocialInt2, ...
?
Loc3, Phone3, SMS3, Acc3, SocialInt3, ...
?
Knowledge
Transfer Learning
Model_3

Figure 2.10: Transfer learning with self-training, proposed method.

These hidden variables are artificially created and used as intermediate information to build better models. In our case, we know in advance exactly how many variables to use and we have some information (values) for these variables, which allow us to produce better models.

Another related technique is precisely semi-supervised learning, where there is some labeled data and a normally larger set of unlabeled data. In our case, what we are missing is not the class labels, but a large proportion of information of useful features that can be used to build a better predictive model. What we propose is to use the available information to fill-in the missing data for some of the attributes.

Normally when there is some missing data, researchers have used imputation methods. These methods try to complete missing data using, for instance, the most common value, the most probable value given the class, or induce a model to predict the missing values using all the information from features and the class. In our case, we are not using class labels for the induced intermediate models, we target the process to very specific features (those involving the intervention from the user) and assume that reliable models can be built from available data (in our case from information obtained from smartphones).

Algorithm 3 Transfer Learning used in our research with four different transfer learning strategies.

```
Let $D_T$; dataset from target user
Let $\{D_1, \ldots, D_n\}$; datasets from other users
Let $M_{all} = \{M_1, \ldots, M_n\}$; induced models from other users
Let $Th$ = threshold value
Induce model $M_T$ using $D_T$
for each $M_i \in M_{all}$ do
  Find similarity value with $M_T$ $(sim(M_T, M_i))$
end for
Sort $M_{all}$ using $sim(M_T, M_i) \mid M_i \in M_{all}$
Use one of the following TL strategies:
if Naïve then
  Select most similar model $M_i$ (first element in $M_{all}$)
  Select data $D_i$ used to construct $M_i$
  Induce new model $M_T$ with $\{D_T \cup D_i\}$
else if Theshold then
  Select the most similar models $M_{sim} = \{\bigcup_i M_i \mid sim(M_T, M_i) > Th\})$
  Select $D = \{\bigcup_i D_i \mid D_i$ was used to induce $M_i \in M_{sim}\}$ )
  Induce new model $M_T$ with $\{D_T \cup D\}$
else if Sampling then
  Select the $K$ most similar models $M_K$ = first $K$ elements in $M_{all}$
  Select $D = \{\bigcup_i D_i \mid D_i$ was used to induce $M_i \in M_K\}$ )
  Let $D' = \{\bigcup_i$ sample $D_i \in D \propto sim(M_T, M_i)\}$
  Induce new model $M_T$ with $D_T \cup D'$
else if Ensemble then
  Select the $L$ most similar models $M_L$ = first $L$ elements in $M_{all}$
  Create a weighted ensemble of models $\{M_T \bigcup_{i=1}^{L} w_i M_i \mid w_i = sim(M_T, M_i) \wedge M_i \in M_T\}$
end if
```

Algorithm 4 Intermediate Models

```
Let $D_1$; dataset (matrix) with more instances (e.g., variables from smartphones)
Let $D_2$; dataset (matrix) with fewer instances (e.g., variables from questionnaires)
Let $Y$; set (column vector) with associated classes (e.g., state bipolar/stress value)
% Build intermediate models
for each variable (column) $x_i \in D_2$ do
  Train a classifier $C_i$ with training data $(D_1, x_i)$
end for
% Create estimated values for $D_2$
for each $C_i$ do
  for each instance (row) $e_j \in D_1$ do
    Use $e_j$ as input to $C_i$ to predict an instance (row) of $\hat{D}_2$
  end for
end for
% Induce final classifier
Train a classifier $C_{final}$ with training data $((D_1 \cup \hat{D}_2), Y)$
```

For training we follow these steps:

1. Use initial data (smartphone + questionnaires) to predict mood variables associated to the questionnaires .
2. Trained a classifier to predict a weighted value (based on accuracy) for each of the variables associated to questionnaires.
3. Use smartphone data and predicted variables to induce a model to predict the episodic state of a bipolar disorder patient or stress levels.

For testing we follow these steps:

1. Use information from smartphones to predict, with intermediate models, a weighted set (based on accuracy) of mood variables.
2. Use information from smartphones and predicted mood variables to predict the final model

In this thesis, we used three variables for bipolar disorder and six variables for stress to characterize information from questionnaires. Consequently, we induce three and six classifiers, respectively, for bipolar and stress applications.

2.8 Chapter Summary

In this chapter, we reviewed some of the most important concepts related to feature extraction and machine learning methods which will be relevant for the approaches described in Chapter 6 and Chapter 5. We presented the algorithms that were used in this research work. Finally, we demonstrated the novelty of using intermediate models and the importance in building final models. In the next chapter we focus on recent works which are related to this thesis.

Chapter 3

RELATED WORK

> *"Ultimately, I hypothesize that technology will one day be able to recreate a realistic representation of us as a result of the plethora of content we're creating converging with other advances in machine learning, robotics and large-scale data mining."*
>
> – ***Adam Ostrow***

There are various applications for semi-supervised learning and transfer learning. Depending on their properties, different models can be derived. The purpose of this chapter is to review some applications of both approaches when addressing scarce data. Section 3.1 is about semi-supervised learning when targeting scarce data. Section 3.1, shows how semi-supervised learning helps to find the best model from a fixed set of models to solve a problem. Section 3.2, describes the transfer learning algorithm together with an interesting application for transfer learning is healthcare. We examine the use of transfer learning for this problem in Section 3.2.

3.1 Semi-supervised learning in scarce data

Semi-supervised learning approaches have been proposed and widely studied in order to target scarce data. We present the most important algorithms in this area, a more extensive survey is presented in (Zhu, 2006).

The main objective of semi-supervised learning is to combine advantages of supervised and unsupervised approaches by learning from both labeled and unlabeled data. Thus, due to their ability of using unlabeled data, semi-supervised learning is an actual topic of interest, within machine learning (Ma et al., 2010).

Semi-supervised learning has been suggested in several research studies (Dempster et al., 1977; Longstaff et al., 2010; Ma et al., 2010) as the right choice aiming to address this issue, which has shown to exploit unlabeled samples to improve learning performance.

However, it is good to note that there exists relatively little work exploring semi-supervised techniques withing the healthcare arena.

Co-training

Co-training and *self-training* are both bootstrapping methods, which belong to so called *"weakly supervised"* learning algorithms. Co-training method is similar to self-training, however, the difference is that co-training uses two classifiers to make predictions from unlabeled data. Similarly, as in self-training method, co-training is a wrapper method that uses two classifiers C_1 and C_2 that can assign a confidence score to their predictions (as shown in Algorithm 5). The two classifiers trained on two data "views" (v_1 and v_2) provide their most confident unlabeled prediction from the training set of each other (*i.e.*, $v_1 \rightarrow L_2$ and $v_2 \rightarrow L_1$).

The success of co-training using the views depends on the following two assumptions (Johnson and Zhang, 2007):

- Each view (v_1, v_2) alone are sufficient to make a good classification, give enough labeled data.
- Both views are conditionally independent given the class label.

The most obvious assumption is the existence of two separate views $v = [v_1, v_2]$. If the two assumptions hold, co-training classifier can learn successfully from labeled and unlabeled data. These assumptions have been examined for natural language processing tasks (Nigam and Ghani, 2000), and some research work has investigated the conditional independence assumption (Johnson and Zhang, 2007), due to its difficulty to find tasks in practice in which it is satisfied.

In cases when the conditional independence assumption is violated, co-training method may not perform well (Chapelle et al., 2006b; Johnson and Zhang, 2007). This means that, despite some theoretical co-training analysis (Balcan et al., 2004) it is merely a mean to know whether two classifiers C_1 and C_2 agree in predicting the same label on the unlabeled instances. The agreement is justified by learning theory, where not many candidate predictors can agree on unlabeled data in two views, the hypothesis space is small (Dasgupta et al., 2002). In situations where a candidate predictor in this small hypothesis space also fits the labeled data, it is less likely to be overfitting ad can be expected to be a good predictor.

Co-training methods make strong assumptions on features splitting. Goldman and Zhou, 2000 demonstrated the performance of two learning algorithms of different type which take the whole feature set. This is essentially used on learners with high confidence instances , identified with a set of statistical tests, in U to teach the other learning and vice versa. Other improvements of Co-training, (Zhou and Goldman, 2004) propose a single-view multiple-learner Democratic Co-learning algorithm. The ensemble of learners are trained separately on all features of labeled data, then make prediction on unlabeled data. If most learners agree on the class of an unlabeled point x_i, then classification uses x_i as a label. x_i and its label is added to the training data, where all learners are retrained again on the actual updated training set. Finally, the best prediction is decided based on majority vote among all learners.

Similarly, Zhou and Li, 2005 propose and advance Co-training, namely 'Tri-training' which uses instead three learners. In situations where two of the learners agree on the classification of an unlabeled instance, the classification is used to teach the third classifier. Strength of this approach avoids the need of explicitly measuring label confidence of any learner. This method can be applied to datasets without different views, or different types of classifiers.

Algorithm 5 Co-Training

```
Input: $\mathbf{L} = (\mathbf{x}_i, \mathbf{y}_i)$; set of labeled instances
U = $(\mathbf{x}_i, ?)$; set of unlabeled instances
    Training set $\mathbf{L}_1$ for classifier $\mathbf{C}_1$, where $\mathbf{L}_1$=L
    Training set $\mathbf{L}_2$ for classifier $\mathbf{C}_2$, where $\mathbf{L}_2$=L
T; threshold for confidence
WHILE   $\mathbf{U} \neq \emptyset$ or $\mathbf{U'} \neq \emptyset$
    Train a classifier $\mathbf{C}_1$ on $\mathbf{L}_1$
    Train a classifier $\mathbf{C}_2$ on $\mathbf{L}_2$
    Classify the unlabeled data with $\mathbf{C}_1$ and $\mathbf{C}_2$ separately
    Add $\mathbf{C}_1$'s most-confident prediction $T$ to $\mathbf{L}_2$
    Add $\mathbf{C}_2$'s most-confident prediction $T$ to $\mathbf{L}_1$
$\mathbf{L}_1 = \mathbf{L}_2 + \mathbf{U'} \Longrightarrow \mathbf{L}$
$\mathbf{U} - \mathbf{U'} \Longrightarrow \mathbf{U}$
```

Semi-supervised SVM (S3SVM)

Semi-supervised approaches differ from each other in the classifier's learning process. Considering the fact that using unlabeled data to learn can help improve the performance of supervised classifiers (*i.e.*, when its predictions provide new useful predicted information), as shown in Figure 3.1. Nevertheless, not always the new included incorrect predictions (*i.e.*, noise) can worsen the new learned model resulting in low performance of the classifier accuracy.

Semi-supervised learning for SVM (S3VM) has been first introduced by (Joachims,

1999) by optimizing the original SVM function (see Equation 3.1).

$$min\left[\frac{1}{2}\cdot||w||^2 + C\cdot\sum_{l}^{i=1}\zeta\frac{d}{i} + C^{*}\cdot\sum_{j=1}^{u}\zeta\frac{*d}{j}\right] \tag{3.1}$$

where u depict the amount of unlabeled data and parameters for unlabeled instances included in the learning phase $(\zeta\frac{*d}{j})$. The margin is measured using $\frac{1}{||w||}$ and minimizing the norm $||w||^2$ which is equivalent to maximizing the margin and satisfying the margin constraint for each data point (Joachims, 1999).

Joachims (1999) demonstrated the performance gap between the supervised SVM and the semi-supervised S3VM, in favour of the latter one. The goal of a S3VM is to find a labeling of unlabeled instances, so that a decision boundary has the maximum margin on both labeled and new added labeled instances. In S3VM, a SVM classifier has to be trained by solving a quadratic programming issue in every iteration (Booch et al., 1999). It is applied to classification tasks with large number of data sets and their computational cost is high (Joachims, 1999). Figure 3.1 (a) shows the support vector machine classifier where a straight line separates two classes and the linear boundary maximizes the geometric margin (*i.e.*, nearest positive (red dots) and negative instances (black dots)). Better decision boundary using S3VM is shown in Figure 3.1 (b) which falls between the unlabeled data. It separates two classes in labeled data. The margin is smaller than the Figure 3.1 (a) and new decision boundary is the one found by S3VMs that is defined by both labeled and unlabeled data.

Chapelle et al., 2006a have proposed an approximation solution to S3VM in order to understand S3VM global optimum. Using the Branch and Bound methods (Welch, 1982) authors finds the global optimal solutions for small datasets, with excellent accuracy. Despite the fact that, Branch and Bound methods are probably not useful for large datasets, results provide some ground truth, and T3VMs potential with better approximation methods.

On the ohter hand, Weston et al., 2006 proposes learning with a 'universum', which is a set of unlabeled data that does not come from two classes. But, the decision boundary is determined by passing through the universum. Authors find similar interpretation to the maximum entropy, where the classifier should be confident on labeled examples, and maximum ignorant on unrelated instances. In this line, Jaakkola et al., 1999 proposes a maximum entropy discrimination method to maximize the margin. The proposes method is able to take into account unlabeled data with SVM as a special case.

Other Semi-supervised Models

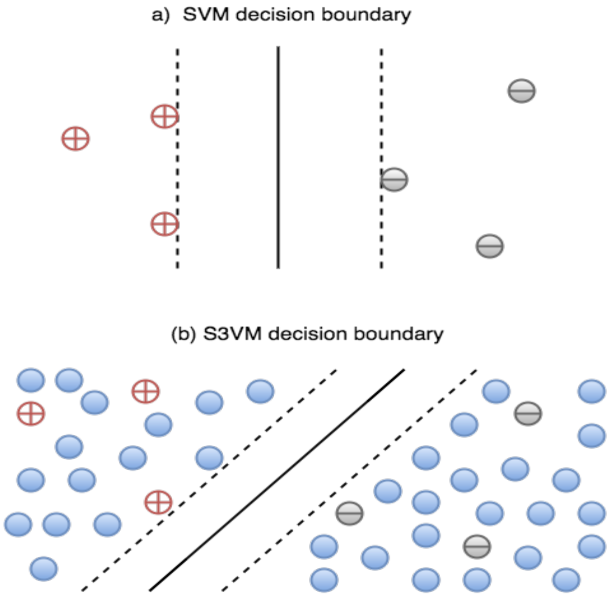


Figure 3.1: SVM vs S3VM, where black and red dots are labeled resource, and blue dots are unlabeled resources. a) Supervised SVM, only labeled data are included. The linear decision boundary that maximizes the distance to only labeled instance is shown in solid line and associated margin is shown in dashed lines, b) Semi-supervised SVM, unlabeled data are associated with the classes and the decision boundary seeks a gap in unlabeled data.

The important semi-supervised learning algorithms used in the literature are demonstrated in Table 3.1. There are other semi-supervised learning methods in the literature including:

- learning from positive and unlabeled data, when there is no negative labeled data (Denis et al., 2002)
- semi-supervised regression (Brefeld and Scheffer, 2006);
- advances in learning theory for semi-supervised learning (Amini et al., 2009)
- inferring label sampling mechanisms (Rosset et al., 2004), multi-instance learning (Zhou and Xu, 2007), multi-task learning (Liu et al., 2008), and deep learning (Ranzato and Szummer, 2008);
- model selection with unlabeled data (Kääriäinen, 2005)
- self-taught learning (Raina et al., 2007) and the universum (Weston et al., 2006), where unlabeled data do not derive from positive or negative classes, but rather from another third class of instances in the same general domain.

Table 3.1: A summary of semi-supervised learners with inductive property of the algorithm.

Approach	**Summary**
– Co-training	Increases prediction consistency among two distinct feature views (v_1, v_2)
– Self-training	Assumes pseudo-labels as true labels and re-trains the model (Rosenberg et al., 2005)
– TSVM, S3V	Margin maximization using density of unlabeled data (Fung and Mangasarian, 2001)
– Gaussian processes	Bayesian discriminative model (Lawrence and Jordan, 2004)
– Semi-supervised Margin Boost (SSMB)	Maximizes pseudo-margin using boosting (Grandvalet, Ambroise, et al., 2001)
– Assemble	Maximizes pseudo-margin using boosting (Bennett et al., 2002)
– Mixture of Experts	Expectation Maximization (EM) based model-fitting of mixture models (Miller and Uyar, 1997)
– EM-Naive Bayes	Expectation Maximization (EM) based model-fitting of Naive Bayes (Nigam et al., 2000)

3.2 Scarce data and transfer learning

The motivation for transfer learning in the field of machine learning was introduced in NIPS-95 workshop on "Learning To Learn" [1] with the focus on building machine learning methods that uses previously learned knowledge. Since then research on TL has attracted attention by different names, such as learning to learn, life-long learning, knowledge transfer, inductive transfer, multi-task learning, knowledge consolidation, context-sensitive learning, knowledge-based inductive bias, meta learning, and incremental/cumulative learning (Thrun and Pratt, 1998).

Demands for transfer learning approaches is described in Chapter 2 where in real-life settings, applications deal with missing labeled data. In some particular fields *i.e.* healthcare, a large amount of expert knowledge is needed. As a result, there is only a very limited amount of data available. Therefore, the reason for making an accurate prediction from the dataset with a lower labeled instance or none is a very crucial problem. TL approaches have been applied in situations where there is not enough labeled instances from the target task available and create an accurate model and reduce the cost. For example, Figure 3.2 shows the difference between traditional and transfer learning techniques. Traditional learning process tries to learn each task from scratch, in contrast TL tries to transfer knowledge from currently built knowledge to a target task.

We provide the relationship between traditional machine learning and current transfer learning settings, such as *Inductive transfer learning*, *Transductive transfer learning*, and *Unsupervised transfer learning.* All three methods requires understanding of TL presented in previous Chapter 2.

[1] http://socrates.acadiau.ca/courses/comp/dsilver/NIPS95 LTL/transfer.workshop.1995.html

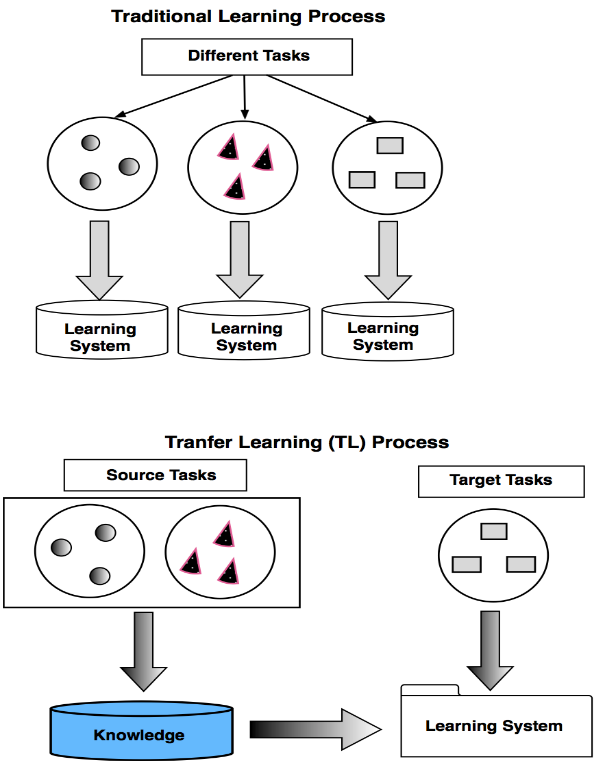


Figure 3.2: Traditional machine learning and transfer learning. Second figure presents the TL process which aim at extracting the knowledge from one or more sources tasks and applies that knowledge gained ot a target task.

– **Inductive transfer learning:** In this setting, there is a difference between the targeted task and the source task, regardless of whether the domain is the same or not. Labeled data in the target domain are required to induce an objective predictive model $f_t(\cdot)$ for use in the target domain. There are two categories of an inductive transfer learning setting:

1. In situations where a large number of labeled data in the source domain are available, the inductive TL is similar to the multi-task learning setting (Caruana, 1998). Nevertheless, inductive TL aims at achieving better performance in the target task by transferring knowledge from the source task, on the other hand multi-task learning tries to learn target and source task simultaneously.
2. Second situation is where no labeled data in the source domain are available. The inductive TL learning is similar to the Self-learning method proposed in (Raina et al., 2007). Using this method, the label spaces between the source and target domains may be different, however, the information of source domain cannot be used directly. Thus, it is relevant to the inductive TL setting where the labeled data in the source

domain are unavailable.

Inductive Transfer with Scarce Data

This setting can be also viewed as a way to offset difficulties posed by tasks that involve semi-supervised learning. In scarce data, if there are small amounts of class labels for a task, treating it as a target task and performing inductive TL setting from a source task could lead in building accurate models. These methods aim at boosting a target task from the source task, even though the both datasets are assumed to come from different probability distributions.

Research work in (Dai et al., 2007b) has investigated Bayesian transfer methods to address scarce data of a target task data. The advantage of using Bayesian TL method is the stability that a prior distribution can afford in the absence of large datasets. Evaluating a prior from related source tasks, Bayesian TL methods prevent the over-fitting that would tend to occur with limited data. Dai et al., 2007a demonstrated TL in a boosting algorithm using large number of datasets from a previous learned task to supplement small amount of dataset. Boosting is another approach for learning several weak classifiers and combining them to build a stronger classifier (Freund and Schapire, 1995). Authors weight source task data according to their similarity to the target task data. This method allows classifiers to leverage source task data that is relevant to the target task while paying less attention to data that appears less relevant.

TL in unsupervised and semi-supervised learning setting is proposed in (Shi et al., 2008). Authors assume that a reasonably sized dataset exists in the target task, however, there are large amounts of unlabeled data due to the cost of having an expert assigning labels. They proposed using an active learning approach to address this problem, where the target learner requests labels for data only when necessary. The classifiers are built with labeled data, including source task and estimate the confidence with which these classifiers can label unknown instances. In cases where confidence is too low, they suggest requesting an expert for labeling.

– Transductive transfer:

In the transductive TL setting, source and target tasks are required to be the same, while source and target domains are different. There are no labeled data available in the target domain, however, there are a lot of labeled data available in the source domain. In addition, according to different situations between source and target domains, we can further categorize the transductive TL setting:

1. Where feature spaces between source and target domains are different, $\mathrm{X}_S \neq \mathrm{X}_T$.

2. Where feature spaces between domains are relevant, $X_S = X_T$, however, marginal probability distributions of the input data are different, $P(X_S) \neq P(X_T)$.

– **Unsupervised transfer:**

This setting is similar to Inductive TL, however, in unsupervised TL the target task is different from but related to the source task. Nevertheless, unsupervised TL focus on solving unsupervised learning tasks in the target domain, such as clustering, dimensionality reduction and density estimation (Dai et al., 2008a). These methods are more common in situations where no labeled data are available, similar to source as well in target domain in training.

3.2.1 Research issues of transfer learning

There are several research issues of TL that have gained interest from the machine learning community. We summarize them as follows,

◆ **TL from multiple source domains:**
In previous chapter we have introduced our focus on one-to-one transfer where only one source domain and one target domain exist. But, in real-life settings, we may have multiple source as a task. Yang et al., 2007 proposed algorithms to a new SVM for target domains using SVMs learned from multiple source domains. In (Luo et al., 2008) proposed to train a classifier for use in the target domain by maximizing predictions agreement from multiple sources. Similarly, (Mansour et al., 2009) proposed a framework using linear weighted distribution for learning from multiple sources. The focus of this work is to estimate data distribution of each source to re-weight data from different source domains.

◆ **TL against different feature spaces:**
Another interesting issue in TL is transferring knowledge across different feature spaces. Ling et al., 2008 proposed a method for transfer learning to address the cross-language classification problem. The method aims at solving the problem where there are a large number of labeled English text data whereas there are only a small number of labeled Chinese text documents. Moreover, Dai et al., 2008b proposed a new risk minimization framework based on a language model for machine translation. These method aims at solving the problem of learning heterogeneous data that belong to different feature spaces.

◆ **TL with Active-learning:**
In Chapter 2 we have discussed the aim of TL to build an accurate model with min-

imal human supervision for a target task in order to reduce cost. Several research work have suggested combing active learning and transfer learning techniques in order to improve the learner and to build more accurate model with less human supervision. Liao et al., 2005, proposed novel active learning techniques to select unlabeled data in a target domain to be labeled with the help of the source domain data. Similarly, Shi et al., 2008 proposed using active learning algorithms to select important instances for transfer learning with TrAdaBoost (Dai et al., 2007a) and standard SVM. In (Harpale and Yang, 2010) proposed an active learning framework for the multi-task adaptive filtering problem to explore various active learning approaches to the multi-task adaptive filter to improve the performance.

◆ **TL for new tasks:**
Despite their popularity of TL in classification, clustering, regression tasks, they have been also proposed for other tasks, such as metric learning (Zha et al., 2009), structure learning (Honorio and Samaras, 2010), and online learning (Zhao and Hoi, 2010). Zha et al., 2009 proposed learning a new distance metric in a target domain by leveraging pre-learned distance metric from auxiliary domains. In (Honorio and Samaras, 2010), propose a multi-task learning method to learn structures across $MultipleGaussian$ graphical models simultaneously. In the same line, Zhao and Hoi, 2010 investigated a framework to transfer knowledge from a source domain to an online learning task in a target domain.

3.3 Latent variables and scarce data

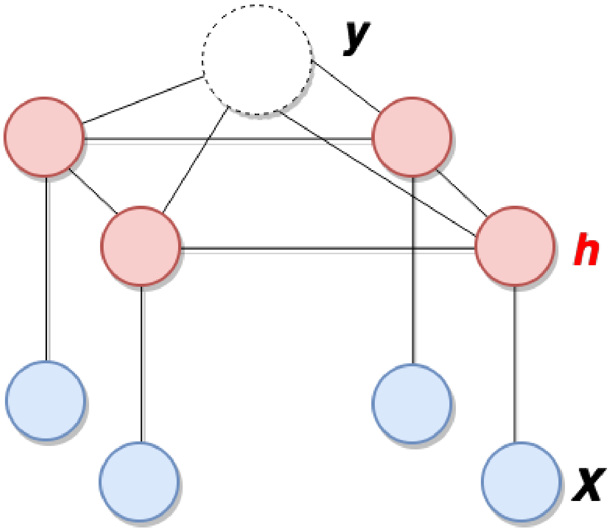
In machine learning field, latent variable models provide classic formulation for several applications.

Definition: *Let D = (x_i, y_i),...,(x_n, y_n) denote the training data, where $x_i \in \chi$ are observed variables (input variables) for the i^{th} instance and $y_i \in \Upsilon$ are the unobserved variables (output variables) whose values are known during training. In addition, latent variables models, denoted by $h_i \in H$. For example, in image processing techniques, we may have a bird images $'x'$ from which we wish to learn a type of bird $'y'$. However, the location of the bird may be unknown and can be modeled as latent variables $'h'$ (as shown in Figure 3.3). Similarly, in healthcare, learning to diagnose a disease based on symptoms or other health signs which can be improved by treating unknown diseases as latent variables. These learning parameters of a latent variable model often requires solving a non-convex optimization problem.*

A learning algorithm proceeds by iterating in two stages, first stage the hidden variables are imputed to obtain an estimate of the objective function that only depends on

w. Second stage includes an estimation of the objective function to obtain a new set of parameters. EM algorithm (Dempster et al., 1977) is one of the most popular learning method for estimation in latent variable models.

Figure 3.3: An example of latent variable model, where x is input variables, y is output variables, and h is hidden variables.



EM Algorithm for Likelihood Maximization:

The objective of this method is to maximize the likelihood (as shown in Equation 3.2):

$$\max_{w} \sum_{i} logPr(x_i, y_i; w) = \max_{w} \left(\sum_{i} logPr(x_i, y_i, h_i; w) - \sum_{i} logPr(h_i|x_i, y_i; w) \right) \quad (3.2)$$

The task for this approach is to use the EM algorithm (Dempster et al., 1977). The EM algorithm for Likelihood Maximization is presented in Algorithm 6, where EM iterates between finding the expected value of the latent variables h and maximising objective in Equation 3.2.

Algorithm 6 EM algorithm for parameter estimation by likelihood maximization.

```
Input $D$=($x_1$, $y_1$, ... , $x_n$, $y_n$), $w_0$, $\epsilon$.
$t \leftarrow 0$
repeat
    Acquire 3.2 under the distribution Pr($h_i$ | $x_i$,$y_i$;$w_t$)
    Update $w_t$+$_1$by maximizing the expectation of objective 3.2,
    where $w_t$+$_1$ = argmax$_w$ $\sum_i$ Pr($h_i$ | $x_i$, $y_i$; $w_t$) logPr($x_i$, $y_i$, $h_i$; w)
    t $\leftarrow$ t + 1
until Objective function cannot be increased above tolerance $\epsilon$.
```

3.4 Chapter Summary

In this chapter, we reviewed recent works that are related to this thesis. We presented the most important related works and compared them by their type of learning, including theoretical guarantees provided and their complexity.

A summary of the limitations found in the state of the art is the following:

◆ Approaches that can be used only for scarce data (Raina et al., 2007; Triguero et al., 2015).

◆ Approaches that are computationally intractable for large scale problems (Raina et al., 2007; Rokach, 2010; Yu and Joachims, 2009; Zhou and Xu, 2007).

◆ Approaches that assume to address scarce data problem (Blum and Mitchell, 1998; Raina et al., 2007; Xiang et al., 2013).

In the next chapters, we present our contributions in addressing scarce data. We start by presenting frameworks used to collect data from subjects that participated in the studies and the features selected for this research work. Then, challenges to address scarce data are presented. We conclude the proposed approach named as Intermediate Models to improve classifiers precision.

Chapter 4

DATA COLLECTION AND ANALYSIS

"We should have lifelong monitoring of our vital signs that predict things like skin or pancreatic cancer so we can eradicate it. We should have personalized medicine; there's a huge amount of innovation possible."

– **Sebastian Thrun**

In this chapter, we provide an overview of the monitoring systems, study setup, and initial data analysis. We begin providing an overview of the trial setup and participants demographics. Then, we provide description of features extracted from the data collected from both systems. We demonstrate the problems that occur in monitoring individuals in long-term using smartphone sensing capabilities. Further, we select the appropriate types of sensors for inferring behaviour changes with respect to users privacy, dealing with scarce data, and the common issues faced using our datasets. Finally, we will close the chapter with our proposed approaches for addressing limitations of scarce data and novel intermediate models proposed to improve the performance of supervised classifiers.

The main contributions of this chapter are as follows:

A.1 *Introductions of the trials and the number of sensory data collected from participants*
A.2 *Methods used to extract features from each type of sensor data acquired*
A.3 *We evaluated the data mining approaches used for this research*
A.4 *Finally, we provide our initial picture of the data and results from data analysis*

The outline of this chapter is as following: the Section 4.1 provides a brief introduction of monitoring system using smartphone sensing modalities. In the Section 4.2 we provide a brief introduction of monitoring system used in bipolar disorder patients, data acquired from the patients in situ, data sources selected for our research, features extracted, and *the initial result from data analysis. Similar, in Section 4.3 we provide details of data collected and analysed. Finally, we provide an overview of the Stress@Work assessments items used to assess employees perceived stress at working environments.*

4.1 Brief introduction of monitoring systems using smartphones

Due to the rapid development of information technologies in healthcare domain, data collection have been shown to play a significant role in improving disease-related knowledge. The new generation of smartphone devices with embedded sensors has created opportunities for exploring new context-aware services and this kind of data can be useful. Despite the advances of sensing systems, there are several challenges that must be targeted to overcome. These challenges revolve around scarcity of data, and missing labeled measurements that limit the systems to have an accurate classification of their users.

The problem of collecting large-scale training data is a common problem. Continuous inference of human behaviour using sensory data measurementsand self-assessments scales (*e.g.*, wellbeing, psychological state) from individuals is itself relatively simple from a technical point of view. However, in practice collecting large sample of data from individuals as they go in their real-life activities requires a lot of effort. Current systems, still suffer from both practical limitations and a number of technological shortcomings, for instance, battery drain causes a significant problem in data collection, the application crashes, the application hung due to system memory, and others.

In order to have an accurate self-care health monitoring system, participants are requested to provide reliable training data that are valuable information for classification accuracy. This provides a clear evidence that obtaining efficient learning model is a crucial issue when it comes to human monitoring. However, in uncontrolled settings labeling data is not nearly as easy due to the time and effort for individuals to manually provide labeled data. This problem is even more expressed when it comes in monitoring mental disorder or even the individuals perceiving stress due to their condition. Under this scenario labels are sometimes unreliable, however, the information provided contain valuable information for classification.

The main problems in real-world scenarios for self-monitoring systems can be summarized as follows:

- ◆ Most of the existing systems are built under the supervised setting where labeled data are crucial for training the model.
- ◆ Having sufficient labeled instances require more effort and it is time consuming.
- ◆ These systems suffer from its dependence on the accuracy of the users labeled data.

- Self-monitoring requires the active involvement and motivation of users (*i.e.*, reminders, feedback) which sometimes may lack.
- Most of the systems do not use the unlabeled instances, however, these instances can also give important information.

4.2 Monitoring systems used in bipolar disorder patients

MONARCA (MONitoring, treAtment and pRediCtion of bipolAr disorder episodes) is an EU project from the FP7 framework program[2]. The main goal of the project was to develop and validate solutions for multi-parametric, long term monitoring of behavioural and physiological information relevant to bipolar disorder. The system consisted of 5 components: smartphone, a wrist worn activity monitor, a novel sock integrated physiological (GSR, pulse) sensor, a stationary EEG system for periodic measurements, and a home gateway. In order to successfully accomplish the goals of the project, there were 2 hospitals and 7 technical universities involved, and 3 companies responsible for the business model and the integration of the final system into the existing clinical work-flows. At CREATE-NET[3], we focused on the analysis of the smartphone data gathered during the trials in one of the hospitals.

4.2.1 Trial setup in bipolar disorder monitoring

The study group consisted of 10 patients (9 female and 1 male). As inclusion criteria, each of the patients had to be diagnosed with bipolar disorder (with frequent changes of episodes), age between 18 and 65, ability and are willing to operate modern smartphone devices. The patients were categorized by the ICD-10, F31 classification (by the International Classification of Disease and Related Health Problems) and were selected from the ward's psychiatrists that are capable of dealing with the requirement of the study.

The trial was uncontrolled, not randomized, mono-centric, prolective, observational study. Each patient was given a personal smartphone to use in any way they wanted. There were no constraints of any kind placed upon the patients, with respect to holding the phone in a specific manner or at a specific place in the body or otherwise. The phone had the continuous sensing application (developed in German Research Center for Artificial Intelligence (DFKI) [4] installed that recorded data on the phone memory and transmitted the data periodically to a dedicated server. All sensing modalities were sampled, including microphone, accelerometer, GPS, WiFi access points, Bluetooth, SMS, phone calls and

[2] http://www.monarca-project.eu

[3] http://www.create-net.org/projects/4/1026/MONARCA

[4] https://www.dfki.de/web/intelligent-solutions-for-the-knowledge-society

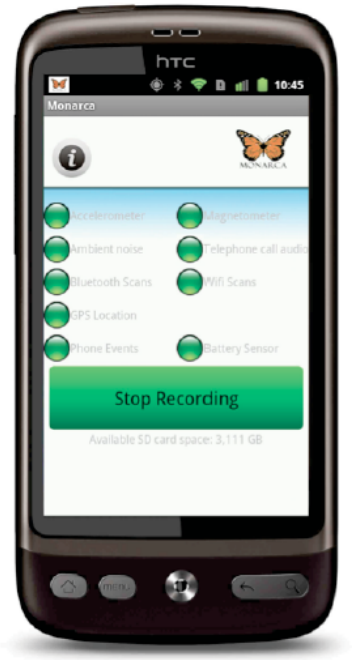


Figure 4.1: Patient monitoring application in bipolar disorder.

their duration. The application, shown in Figure 4.1 ran continuously in the background, sampling these sensors and was set to start automatically on phone start up.

Patient monitoring application was designed to measure two aspects, namely patients' internal affective states, through the use of questionnaires; and, objective behaviour, through sampling of phone sensors. The application has been developed in close cooperation with the psychiatrics in order to capture relevant aspects of the disease. In order to increase patients' motivation to provide daily experience sampling, the application provides alarms and reminders to fill out the questionnaire at a predefined time in the evening. Through the questionnaires the patients were able to provide their current state as well as activities they performed during the day, estimate their sleeping hours as well as quality, time spent outdoors and their social-activities.

4.2.2 Patient psychiatric evaluation

Psychiatric assessment and the psychological state examination were performed every 3 weeks over a period of 12-weeks at the psychiatric hospital Hall in Austria (TILAK - Department of Psychiatric, State Hospital, Hall in Tyrol, Innsbruck). The psychiatrists have set the interviews for the patients in such a way to reduce memory effect, which prevents having biased evaluation outcomes. To improve the scarcity of ground truth, between scheduled interviews well trained and experienced clinicians talked collaboratively with patients about treatment by phone. During the examination, four standardized scales were used from clinical psychologists.

The clinicians used the following standard scales during the assessment of the patients:

- **Hamilton Depression Scale (HAMD)**: HAMD scale has been applied to rate the severity of depression in patients through assessment of a range of symptoms. The higher the magnitude of symptoms, the higher is the scale of severity of depression (cut-off value: $\geq$8)

- **Young Mania Rating Scale (YRMS)**: YRMS is most frequently utilized rating scale to assess manic symptoms. The baseline scores can differ in general, depending on the patients' clinical features such as depression (YMRS=3) and for mania (YMRS=12).

In order to evaluate the patients, the HAMD and YRMS scores were normalized in a scale of -3 to +3 where the former indicates Severe Depression and the latter indicated Severe Mania, with intermediate steps of depressed, slightly depressed, normal, slightly manic and manic.

Table 4.1: Psychiatric Evaluation (PE) Scores during the trial.

P.ID	**1stPE**	**2ndPE**	**3rdPE**	**4thPE**	**5thPE**
P0101	+2	+1	+1	+0.5	**0**
P0201	**-1**	**0**	**0**	**0**	**-3**
P0302	**-3**	**-2**	**0**	**0**	**0**
P0402	**-3**	–	**-3**	–	**-3**
P0502	**0**	–	**-3**	**-2**	0
P0602	**-0.5**	**-1**	**0**	**quitted Trials**	–
P0702	**-2.5**	**-2**	**-0.5**	**-2**	**-2**
P0802	**-3**	**-1**	0	–	0
P0902	**0**	**0**	**-2**	**-1**	**-2**
P1002	+1	**-1**	**-2**	**-2.5**	–

Psychiatric evaluation scores of the patients are shown in Table 4.1 using the normalized scale.

None of the patients had rapid relapses where their state did not change within a few days but at least one or more weeks. According to the professional psychiatrist, it was acceptable to set the ground truth assessment values 7 days before the examination and 7 days after the examination and it was adjusted (extended or shortened) according to stable or unstable daily subjective self-assessments.

4.2.3 Completeness of study

There were five measurements points for 10 bipolar disorder patients. A patient (P0402) did not show any changes in their episodic state during entire trial. As such, their data was of no use in respect to classification state and are discarded. Further, the patient (P0602) drop out of a clinical trial due to the condition faced at that period. In our studies, we have analysed the data during phone-conversation, however, patients (P0101, P0802) did not use the smartphones for phone-conversation. Furthermore, patient P0502 did not have sufficient phone-conversations to be used in the classification model.Therefore, only 5 patients (P0201, P0302, P0602, P0902, P1002) provided sufficient data point for and different classes due to their relapses ($i.e$, experiencing more than two episodic changes) to make our studies possible for classification to their state.

4.2.4 Scarce data and missing information in monitoring system

A number of challenges plagued the trial, most prominent of which was patient compliance. Considering that the trial was conducted under uncontrolled conditions, during normal daily life of the patients, it was impossible to ensure that the patients always carried the phone with them. In addition, many practical challenges have been faced. Some patients switched-off sensing application at certain occasions or forget charging the smartphone over night, creating gaps in available data. This increased significantly a number of missing data for entire recording period.

As discussed in previous section, ground truth was available every three weeks and increased a number of unlabeled instances between psychiatric evaluations. The inability of supervised learning approaches to endure with unlabeled training data reduced even more the number of available days. The actual amount of sensor data available lies between 19 and 71 datasets per patient per sensor modality. Fortunately, self-reported mood were assessed on a daily basis on the smartphone, allowing us to draw the knowledge upon patient's behaviour and self-reported mood to extend the ground truth periods.

We believe that proposed approaches can be effective in overcoming many of the obstacles to smartphone sensing. The following chapter of this thesis prove the strength of using semi-supervised learning and intermediate models, to overcome the challenges on handling scarce information.

4.2.5 Quantifying physical activity in bipolar disorder patients

In order to quantify the level of activity we use accelerometer sensor data acquired from the smart phone. We have captured 3-axial linear acceleration continuously at a rate of 4Hz to 10Hz, which varied due to Android system operating conditions, such as system load and battery levels. However, this sampling rate was sufficient to infer physical activity levels of patients. The accelerometer signals were re-sampled at fixed rate of 5 Hz (25.6 second). For each patient there was an average of 2 GB of raw accelerometer data. Physical activity levels were estimated using pre-processed accelerometer data. Acceleration magnitude (namely Signal Vector Magnitude 4.1) vector was calculated as square root of sum of squares of individual acceleration axis, which allowed calculation of physical activity levels to be invariant to phone orientation, which due to unconstrained nature of the trial, phone orientation is unknown. The variance of the magnitude (as shown in Equation 4.17) on each *n=128* samples provided an activity score, which was set within a threshold of three states, namely 'none', 'moderate' and 'high' activity as detailed in FUNF framework (Aharony et al., 2011).

$$\boldsymbol{SVM} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{x_i^2 + y_i^2 + z_i^2} \tag{4.1}$$

$$\boldsymbol{varSum}\textit{(n)} = ((SVM(n) - avgSVM(n))^2 - (\frac{n}{n-1}) - 2SVM(n) \tag{4.2}$$

For this research we were interested in change of overall activity levels, therefore we have combined the two active states (*'moderate'* and *'high'*) to produce a single score. In the sections that follow, we provide results of our initial analysis of overall activity levels and also the results of intervals, where monitored days were divided in daily intervals. It is important to note that for this analysis, we have excluded the days in which the patient went to the clinic for the psychiatric evaluation. This is because during the assessment there would be physical activity recorded, which may not correspond with the natural behaviour of the patient and thus would have biased our results.

4.2.6 Classifying episodic states of the patients with bipolar disorder

As discussed in previous section there were no constraints of any kind placed upon the patients, with respect to holding the phone in a specific manner or at a specific place in the body. Considering the fact that the trials were conducted under uncontrolled conditions in real life activities, in this research we focus on analysing accelerometer raw data and the speech features extracted from microphone during the phone conversation, when we are almost sure that the patients are holding their smartphone. We believe that both sensing techniques have their own advantages, complement each other, and can provide adequate information for classifying the course of mood episodes or relapse of a patient. In our experiments, we also included information from the self-assessment questionnaires relevant to motor activity, such as self-reported psychological state, physical state, and activity level.

We analysed the information collected and selected those patients with enough data recorded during their phone conversation and who represent different severities of disease on their psychiatric evaluation scores.

Table 4.2: Number of calls and class associated to them based on psychiatric evaluations. There is also additional data (last column) where there is no class associated.

Patient	**Severe Depression**	**Moderate Depression**	**Mild Depression**	**Normal**	**Mild Manic**	**Total**	**Additional Data**
P0201	36	–	113	149	–	**298**	**435**
P0302	135	–	–	99	–	**234**	**199**
P0702	–	112	39	–	–	**151**	**116**
P0902	–	142	–	161	–	**303**	**178**
P1002	–	–	35	–	162	**197**	**28**
All	**171**	**254**	**187**	**409**	**162**	**1183**	**956**

Table 4.2 shows the number of class and class associated to them based on psychiatric evaluations. It can be seen that we have a different number of calls per patient and per episode. The table also shows additional data (last column) indicating the numbers of phone calls that we have that are not associated to any episode as they were performed outside the 7 days window of the psychiatric assessments.

4.2.7 Feature selection

Feature selection from smartphone sensory data is probably the most important factor to consider in order to improve the recognition performance of machine learning tools. In the following subsections, we describe the most representative techniques for extracting time and frequency domain features from accelerometer raw data and prosodic and energy features extracted from speech.

4.2.8 Accelerometer signal features in Time-Domain (TD)

In order to quantify motor activities from the smartphone, acceleration readings collected during conversation (including picking up the phone, starting and finishing the call, and replacing the phone into the holder) were used in our analyses. These periods during conversation determine meaningful changes of acceleration values. We captured 3-axial linear acceleration continuously at rates, which varied due to Android system operating conditions, such as system load and battery levels. In this research, the accelerometer signals were re-sampled at a fixed rate of 5 Hz. The accelerometer features proposed in this research, shown in Table 4.3, are quite popular amongst practitioners in the field, and were used as the basis for identifying periods of activity. To reduce the effect of spikes and noise from the accelerometer signal, statistical metrics such as mean 4.3a), variance 4.3b),

Table 4.3: Features selected for the accelerometer sensor signals.

Time Domain	**Frequency Domain**
(1) Magnitude	(1) FFT Energy
(2) Signal magnitude area	(2) FFT Mean Energy
(3) Root-Mean-Square (RMS)	(3) FFT Std.Dev Energy
(4) Variance Sum	(4) Peak Power
(5) Curve Length	(5) Peak DFT Bin
(6) Non Linear Energy	(6) Peak Magnitude
(7-14) For the 3 axes:	(7) Entropy
Variance, Mean, Max, Min,	(8) DFT
Std. Dev., Absolute,	(9) Freq.Dom. Entropy
Median, and Range	(10)Freq.Dom. Entropy with DFT
(15-20) Mean and Std. Dev. of X, Y and Z axis.	
For all 20 features, we obtained the **Min, Max, Mean**	For all: **Min, Max, Mean**
Total: **60**	Total: **30**

and standard deviation 4.3c), where x(i) represents sum of three axis are applied over a window of approximately 26 seconds (non-overlapping fixed length windows of *N=128* samples).

$$\begin{aligned} &a)\ \mu = \frac{1}{N}\sum_{i=1}^{n} x(i) \\ &b)\ \sigma^{\mathbf{2}} = \frac{\sum(x_i - \bar{x})^2}{N-1} \\ &c)\ \sigma = \frac{1}{N-1}\sum_{i=1}^{N}(x(i) - \mu)^2 \end{aligned} \tag{4.3}$$

Other features included the root-mean-square (RMS) acceleration for the period of conversation, as an indication of the time-averaged power in the signal. The RMS of a signal x_i, y_i and z_i represents a sequence of n=128 discrete values obtained using Equation 4.4.

$$\boldsymbol{RMS} = \sqrt{\frac{x_2^2 + x_2^2 + x_3^2 + ... + x_2^n}{n}} \tag{4.4}$$

The RMS results demonstrate differences in the motor activity during the phone conversation. The lower the RMS value, the lower the motor response which is manifested in depressed patients, whereas patients in the manic phase show elevated levels, as shown in Figure 4.2 b).

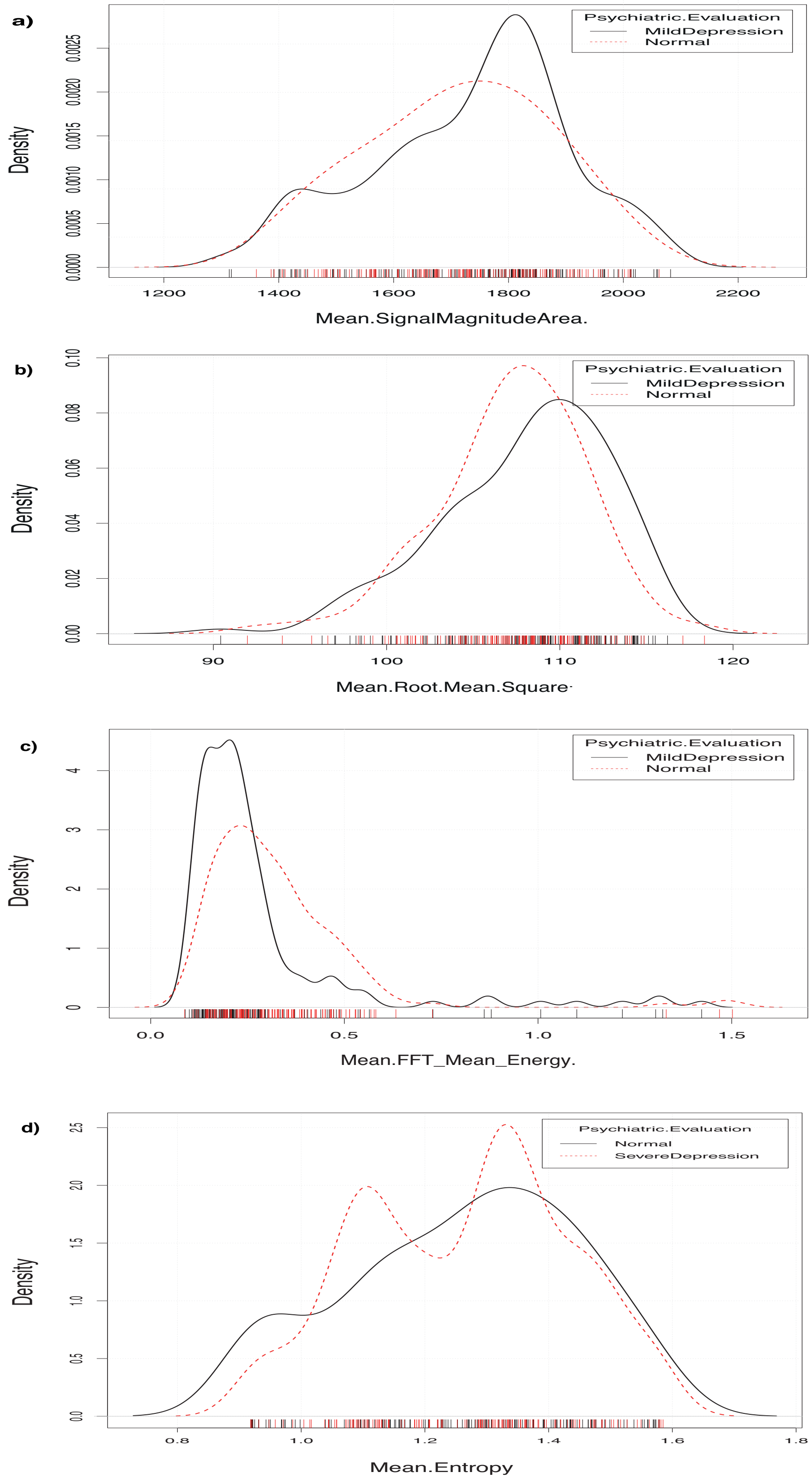


Figure 4.2: Overall mean values of a) RMS (p0201), b) SMA (p0201), c) energy (p0302) and d) entropy (p1002) with psychiatric evaluation.

Another suitable measure for phone activities is the normalized signal magnitude area (SMA) that was used as the basis for identifying periods of activity during phone conversations, where $x(t), y(t)$, and $z(t)$ are the acceleration signals from each axis with respect to time t as denoted by Equation 4.5.

$$\boldsymbol{SMA} = \frac{1}{t}\int_o^t |x(t)|dt + \int_o^t |y(t)|dt + \int_o^t |z(t)|dt \tag{4.5}$$

An example using SMA is presented in Figure 4.2 a) where changes of motor activity can be compared in two states of the disease, transition from mild depressive state to normal state. The graph includes of number of phone calls in both states (n=140).

Also, a feature like Signal Vector Magnitude (SVM) (Jeong et al., 2007) has been used to measure the degree of activity intensity and velocity of phone movement during the phone conversation and was obtained using Equation 4.6. In addition to SVM, we computed the Variance Sum (Aharony et al., 2011), that using the equation shown 4.7, where n represents the window size and $avgSVM$ the mean of the SVM of that window size:

$$\boldsymbol{SVM} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{x_i^2 + y_i^2 + z_i^2} \tag{4.6}$$

$$\boldsymbol{varSum}(n) = ((SVM(n) - avgSVM(n))^2 - (\frac{n}{n-1}) - 2SVM(n) \tag{4.7}$$

Furthermore, in order to capture abrupt changes of phone activity during the phone conversation we used Averaged Non-linear Energy feature and Curve Length (CL) (Mukhopadhyay and Ray, 1998) feature using Equations 4.8 and 4.9.

$$\boldsymbol{CurveLength} = \sum_{i=1}^{n} |x_{i-1} - x_i| \tag{4.8}$$

$$NonE_i = x_i^2 - x_{(i-1)}x_{(i+1)};\ avgNLE = \sum_{i=2}^{n-1}\frac{NonE_i}{n-2} \tag{4.9}$$

4.2.9 Accelerometer signal features in Frequency-Domain (FD)

The signal and the distribution of signal energy over the frequency-domain are also popular choices in signal analysis. In this research, we used frequency-domain techniques to capture the repetitive nature of an accelerometer signal. These repetitions are often correlated to motor activity changes, which are capable of capturing distinctive pattern of movements in bipolar disorder patients during phone conversations. We applied the Fast

Fourier Transform (FFT) on acceleration segments. Similarly as in TD, we used time window of approximately 26 seconds (non-overlapping fixed length windows of n=128 samples), which enabled fast computation of FFT's that produces 128 components for each 128-sample window. Since our goal is to investigate the activity signatures, energy features were used to assess the strength of motor acts. The features in frequency-domain that are given in Table 4.3 have been used to determine the intensity of the signal. Total Energy of the acceleration signal was calculated as the squared sum of its spectral coefficients (sum of the squared discrete FFT component magnitudes of the signal) normalized by the length of the window. Using this metric, we were able to capture the intensity of the activity obtained using Equation 4.10 component magnitudes of the signal.

$$\textbf{Energy} = \sum_{j=1}^{(n/2)+1} y[j]^2 \tag{4.10}$$

Figure 4.2 c) shows an example of the total energy values of patient P0302 during phone conversations with different episode. As can be appreciated, the patient shows an increase level of motor activity in normal state compared to depressive states.

In order to determine the highest magnitude of all frequencies, frequency magnitude was measured using the real and imaginary components of the FFT values (using Equation 4.11). Frequency magnitude values below the cut-off and above the Nyquist rate (Nyquist-Rate=window-length/2) where nullified by keeping the peaks obtained in the window. Data has been normalized using Equation 4.12 and multiplied by 2 to maintain the same energy. Furthermore, feature values obtained from entropy metric were measured using the normalized information entropy of the discrete FFT coefficient magnitudes by excluding the gravitational component, so called DC component of FFT (using Equation 4.13). Figure 4.2 d) shows an example of mean entropy values for patient P1002.

$$\textbf{Magnitude} = \sqrt{FFT.real^2 + FFT.imag^2} \tag{4.11}$$

$$\textbf{Normalized} = Magnitude * 2/windowLength \tag{4.12}$$

$$\textbf{Entropy} = \sum_{j=1}^{(n/2)+1} c_j \cdot log(c_j), \text{ where } c_j = \frac{|y_i|}{energy} \tag{4.13}$$

$$\textbf{Peak}_{Freq} =_{j}^{argmax} \; |y_i| \tag{4.14}$$

$$\mathbf{Peak}_{energy} =_{j}^{max} \quad |y_i| \tag{4.15}$$

Together with the FFT Energy mean, FFT Energy standard deviation, FFT energy, DFT (Discrete Fourier Transform), and frequency magnitude, Entropy (Cover and Thomas, 2012) is helpful in discriminating activities that differ in complexity. In our research, using this feature helped us to distinguish signals that have similar energy values with different motor activity patterns. Furthermore, we also investigated the largest signal peak using Peak Power Frequency that was compared against the baseline values (Equations 4.14 and 4.15).

4.2.10 Feature selection and extraction from speech during the phone conversation

Previous work have shown scientific evidence that speech features can be used as an indicator of bipolar disorder (Moore et al., 2003; Moore et al., 2008). In this regard, speech production is one physiological function that has been reported to affect motor retardation in bipolar patients. The application developed for our research, records speech signals from microphone only during the phone conversation with a sampling rate of 44Hz and 16 bits amplitude quantization. Algorithms were developed to scrabbled/stretched the actual signal to avoid its original reconstruction while keeping the required properties for analysing the voice. In the current research work, we extracted acoustic features from the speech signal using OpenEar (Eyben et al., 2009) and Praat (Boersma, 2002). We evaluated features that have been successful in previous work (Pérez-Espinosa et al., 2012). Table 4.4 shows the acoustic features that were included in this research. We divided the features in two types: prosodic and vocal tract spectrum.

Table 4.4: Selected speech features relevant to bipolar disorder states.

Group	Feature Type
Prosodic:	
Energy, Times	LOG energy, Zero crossing rate
PoV, F_0	Probability of voicing, F_0
Spectral:	
MFCC	MFCC
MEL	MEL spectrum
SEB	Spectral energy in bands
SROP	Spectral roll of poing
SFlux	Spectral flux
SC	Spectral centroid
SpecMaxMin	Spectral max and min with DFT

The features that were extracted from the patients' speech data include the first-order functional of low-level descriptors (LLD) such as FFT-Spectrum, Mel-Spectrum, MFCC, Pitch, Energy, Spectral and LSP.39 functionals such as Extremes, Regression, Moments, percentiles, Crossings, Peaks, and Means. Prosodic features have been shown to provide rich source of information in speech such as pitch, loudness, speed, duration, pauses, and rhythm that could be used to detect the state of mind of patients during phone calls, *i.e.*, when patients are in severe depressive state to normal or from moderate depression to normal states (Moore et al., 2003).

The second types of features were spectral features, which provide accurate distinction to a speaker's voice when prosodic aspects are excluded. We included the most popular voice quality descriptors shown in Table 4.4. With these types of features, we were able to distinguish periods of speech from patients, such as duration of speech segments, number and type of pauses (*i.e.*, long, medium, and short), and overlapped or non-overlapped speech during conversations. We also measured the reaction and response time during the conversation time. We use the terms Number- and Duration of long pauses during the conversation to refer to the phone rate over the total conversation session, with times when the speech is not active (pauses) included in the total conversation session. Motivated by the clinical work carried out in studying bipolar patients in (Moore et al., 2008; Naranjo et al., 2011), we examined the association between long speech pauses in depressive patients and speech increments in manic phase during the phone conversation with their psychiatric scores, as shown in Table 4.7.

Table 4.5: Selected speech features relevant to bipolar disorder states.

Emotional Features	**Spectral Features**
(1) **Percentage of Angriness** (2) **Percentage of Nonconformity** (3) **Percentage of Happiness** (4) **Percentage of Equanimity**	(1) **Number of speech segments** (2) **Number of short pauses** (3) **Number of medium pauses** (4) **Number of long pauses** (5) **Total duration speech in call** (6) **Total duration not overlapped speech** (7) **Total duration overlapped speech** (8) **Quality of Service** (9) **Duration of medium pauses in call** (10)**Duration of long pauses in call**
Total: **4**	Total: **10**

Table 4.6 provides an overview of phone conversations during the trial. This table shows the overall number and average duration of phone conversation between the psychological evaluations in a daily basis. Since we focus on understanding meaningful information around the phone conversation, we keep accelerometer reading one-minute

before the phone conversation, the readings from the entire duration of the call, and one minute after the conversation ended. Phone calls of less that 10 seconds were discarded in our experiments.

Table 4.6: Overall number and duration of phone calls (Incoming, Outgoing) between the psychiatric assessments (Mean±SD)

Patient ID	**1st-2ndPE**	**2nd-3rdPE**	**3rd-4thPE**	**4th-5thPE**
P0201	400 (8.76±5.38)	204 (5.1±3.7)	153 (4.02±3.21)	193 (5.36±3.79)
P0302	169 (6.76±3.4)	119 (5.66±3.46)	158 (7.53±4.52)	85 (5.31±3.33)
P0702	121 (6.1±4.33)	50 (5.0±3.01)	125 (7.73±5.47)	119 (6.4±4.92)
P0902	172 (10.06±7.09)	108 (8.71±5.76)	185 (5.44±4.85)	–
P1002	130 (13.16±8.01)	216 (11.36±12.6)	–	–

Table 4.7: Relationship between duration and number of long pauses in phone calls and psychiatric assessment scores (*n/a - not applicable, since the patient did not experience a second depressive episode).

P.ID.	**Avg. Duration / Avg. Long Pauses (Score)**	**Avg. Duration / Avg. Long Pauses (Score)**	**Difference (%)**	**Avg. Duration / Avg. Long Pauses (Score)**	**Difference (%)**
P0201	57.56/ 0.52 (MiD)	39.77/ 0.28 (N)	**-30.90/ -46.15**	74.74/ 0.57 (SeD)	**87.93/ 103.57**
P0302	130.86/ 1.15 (SeD)	87.95/ 0.87 (N)	**-32.79/ -24.34**	n/a	n/a
P0702	54.19/ 0.53 (MoD)	143.66/ 1.32 (N)	**165.10/ 149.05**	119.17/ 0.98 (MoD)	**-17.04/ -25.75**
P0902	95.64/ 0.75 (N)	130.86/ 1.05 (SeD)	**26.91/ 28.57**	n/a	n/a
P1002	85.96/ 0.62 (MiM)	222.97/ 1.51 (MiD)	**73.27/ 143.54**	n/a	n/a

– **(MiD)**=Mild Depression;
– **(N)**=Normal;
– **(SeD)**= Severe Depression;
– **(MoD)**=Moderate Depression;
– **(MiM)**=Mild Manic.

As can be seen from Table 4.7, average pauses and response delays in depressive state were inserted, in general, more often than during non-depressive state. This decrease can be seen across patients P0201, P0302, and P0902. In patients P0201 and P0302 it is more noticeable, where the average of decrease of phone call duration/average number of long pauses between the words went from 57.56(sec.)/0.52 during a depressive state to 39.77(sec.)/0.28 during a normal state (P0201); and patient P0302 where the average decrease of phone call duration and number of long pauses went from 130.86(sec.)/1.15 during a depressive state to 87.95(sec.)/0.87 during a normal state. In Figure 4.3 c) and Figure 4.4 a) we present the distribution of overall speech segments in conversation by mood episode of the patients. The speaking rate is significantly reduced during depressive periods as well as the duration of continuous speech segments.

In contrast to patients P0201 and P0302, where the transition of their state was from depression to normal phase, patient P0902 had a noticeable decrease number of long

Figure 4.3: Overall mean values of a) number of long pauses (p1002), b) duration of long pauses (p1002), c) number of speech segments (p0702).

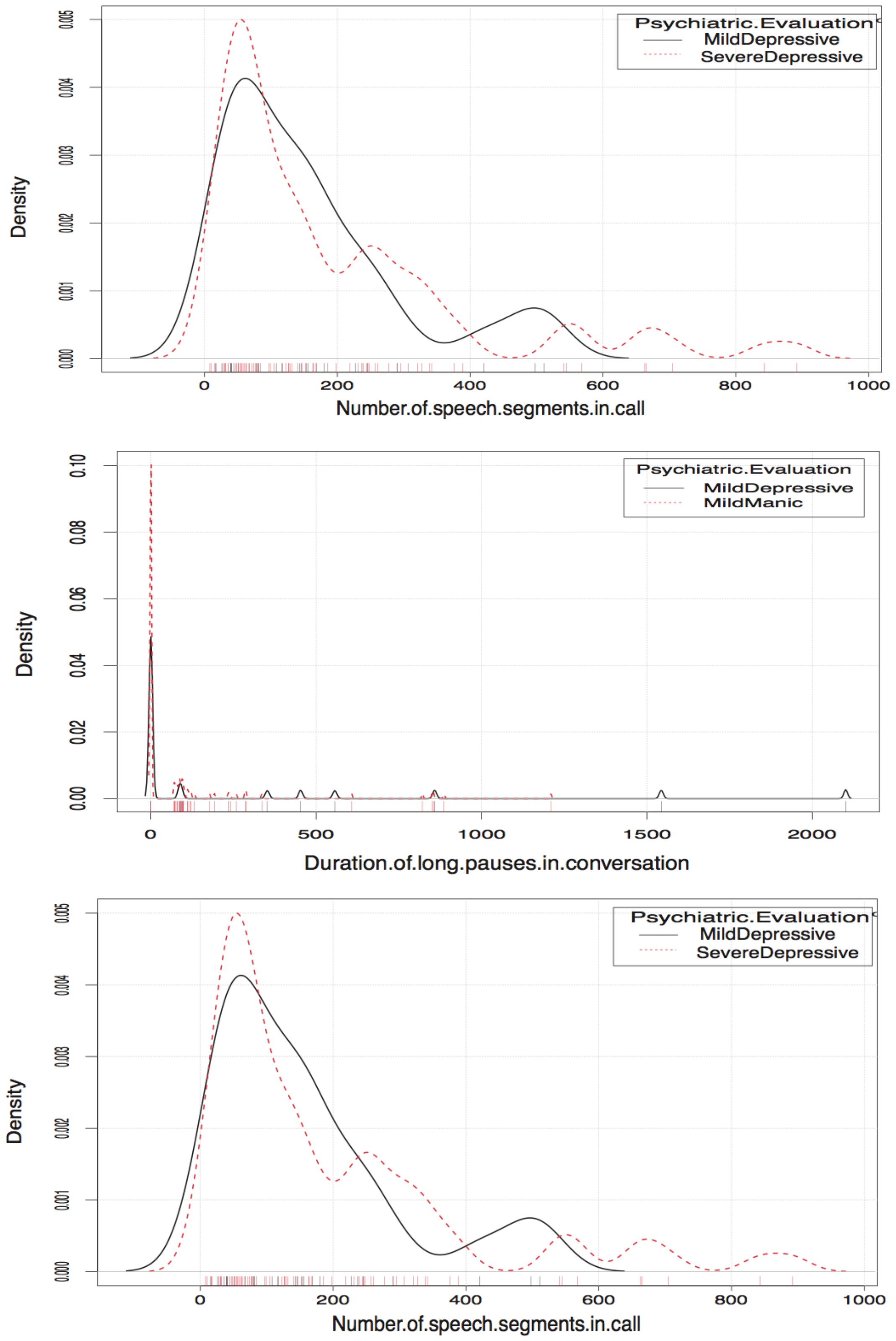
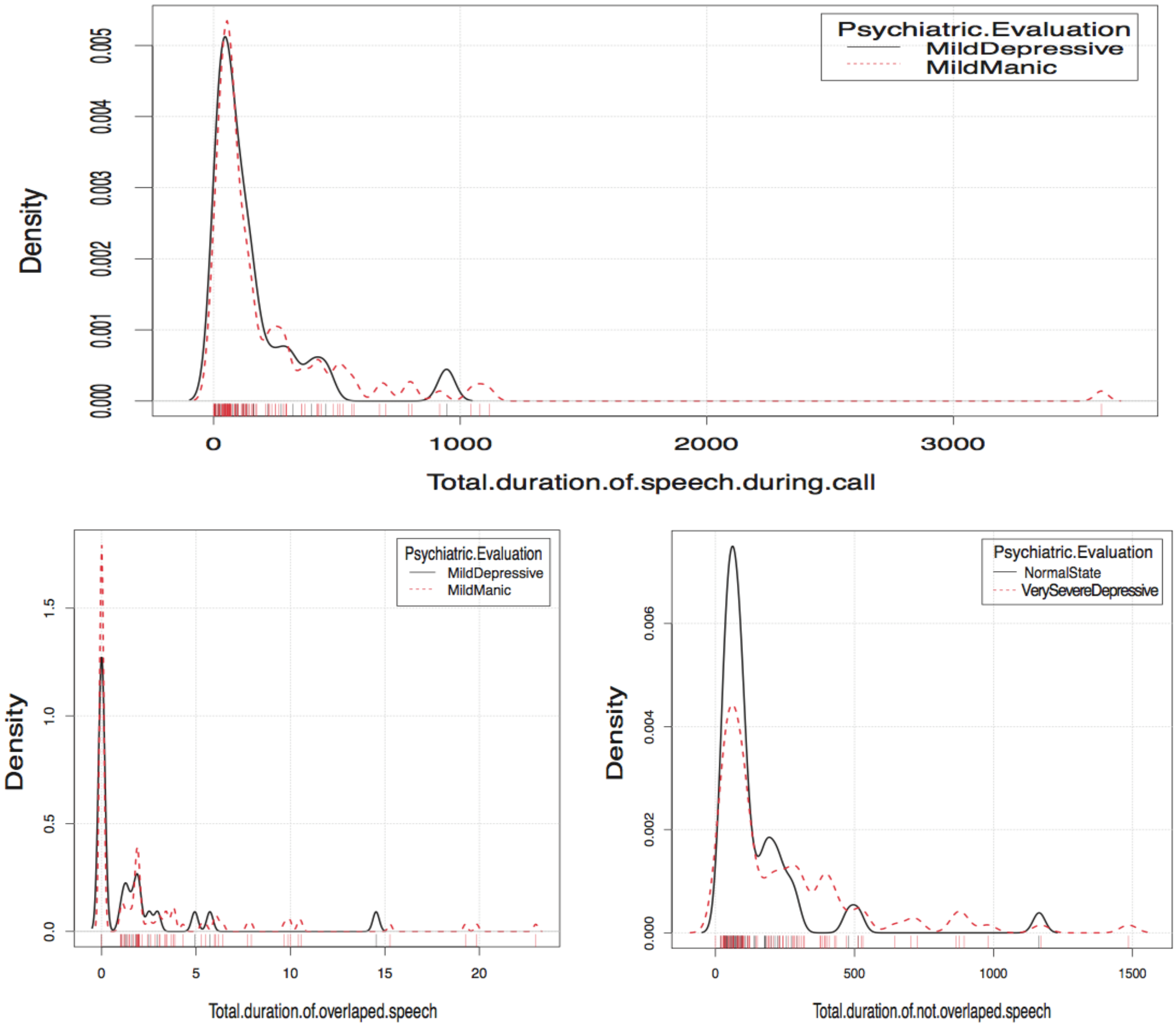


Figure 4.4: Overall mean values of a) total duration of speech segments (p0201), b) duration of overlapped speech (p0302) -on the left, c) duration of not overlapped speech with psychiatric evaluation (P1002)- on the right.



pauses during the phone calls as he went from a normal state to a depressive state. As such, there was a 26.91%/28.57% increase average duration of phone call duration and number of long pauses due to the transition to a depressive episode.

For the patient that experienced a manic episode, P1002 we can see a reverse trend, where the patient had decreased his average of long pauses, in accordance with the study reported in (Vanello et al., 2012). Average duration and number of long pauses were increased to 73.27%/143.54% during the depressive episode. Figure 4.3 a) and Figure 4.3 b) provide the proportion of number/duration of long pauses between transitions from a manic episode to a depressive episode (P1002). We also studied speech overlapping, voice quality and emotional features during phone conversations. Voice quality measures active speech frames, which were determined according to an energy-based speech activity. We explored the regularity and the responses from both active speakers during a phone conversation. Speech-overlapping was used to see the regularity during the conversation. Figure 4.4 b) and 4.4 c) present a comparison between non-overlapped in depression (P0302) and overlapped speech from patients in manic episodes (P1002).

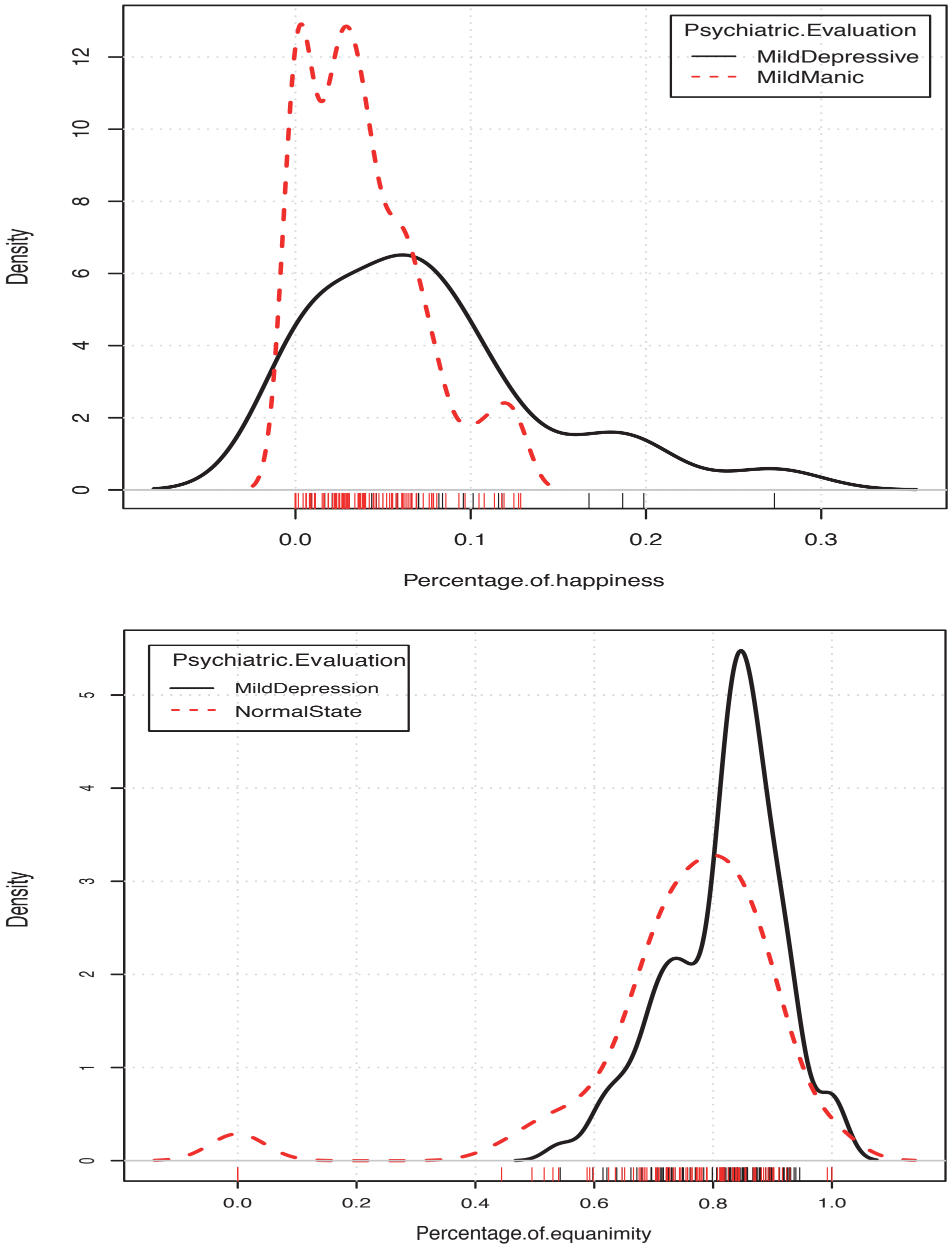


Figure 4.5: Distribution of percentage of: a) happiness (p1002), and b) equanimity (p0201) features by psychiatric evaluation.

Table 4.8: Number of features used in the experiments.

Feature	**Number**
Accelerometers:	
1) Time-Domain	60
2) Frequency-Domain	30
Audio:	
1) Emotional	4
2) Spectral	10
Questionnaire	3

The effects of emotional expression on speech are an interesting feature in bipolar disorder. Emotional state has been reported in previous studies, by identifying changes in muscle tension and in breathing. In our previous work (Pérez-Espinosa et al., 2012), we have explored emotional state features from speech (*i.e.*, happiness, angriness, nonconformity, and equanimity). In clinical reports that have investigated the symptoms in a manic episode, such as in (Vanello et al., 2012), patients were characterized by extreme happiness and hyperactivity. Similarly, in our research we found a different percentage of happiness extracted from speech in manic episodes, while in a depressive state we found lower percentage of happiness, as shown in Figure 4.5 a). Equanimity feature has also shown lower percentage in mild depression, whereas in normal state we found lower percentage of equanimity during the phone conversation (as shown in Figure 4.5 b)).

For our experiments we also used information from the questionnaires in terms of three attributes: (1) Physical, (2) Activity, and (3) Psychological condition, whose values range from 1=low to 5=high. A summary of all the attributes used in the experiments is given in Table 4.8. In the experiments we tested different sets of these attributes.

4.3 Monitoring Stress@Work

TurnOut-Burnout is a project for monitoring Stress@Work funded from EIT ICT Labs. The main goal of the project was to use unobtrusive technologies for monitoring (*i.e.*, smartphones) behavioural information and detect burnout in the early phases of the so-called burnout cascade.[4] The system consists of a smartphone and web-server to visualize their daily behaviour patterns. Acquiring data from the employees' life are used to generate recommendations for people at risk of getting a burnout. The aim of the project ws to create prototype services for early burnout recognition as well as recommendation

[4] http://www.create-net.org/projects/4/2716/Turn-Out%20Burnout

services for people who are at risk to get a burnout.

4.3.1 Introduction to Stress@Work

Stress is a physiological response to mental, emotional, or other physical challenges that humans confront in their real-life activities, including in their working environments. Continuous exposure to stress may lead to serious health problems, such as causing physical illness through its physiological effects, behaviour changes, and social isolation issues (Glanz et al., 2008; Korabik et al., 1993; Maslach et al., 2001). All these negative effects are known to affect the wellbeing of a person at workplace. As a consequence, a long-term exposure to stress typically leads to job-burnout, a state that leads to mental and physical exhaustion Maslach et al., 2001.

Over the last four decades there has been rising concern in many countries about the growth and consequences of work related stress and burnout. Recent reports show that stress is ranked as a second most common work-related health problem across the members of the European Union Milczarek et al., 2009; the same report shows that individuals with high levels of stress were accompanied by physical and psycho-social complaints and decreased work-control for the requirements placed on them.

To date, current approaches for measuring stress rely almost exclusively on self-reported questionnaires Näätänen and Kiuru, 2003, which are subjective and cannot provide immediate information about the state of a person. Therefore, a continuous stress monitoring with the use of current technology may help to better understand stress patterns and also provide better insights about possible future interventions. On the other hand, to get more information about human behaviour patterns through the use of technology requires use of less obtrusive and more comfortable devices as they measure real-life activities. Several works have shown that smartphones are an appropriate tool to collect relevant data used to classify specific human behaviour, such as Al-Mardini et al., 2014; Guidoux et al., 2014, therefore in our work we have used smartphones as non obtrusive approach to collect relevant behaviour data relative to stress levels.

4.3.2 Study demographics

In total, 30 employees from two different organisation in Trento, Italy, were selected for the study. Table 4.9 provides the summary of employees' demographics characteristics. We can note that there is a balanced mix of gender, age and education level, marital status and number of children among the subjects. The respondents in the sample comprised 16 (60%) male and female 14 (40%); married 15 (50%) and not married 15 (50%), and age ranged from 26-30 (16.67%), 31-40 (60%) and above 40 (23.33%). The participants had different educational background, where 10 (33.33%) had an academic degree, 11 (36.7%) had bachelor degree and 9 (30%) had high school education.

Table 4.9: Study demographics of the subjects in our research.

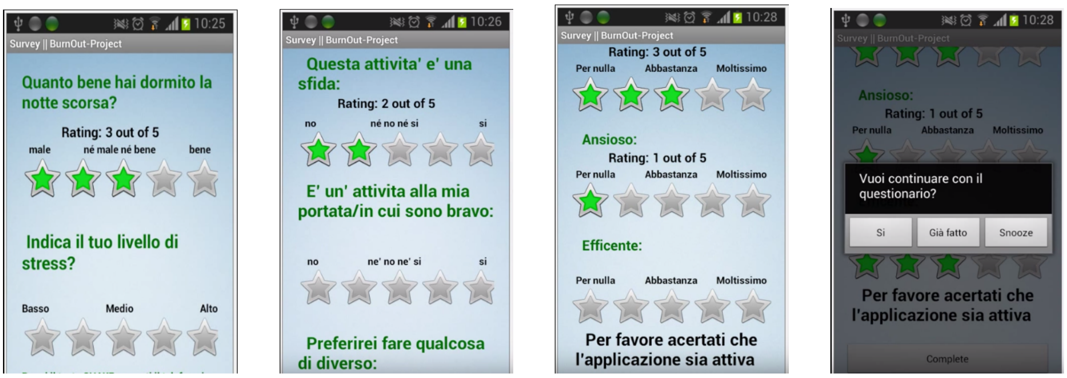
Variable	Characteristics	Nr.	(%)
Gender	Male	**18**	(60.00%)
	Female	**12**	(40.00%)
Education	High-school graduate	**9**	(30.00%)
	Bachelor degree	**11**	(36.67%)
	Graduate degree	**10**	(33.33%)
Age	26-30	**5**	(16.67%)
	31-40	**18**	(60.00%)
	>40	**7**	(23.33%)
	Mean (±SD)	**37.46**	(±7.15)
Marital status	Married	**15**	(50.00%)
	Never married	**15**	(50.00%)
No. of children	None	**17**	(56.67%)
	1-2	**10**	(33.33%)
	3-4	**3**	(10.00%)

4.3.3 Trial description

Data was collected from a group of 30 subjects in the course of 8 weeks. Considering the fact that the data collection period covered the months of November and December (where the employees have to finalize yearly objectives), we could ensure that the data contained behavioural changes from elevated stress levels. Our data collection framework was based on a server-client architecture built around the Samsung Galaxy S3 mini 32GB smartphone[1]. During the study, subjects used the smartphone in daily basis as their own phone (including working hours).

There were no restrictions placed on users regarding the handling of their smartphones, so our analysis is framed under usual/realistic conditions. The application developed to collect data was running continuously as a background application. The application started automatically at 9am at working days (Monday-Friday) without any interaction with the user. In order to understand users' mood and stress levels, the app prompted users to fill in a questionnaire at three different times of the day: at 9am (at the beginning of the work hours), at 2pm (after lunch break) and at 5pm (at the end of the work hours). The questionnaires appeared automatically and the user had the option to answer the questions or snooze the questionnaire for later. The questionnaire consisted of 14 questions that were answered is around one minute. Some examples of screenshots of the questionnaire are shown in Figure 4.6.

[1] We did not consider using other devices like smart watches are they are currently more expensive and less available among the population.



Quanto bene hai dormito la notte scorsa?

Rating: 3 out of 5

Indica il tuo livello di stress?

Questa attivita' e' una sfida:

Rating: 2 out of 5

E' un' attivita alla mia portata/in cui sono bravo:

Preferirei fare qualcosa di diverso:

Rating: 3 out of 5

Ansioso:

Rating: 1 out of 5

Efficente:

Per favore acertati che l'applicazione sia attiva

Ansioso:

Rating: 1 out of 5

Vuoi continuare con il questionario?

Per favore acertati che l'applicazione sia attiva

Figure 4.6: Examples of screen shots of the questionnaire.

Even when questionnaires appeared automatically no compulsory actions (such as blocking the phone until answering) were taken, therefore users had the possibility to ignore them. This resulted in incomplete information of two types: missing questionnaires in a day (possibly because users decided to ignore them) and missing questionnaires for a complete day (possibly because high work load). From the complete set of 30 users feature extraction was performed for two types of variables.

- The first group of variables includes information of user's behaviour during work hours, these are called *objective variables*.
- The second group contains subjective information obtained from the questionnaires which reflects the mood, work-demands/control and perceived stress of the user, these are called *subjective variables*.

Extracted data for everyday was divided into two intervals: from 9am to 2pm, and from 2pm to 5pm, referring to the subjective variables (considered as ground truth) acquired from questionnaires.

Now we present a summary of the demographics of the 30 subjects in the study. Then, we present the variables that correspond to stress and mood (subjective variables). Finally, we present the features extracted from smartphone usage (objective variables).

4.3.4 Employees state evaluation

The first type of data includes subjective information related to subjects' stress and mental state. In order to get insights in the working environments and job-demands of employees during working days, we developed a questionnaire in a smartphone application to assess several psychological working variables related to work stress. The questionnaire is clinically validated to capture subjects perceived stress and mood states of the employees at work. Three times a day the questionnaires appeared automatically (9am -at the beginning of the work, 2pm -around noon, and 5pm -before leaving workplace). The questionnaire was derived from the POMS (Profile of Mood State) scale (McNair et al., 1971) which has two dimensions related to affect of mood states, including, "Positive Affect (PA)" (*e.g.*, Cheerful, Energetic, Friendly) and "Negative Affect (NA)" (*e.g.*, Tensed, Anxious, Sad, Angry) and the rest measures disengagement from work. The PA, NA and disengagement from work items were presented in mixed order.

Each item had five response alternatives, which assessed five stress-related factors on a Likert scale ranging from 1 (absolutely agree) to 5 (absolutely disagree). The answers were stored on the mobile device and constituted part of the analysis. For the purpose of our analyses score distribution has been segmented into three regions, which in our case correspond to three ordinal classes: (*"low"* or *"poor"*), when score<3; (*"moderate"* or *"fair"*), when score = 3; and (*"high"* or *"sufficient"*), when score>3.

Table 4.10: Subjective variables: overall percentage self-reported questionnaires (exhaustion and disengagement from work) by Perceived Level (High, Moderate, Low) and Number of Subjects.

Variable	Level	Nr.Response(%)	Nr.Subjects	Variable	Level	Nr.Response(%)	Nr.Subjects
Perceived Stress	*High*	325 (22.18%)	27	**Perceived Job-control**	*High*	612 (**41.77%**)	30
	Moderate	515 (35.15%)	30		*Moderate*	604 (41.23%)	30
	Low	625 (**42.66%**)	30		*Low*	249 (17.00%)	27
Perceived Job demand	*High*	741 (**50.58%**)	29	**Perceived Energy**	*High*	357 (24.37%)	28
	Moderate	357 (24.37%)	30		*Moderate*	756 (**51.60%**)	30
	Low	367 (25.05%)	24		*Low*	352 (24.03%)	28
Tensed	*High*	118 (8.06%)	19	**Anxious**	*High*	128 (8.74%)	18
	Moderate	280 (19.11%)	28		*Moderate*	279 (19.04%)	3
	Low	1067(**72.83%**)	30		*Low*	1058(**72.22%**)	30
Cheerful	*High*	274 (18.70%)	28	**Friendly**	*High*	463 (31.60%)	27
	Moderate	756 (51.60%)	30		*Moderate*	692 (**47.23%**)	30
	Low	435 (**29.70%**)	30		*Low*	310 (21.16%)	29
Angry	*High*	83 (5.67%)	11	**Sad**	*High*	28 (1.91%)	10
	Moderate	186 (12.70%)	5		*Moderate*	112 (7.65%)	30
	Low	1196 (**81.63%**)	30		*Low*	1325 (**90.44%**)	12
Sleep quality	*Sufficient*	886 (**60.48%**)	30				
	Fair	313 (21.37%)	28				
	Poor	266 (18.15%)	24				

The first section of the questionnaire, collected information about occupational health outcomes of the participants: i) job induced stress, ii) job-control, iii) job-demand and iv) energy perceived during working days. The second section contained several widely used scales to measure mood: the existence of tensions and pressures growing out of job requirements, feelings of anxiety, cheerfulness, friendliness, sadness, angriness, and quality of sleep. In Table 4.10 we provide overall response rates of completed questionnaires on work-relevant stress from all participants throughout the entire study using the 3-point scale defined earlier. We obtained 1455 completed questionnaires, which represented a response rate of 79.97%. It is worth mentioning that in this research work we include only self-reported questionnaire items obtained at ~2pm and ~5pm, since we are interested in exploring the relation of stress, moods, and job-performance with respect to the objective variables measured in the previous working hours. We did not include data of the questionnaires at 9 am because we started to get data from the smartphones exactly at that time and could not relate (almost) any information to this questionnaire. It can be noted that employees perceived increased workload and stress, since almost all the respondents perceived a *moderate* (35.15%) to *high* (22.18%) stress level throughout the entire monitoring period.

Regarding how stress impaired productivity of the employees, almost all of them (29 out of 30) reported that at some point their job tasks and job responsibilities as highly demanding (50.58%) throughout the entire monitoring period (marked with red-colour in Table 4.10). This is important since prolonged exposure to certain job-demands has been shown to lead employees to variety of health issues, such as mental and physical disorder (Maslach et al., 2001). In response to work-related stress, 19 employees felt themselves *High - Tensed* at some point of the study, 18 respondents felt *High - Anxious*, 11 of respondents have reported *High - Angriness* (5.67%), which shows that a large group of subjects showed negative moods. Finally, a relevant physical reaction to stress is a *Poor - Sleep Quality*, which was felt by 24 of the respondents.

4.3.5 Employees Evaluation

The second type of data which provides objective measures associated with users' behaviour was collected from sensors embedded on the smartphones used in this research. From the analysis presented in Section 4.3.4 we concluded that 4 categories were needed to perform a proper assessment of subjects stress: physical activity, location, social interaction and smartphone usage. From these categories we extracted 18 features using 9 sensors, as shown in the Table 4.11.

4.3.6 Physical Activity Level - (pACL)

The potential role of physical activity (and its relation with sedentary behaviour) in the development of psychological complaints has received increased attention during the last decades (Bernaards et al., 2006; Fleshner, 2005; Penedo and Dahn, 2005). On the one hand, psychological stress has been reported as a factor in reducing frequency, intensity, and duration of physical activity (Lutz et al., 2010) by inducing specific physical responses such as tiredness, weakness, and fatigue (Spielberger et al., 2003). On the other hand, research studies have acknowledged physical activity as a psychological de-stressor (Proper et al., 2003) since an active lifestyle is associated with health benefits (Fleshner, 2005). Most related research has used mainly self-reported questionnaires to address the association between physical activity and psychological wellbeing. In contrast, we wanted to

Table 4.11: Objective variables divided in four categories. Sensors and features extracted from smartphone usage on every subject in the study.

Category	**Sensors**	**Features**
1. **Physical Activity Level**	*Accelerometer*	1) 3-axis Magnitude
		2) Variance Sum (Aharony et al., 2011)
2. **Location**	*Cellular*	3) CellID and LACID (Number of clusters (DBSCAN) (Birant and Kut, 2007)
	WiFi	4) Access Points (Number of clusters (DBSCAN) (Birant and Kut, 2007))
	Google-Maps	5) Latitude and Longitude (Number of clusters (DBSCAN)(Birant and Kut, 2007), Haversine (Robusto, 1957))
3. **Social Interaction**	*Microphone*	6) Proximity based on verbal interaction (Pitch (Hedelin and Huber, 1990), Mel-MBSES (Harris, 1978))
	Phone Calls	7) Number of Incoming Calls
		8) Number of Outgoing Calls
		9) Number of missed Calls
		10) Duration of Incoming Calls
		11) Duration of Outgoing Calls
		12) Most common Contact-Calls
	SMS	13) Number of Incoming SMS's
		14) Number of Outgoing SMS's
		15) Length of Incoming SMS's
		16) Length of Outgoing SMS's
		17) Most common Contact-SMS
4. **Social Activity**	*App usage*	18) Number of used applications (Social, System)
		19) Duration of used applications (Social, System)

investigate the association between objectively measured physical activity and perceived psychological stress.

We assume that most forms of physical activity (such as mini-breaks and lunch breaks) would reduce the level of stress and increase the positive mood of the subjects. To analyse physical activity, we measure it using accelerometer signals from the smartphones. For this research, we captured 3-axial linear acceleration continuously at a rate of 5Hz, which was sufficient to infer physical activity levels of subjects. Similar to the work in (Aharony et al., 2011), we measured the variance sum of 26 seconds (non-overlapping fixed length windows of *n=128* samples) accelerometer readings, providing the activity levels of *high*, *low*, and *none* using the magnitude of the signal (as shown in Equation 4.16), and the variance sum (varSum) in Equation 4.17:

$$\mathbf{Mag} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{x_i^2 + y_i^2 + z_i^2}; \tag{4.16}$$

$$\mathbf{varSum}(n) = ((Mag(n) - newAvg_{(Mag)}(n))^2 - (\frac{n}{n-1}) - 2Mag(n); \qquad (4.17)$$

We define three ranges of *percentage of physical activity level (pACL)* as follows: *high*-(h) when varSum≥7, *low*-(l) when 3≤varSum≤7, and *none*-(n) when varSum<3; using Equation 4.18:

$$\mathbf{pACL}_{(h,l,n)} = \frac{\textit{Number of High Activities (h)}}{\textit{Total Classified Activities (h,l,n)}} X100\% \qquad (4.18)$$

4.3.7 Location patterns

Additional sources of stress can produce behaviours such as frequent smoking, caffeine consumption and skipping lunch (Conway et al., 1981), which are known to contribute to health issues. For this reason, we analyse locations of subjects with the focus in understanding frequent locations changes during working hours. For example, we assume that during the days with high job-demands and high-stress, subjects tends to reduce changing locations or skip lunches due to their responsibilities or deadlines for delivering their work.

In order to measure location changes, we retrieved 3 important sources: (i) the list of WiFi networks available with their respective BSSID address, (ii) cell tower locations (CID, LAC-ID) and (iii) Google Maps locations information (latitude, longitude). In order to preserve the battery life of the smartphones, we have intentionally not used the GPS sensor. Using the location information we cluster locations from each source using the DBSCAN algorithm (Birant and Kut, 2007), which is an algorithm mainly used for clustering spatio-temporal locations. For Google location information, we clustered locations with maximal diameter of 300 meters (using latitudinal and longitudinal coordinates and the Haversine distance equation (Robusto, 1957)) where the subjects stay for more than 15 minutes and measured the amount of locations in each day. For Cell Tower information and WiFi networks we clustered location information on an hourly basis. Our objective is to test whether subjects show changes of location in each interval (9am-2pm and 2pm-5pm). For this we compared locations every hour counting +1 when different clusters appear with respect to the previous hour.

4.3.8 Social Interaction (SI)

Perceiving stress in everyday activities evokes a number of emotional responses that may affect interpersonal relations and social ties (AIS, 2015). Several works have reported that continuous stress may reduce social wellbeing in the long-term (Cohen and Wills, 1985). As a result, lowered social functioning (AIS, 2015) may predict decreased mental and physical health (Singh-Manoux et al., 2005). For example, social withdrawal has been used as one of the diagnostic criteria for post-traumatic stress disorder. On the contrary, being socially active has been found to reduce stress by providing a sense of security, enhancing self-confidence, and buffering the impacts of a stressful situation on individuals (Cohen and Wills, 1985).

In the last decades, monitoring social interaction has attracted significant attention (Vinciarelli et al., 2009). Social behaviour encompasses skills from social recognition and many distinct types of interaction. Previous studies monitored speech articulation aiming at inferring stress using smartphones. However, these works have been performed on controlled experimental (laboratory) studies (Lu et al., 2012).

In contrast, in this research we investigate the effects of stress on social behaviour derived from continuously recorded and classified human voice (from smartphone's microphone) in real working environments. Moreover, since social interaction includes not only face to face conversations but also phone conversations and messages, another important social aspect that we have taken into account are the employees phone conversations and SMS logs. For this, we investigate the number of conversations (incoming, outgoing and missing), SMS messages (incoming and outgoing), and unique common *called* and *calling* contacts, compared with the perceived stress on a daily basis. In order to protect users privacy, all phone call events where anonymized where we register only the five last numbers of each calling or called contact. In detail, we measured two aspects of social interaction:

- **Speaker Recognition:** Recent work in stress detection suggest to use Bluetooth embedded sensor on smartphones for measuring social-proximity (Bogomolov et al., 2014). However, this method poses several disadvantages since the users may not carry the phones all the time. Second, Bluetooth scans have time limits, which restricts the estimation of social-interaction.

  In contrast, in this research we use the microphone embedded on the smartphones for better and accurate recognition of verbal interactions, namely social-interactions. We have extracted two main audio features (Pitch (Hedelin and Huber, 1990) and Mel-Multi-Band Spectral Entropy Signature, Mel-MBSES (Harris, 1978)) to obtain a higher accuracy in speech activity recognition.

  In this research, two conditions required for processing audio on smartphones: i) measuring pitch within the range of human voice (40 Hz to 600 Hz), and ii) recognizing human voice from the captured frames using the MEL-MBSES coefficients and Support Vector Machine (SVM) classifiers (Vapnik et al., 1997). We built a SVM (Vapnik et al., 1997) classifier using MEL-MBSES coefficients trained on frames coming from 3 minutes of voiced data and 3 minutes of background data. The training set for the SVM consisted of positive vectors (speech) and the negative vectors (non-speech or background). We sampled audio frequency of 8000Hz and set a frame every 256 samples where we calculated Pitch and Mel-MBSES features for each frame, then each frame is labeled either as human voice or not a human voice. Approximately every 0.7 second (7 out of 30 frames) must be detected as voice in order to indicate voice activity in that audio segment. We measured percentage of social-interaction based on the total duration (hourly, daily, weekdays) of conversations as shown in Equation 4.19:

$$\boldsymbol{SI} = \sum_{i=1}^{n} \frac{\mathbf{True}_{Classified}}{\mathbf{Total}_{Classified}} \times 100\% \tag{4.19}$$

It should be noted that since there were no restrictions on the use of the smartphones, in some cases these were placed inside pockets. The smartphone can still be used to recognise voice in these cases, although the information is less reliable and only works at reduced distances. This may result on underestimating our results for social interaction.

◆ **Phone-Call and SMS behaviour:** Since calling and texting messages (SMS's) behaviour could be an important source to infer stress-relevant factors we consider phone calls in terms of: *number*, *duration* and *most frequent number (on a daily basis)* of *incoming, outgoing* and *missed*. Furthermore, for SMS's, we measure *number* and *length* (incoming and outgoing). These features may serve as a source of stress, for example understanding phone-call behaviour from subjects that contact different persons more frequent during stress-less periods in comparison with stress-full times. In order to find the most common called/calling ID in each interval (9am-2pm and 2pm-5pm) we used $\text{argmax}_{(Call)} = \sum_{i=1}^{n} countmax(CallID)$ and $\text{argmax}_{(SMS)} = \sum_{i=1}^{n} countmax(SMSID)$ for most frequent Call and SMS's respectively. In order to remove ties among ID's that have the same number of calls, we proposed a scoring model *Score* for both calls and SMSs:

$$\mathbf{Score}_{(Call)} = \frac{duration(CallID)}{countmax(CallID)} \text{ and } \mathbf{Score}_{(SMS)} = \frac{length(SMSID)}{countmax(SMSID)}$$

4.3.9 Social Activity

Finally, another aspect that may have impact on the stress levels is application usage of the smartphones. Our first intention was to explore the impact of smartphones usage during working days and to investigate whether their usage were more likely to view them as a positive influence in balancing their work and personal life. For this, each time and employee uses an app, our software captures the event and stores it together with the duration and time-stamps. With this information we were able to extract the following data: number of application used per interval and duration of their usage. Applications were divided in two categories:

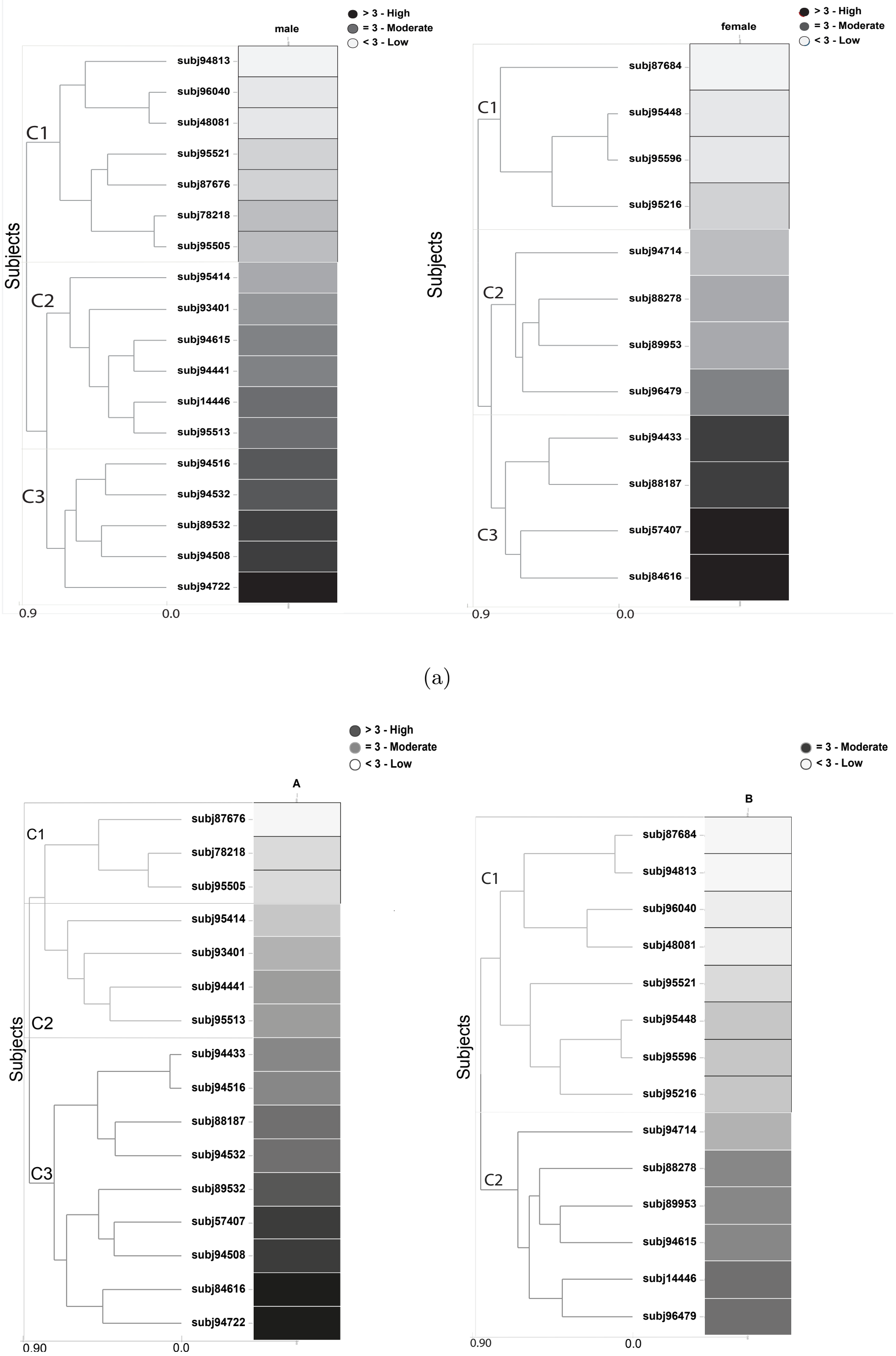
- System apps: pre-installed apps like Camera or Calendar, Web-browsing, E-Mail client.
- Social apps, such as Viber, WhatsApp, Facebook, Skype and other user downloaded apps (*e.g.*, games other entertainment apps).

4.3.10 Analysis of information

Using the features presented in Section 4.3.3 we retrieved the data from all the participants in the study. First, data was filtered discarding information from weekends and hours not in the range 9:00am-5:00pm (representing the working hours). Recall that this range is closely related with the ground truth information acquired from self-assessments (Section 4.3.4). After the data was filtered, different techniques were used to perform a thorough analysis: (i) we started using hierarchical clustering (Section 4.3.11), (ii) then correlation analysis (Section 6.1.1), and (iii) finally, we performed variable importance analysis (Section 6.1.2.1).

4.3.11 Diversity and similarity of stress level within subjects

Hierarchical clustering was used to analyse the participants self-reported stress on a daily basis. We used Ward's method (Ward Jr, 1963) to perform the hierarchical clustering of self-reported stress using the half-square euclidean distance between subjects. Euclidean distance is always greater than or equal to zero. Measurements would be ≈ 0 for identical subjects and ≈ 1 for subjects that show less similarity. Figure 4.7 present dendrograms about the perceived stress level divided by gender and organisation. Each dendrogram is ordered by clusters, and inside each cluster they are ordered by mean values of perceived stress level. From these figures we can note that gender do not easily determine the stress level since both of them show a great variation of perceived stress, however, as we will see, at least in these experiments, there is a higher percentage of women in the high



(a)

(b)

Figure 4.7: Dendrograms obtained by computing similarities and diversity between perceived stress level of each subject (a) by Gender and (b) by Organisation. Three major clusters can be noted, colour boxes correspond to average stress for different subjects.

stress group. In contrast, when clustering by organisation we can see that subjects in organisation A showed in average a higher stress than those in organisation B.

It is interesting to note that organisation A is an IT organisation, while B is a social support organisation. In Table 4.12, we provide an overview of clustering results based on gender. Cluster analysis yielded 3 distinct clusters (C1, C2 and C3) which represent *low*, *moderate*, and *high* stress levels. Note that women show a uniform distribution across stress levels and men showed slightly more subjects with low stress. We also performed clustering within the organisations, which is shown in the Table 4.13. The results show that stress was different between organisations. For example, in organisation A, all women (4) showed high stress levels. In contrast, in organisation B, half of the women showed low stress and half of the women showed moderate stress levels. Again, in this company, there are slightly more men with low level of stress.

Table 4.12: Perceived stress level from dendrogram analysis by gender. Three major clusters can be noted based on perceived level of stress.

Cluster (Stress-Level)	**Men (Nr./%)**	**Women (Nr./%)**
C1 (*low* < 3)	7/18 (38.89%)	4/12 (33.33%)
C2 (*moderate* $= 3$)	6/18 (33.33%)	4/12 (33.33%)
C3 (*high* > 3)	5/18 (27.78%)	4/12 (33.33%)

Table 4.13: Perceived stress level from dendrogram analysis by Gender within Organisations. Three major clusters can be noted based one perceived stress.

Cluster (Stress-Level)		**Organisation A**	**Organisation B**
C1 (*low*<3)	**Men:**	3/12 (25.00%)	4/6 (66.67%)
	Women:	0/4 (0.00%)	4/8 (50.00%)
C2 (*moderate*=3)	**Men:**	4/12 (33.33%)	2/6 (33.33%)
	Women:	0/4 (0.00%)	4/8 (50.00%)
C3 (*high*>3)	**Men:**	5/12 (41.67%)	0/6 (0.00%)
	Women:	4/4 (100.00%)	0/8 (0.00%)

Table 4.14: Perceived Stress Level from dendrogram analysis by Response Intervals ([9am-2pm], [2pm-5pm]). Three major clusters can be noted based on perceived level of stress and transition of perceives stress into intervals.

Cluster (Stress-Level)	**Intervals**
C1 low→low; *low⟷moderate*	11/30 (36.67%)
C2 moderate⟷low; *moderate⟷moderate*	12/30 (40.00%)
C3 high⟷moderate; *high⟷high*	7/30 (23.33%)

Finally, we clustered self-reported stress changes within intervals (9am-2pm and 2pm-5pm) as shown in Table 4.14. For example, low ⟷ moderate, means that subjects in

Table 4.15: Overall average percentage of physical Activity Level (pACL) by Intervals (9am-2pm and 2pm-5pm) and Perceived Stress Level (SL) [High, Moderate, Low].

Distribution of pACL by (Gender, Age, Education, Marital Status and Organisation)	**pACL** [9am.-2pm.]	**pACL** [2pm.-5pm.]	**High (SL)**	**Moderate (SL)**	**Low (SL)**
– **Male**	**18.03**	**21.34**	16.29 (*)	16.68	**23.60**
– **Women**	15.66	18.74	10.57 (**)	15.37	**18.89**
– **26-30** (28.6±1.95)	12.89	15.48	12.45	13.65	**17.83**
– **31-40** (35.33±2.4)	17.50	21.00	12.87	16.22	**21.97**
– >**40** (49±2.52)	**18.69**	**21.66**	17.61	18.20	**21.90**
– **High school graduate**	17.01	21.40	16.84	16.84	**18.77**
– **Bachelor degree**	**19.22**	**23.52**	11.70	17.48	**29.19**
– **Graduate degree**	14.78	15.54	12.64	14.86	**16.51**
– **Married**	**20.51**	**25.48**	17.71	19.53	**26.73**
– **Never married**	13.36	14.78	10.23	13.39	**16.31**
– **A.**	12.17	15.50	12.21	10.77	**17.33**
– **B.**	**22.45**	**25.49**	18.39	23.93	**24.21**
– **Overall (Mean±SD) of pACL (%)**	17.06 (±12.01)	**20.14** (±13.12)	16.43 (±16.42)	16.46 (±12.30)	**19.65** (±12.85)

(*) **16/18** - male subjects perceived high stress.
(**) **11/12** - female subjects perceived high stress.

the clusters showed low stress levels in the first interval and the changed to moderate in the second interval or that moderate changed to low. In this case, 23.33% of the subjects showed at least a high level of stress in their daily activities (*high*⟷*moderate* or *high*⟷*high*) and 2/3 of the subjects (63.33%) showed levels between moderate and high. It is important to note that employees did not perceive drastic changes of stress, from *low*⟷*high.* Now we present a more detailed analysis for each category of objective variables and its relation with mood, and specifically with perceived stress levels.

As a summary of this first set of experiments, we can note that with our current data: (i) there is a slight bias in men towards lower levels of stress in their working environments, (ii) there is a clear difference between stress levels in companies, where an IT company showed higher stress levels than a social support company, (iii) about 2/3 of the employees perceived moderate to high stress and 23.33% perceived high stress, and (iv) there were no drastic changes between levels of stress.

4.3.12 Physical activity levels

Table 4.15 presents overall percentage of physical activity level with respect to perceived stress level (*High*, *Moderate*, and *Low*) on a daily basis for all 30 participants compared with demographic characteristics (age, gender, education, marital status, number of chil-

Table 4.16: Overall average percentage of activity level (mean ± Std.dev.) during working days and perceived level (SL) of Stress (H-*high*, M-*moderate*, L-*low*) by Gender.

	Men			***Women***		
	H (SL)	**M** (SL)	**L** (SL)	**H** (SL)	**M** (SL)	**L** (SL)
Monday:	**24.3±22.2**	16.2±16.2	21.6±18.0	12.3±12.1	13.0±7.0	**21.6±22.4**
Tuesday:	10.0±6.5	17.5±16.6	**22.2±14.4**	6.2± 3.1	12.3± 6.3	**16.5±7.7**
Wednesday:	18.0±19.8	19.8±18.3	**22.5±18.5**	12.6±8.4	13.2±7.7	**14.6±7.6**
Thursday:	19.0±20.7	20.7±18.6	**24.3±18.7**	9.6±8.0	**17.9±12.4**	14.3±13.9
Friday:	14.9±17.3	15.9±19.6	**20.4±19.5**	11.4±12.4	**17.7±13.6**	13.8±8.0

Table 4.17: Overall average percentage of activity level (Mean ± Std. Dev.) by Job-Demands, Job-Control, Energy and Sleep-Quality perceived level (PL) with respect to Gender.

	Men			***Women***		
	H (PL)	**M** (PL)	**L** (PL)	**H** (PL)	**M** (PL)	**L** (PL)
Job-Demand	20.0 ± 16.7	17.9±14.8	**22.3 ±22.1**	**16.1±6.0**	13.1±8.4	11.7±6.2
Job-Control	**19.0±14.6**	18.0±14.8	18.9±16.0	14.2± 6.5	**16.7±6.2**	13.3 ± 8.7
Energy	**23.7±15.6**	19.9±14.8	17.8± 14.6	14.1± 9.7	15.3± 6.1	**16.8±9.7**
Sleep-Quality	20.9± 14.3	22.1±15.8	**22.7± 17.4**	**16.6±6.9**	15.9±6.8	15.8± 6.9

dren and organisations). Activity levels were normalized for each interval (9am-2pm and 2pm-5pm) or for a complete day.

Some conclusions are:

- pACL during lower perceived stress times was associated with higher activity (19.65% of activity). In contrast, a high perceived stress showed less activity (16.43%).
- Subjects were more active during the second interval (2pm-5pm), with 20.14% pACL compared to 17.06% in the first interval.
- Following age, education level, and marital status, participants that reported *high* and *moderate* stress levels were associated mostly with lower pACL than when they have *low* stress.
- The age group of (≥40) showed more activity level than the rest when they perceived high stress level.
- The group of married subjects showed more activity than the never married group no matter their perceived stress level.

Furthermore, separating overall activities into working days allowed us to compare pACL in different days of the week (as shown in Table 4.16). Results show that men have a higher pACL only on Mondays when they perceive high stress. In contrast to women that do not show a high level of pACL when they perceive high stress.

As described in Section 4.3.4, it is important collecting information about occupational health, such as job-demands and job-control. In this regard, Table 4.17 shows mean scores on perceived job-demand, perceived job-control, perceived stress, and perceived energy for the respondents. Low perceived job-demand was associated with higher physical activity level (22.3%) for male participants. In contrast, women showed increased activity levels when they perceive high job-demands. Men participants with higher pACL perceived higher energy. In contrast, women with higher pACL showed lower energy. In summary, this table shows that in general men and women show different results in terms of perceived emotions with respect to their activity levels throughout the day.

4.3.13 Social Interaction

In contrast to our previous work (Ferdous et al., 2015), where we explored the correlation of total amount of verbal interaction per day with self-reported stress, in this research we expand that analysis, since now we explore the distribution of the verbal-interaction in an hourly basis and working intervals.

In Table 4.18 we present a summary of social-interaction levels, with respect to different characteristics of the employees.

Some findings are the following:

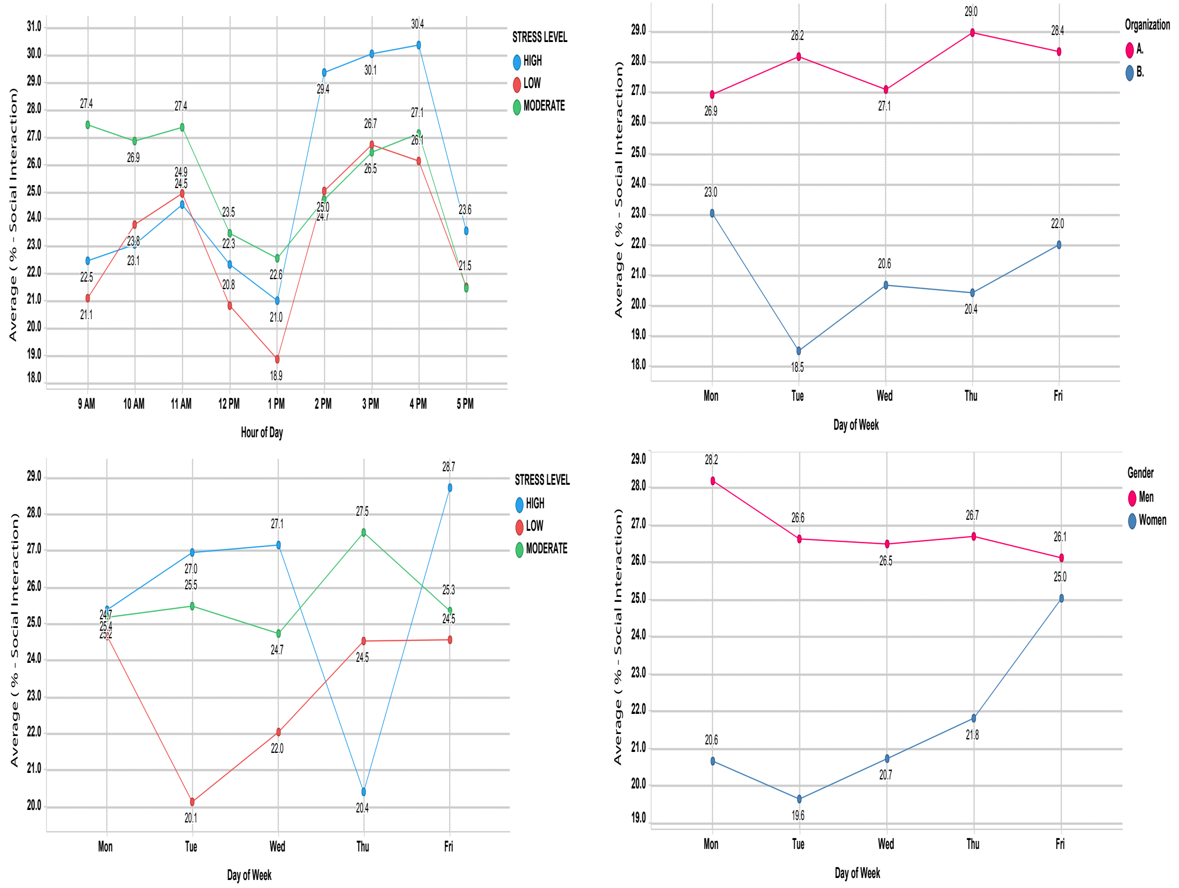
- In average subjects showed higher social interaction in moments of low stress than in moments of high stress.
- However, analysing this data by age group we observe that older (and married) employees showed the opposite behaviour, they increased their social interaction during low levels of perceived stress.
- There is in general more social interaction in the afternoons than in the mornings.
- Another interesting behaviour appears across organisations. In this case subjects in organisation A showed higher social interaction than those in organisation B.

We explore further these measurements. We depict social-interaction as a) percentage in hourly basis in a day, b) per day day of week, c) per hour within organisations and d) per day of the week by gender in Figure 4.8.

- A notable result is a homogeneous behaviour (similar shapes of the curves) of social interaction across stress levels (Figure 4.8 (a)), with higher interaction in the morning for moderate perceived stress and a higher interaction in the afternoon for high perceived stress.
- Another homogeneous behaviour is shown across organisations, where people decrease their social interaction near lunch time (12-13 hrs), see Figure 4.8 (c).
- The peak of the verbal interaction in High-Level of stress is achieved in the afternoon (one hour before the end of the working day).

- When subjects perceive high stress, social interaction drops on Thursdays and then increases again on Fridays, see Figure 4.8 (b).
- The social interaction varies with the perceived stress during the week, except on Mondays where it has similar values with the different perceived stress levels.
- With respect to gender, men showed a more stable social interaction across the weekdays. In contrast, women then to increase their interaction near the weekend, see Figure 4.8 (d).

Figure 4.8: Overall percentage of social Interaction and stress Level a) by working hours, b) by week days, c) by Organisations and d) by Gender.



Social interaction also include phone calls and SMS behaviours. Using the self-reported stress level, we were able to compare the phone activeness from 5767 phone calls and 5911 SMS's. To be noted, that all marketing SMS's or responses from the GSM operators were excluded in this research. In Tables 4.19 and 4.20 we explored the relation of phone calls and SMS's with respect to perceived level of stress. From these tables it can be seen that the number of phone-placed *Outgoing*, phone received *Incoming* and missing calls, was

Table 4.18: Distribution of Social-Interaction (SI) by Response Intervals ([9am.-2pm.], [2pm. - 5pm.]) and Stress-Level (SL)

Distribution of SI by Gender, Age, Education, Marital Status, Organisation	**SI [9am.-2pm.]**	**SI [2pm.-5am.]**	**High (SL)**	**Moderate (SL)**	**Low (SL)**	**Nr.** Employee
– **Male**	**25.67**	**28.75**	**27.88**(*)	27.54	25.74	18
– **Women**	20.17	23.83	**22.88**(**)	22.72	19.79	12
– **26-30** (28.6±1.95)	**25.46**	**29.53**	**28.57**	26.02	26.44	5
– **31-40** (35.33±2.4)	22.96	26.61	24.90	**26.67**	22.34	18
– >**40** (49±2.52)	22.73	24.84	22.97	22.54	**24.32**	7
– **High school graduate**	20.63	25.09	22.97	**26.95**	26.22	11
– **Bachelor degree**	24.23	**28.16**	**29.49**	26.28	23.12	10
– **Graduate degree**	**25.30**	26.94	22.81	**23.37**	21.24	9
– **Married**	21.75	25.02	22.91	21.92	**23.56**	15
– **Never married**	**24.68**	**28.07**	27.61	**28.45**	22.89	15
– **A.**	**26.40**	**30.41**	27.96	**29.64**	25.67	16
– **B.**	20.07	22.49	18.20	20.21	**21.80**	14
– **Overall (Mean±SD) of SI (%)**	23.61 (±10.53)	**26.93** (±11.04)	23.47 (±11.02)	24.58 (±10.47)	**25.28** (±11.67)	**30**

(*) **16/18** - male subjects perceived high stress.
(**) **11/12** - female subjects perceived high stress.

Table 4.19: Number of phone-calls by perceived stress level (SL).

	Nr. Phone Calls	**High SL**	**Moderate SL**	**Low SL**
Incoming:	1696 (100%)	355 (20.9%)	511(30.1%)	**830 (48.9%)**
Outgoing:	2912 (100%)	547 (18.7%)	839 (28.8%)	**1526 (52.4%)**
Missing:	1159 (100%)	220 (18.9%)	405 (34.9%)	**534 (46.1%)**

Table 4.20: Number of SMS's by perceived stress level (SL).

	Nr. SMS	**High SL**	**Moderate SL**	**Low SL**
Incoming:	3767 (100%)	1067 (28.3%)	801 (21.2%)	**1899 (50.4%)**
Outgoing:	2144 (100%)	697 (32.5%)	710 (33.1%)	**737 (34.3%)**

higher when subjects perceive less stress. In the appendix (Tables A.5 and A.6) we show the overall mean number, duration and length of *Outgoing, Incoming, Missed Calls and SMS's (Incoming, Outgoing)* from 30-subjects throughout the entire monitoring period, using demographics of the study and separating into weekdays.

We also analysed the duration and length of SMS's and calls and some interesting observations are the following:

- In stress-full days, in most of the cases *Outgoing* calls have in average shorter duration.
- Longer duration of *Incoming calls* were associated with high perceived stress level.
- Almost in all cases a high number (and length) of *Incoming-SMS* and *Outgoing-SMS* were also related to *high* stress.
- Analysing the conversations by weekdays, high perceived stress was associated with longer duration of *Incoming-Calls* and the length of *Incoming-SMS's*, which in contrary to duration of *Outgoing-Calls* and length of *Outgoing-SMS's* is lower when the employees perceive high stress. Similarly, having high job-demands was associated with lower duration of phone-call and length of SMS's in all categories.

Moreover, in Figures in 4.9 we depict the frequency of the most common contact for phone calls and SMSs (blue line) for every subject. From these figures we note a higher frequency of phone-calls and SMS's with the most common contacted number when they perceive high stress levels (average frequency of most frequent contacts is shown with red line). In contrary, in low and moderate stress the frequency of phone-call is in average lower. These results shows that higher frequency of the phone-calls and SMS's can be an indicator of stress during the working times.

4.3.14 Location changes

Table 4.21: Overall number of clusters obtained from location using the DBSCAN algorithm by perceived Stress Level (SL). Descriptive statistics (Mean±SD) provide information of overall number of clusters retrieved from the 30-subjects throughout the entire monitoring period.

Locations	**Clusters** 9am-5pm Nr.(Mean±SD)	**High**-(SL) Nr.(Mean±SD)	**Moderate**-(SL) Nr.(Mean±SD)	**Low**-(SL) Nr.(Mean±SD)
Cell:	1383 (1.05±0.38)	230 (1.01±0.39)	349 (1.07±0.40)	**527** (1.05±0.33)
WiFi AP's:	2663 (1.40±1.38)	486 (1.42±1.41)	742 (1.49±1.35)	**961** (1.55±1.39)
Google Maps:	628 (0.48±0.78)	143 (0.63±1.01)	158 (0.48±0.90)	**234** (0.46±0.85)

To analyse location changes we measured the number of clusters obtained from different locations throughout the entire monitoring interval (see Table 4.21). From all three sources it is evident that overall subjects tend to reduce visiting different places or going further away from work environments when they perceive high-stress level during working

Figure 4.9: Frequency of the most common contact calls for each subject by perceived stress level.

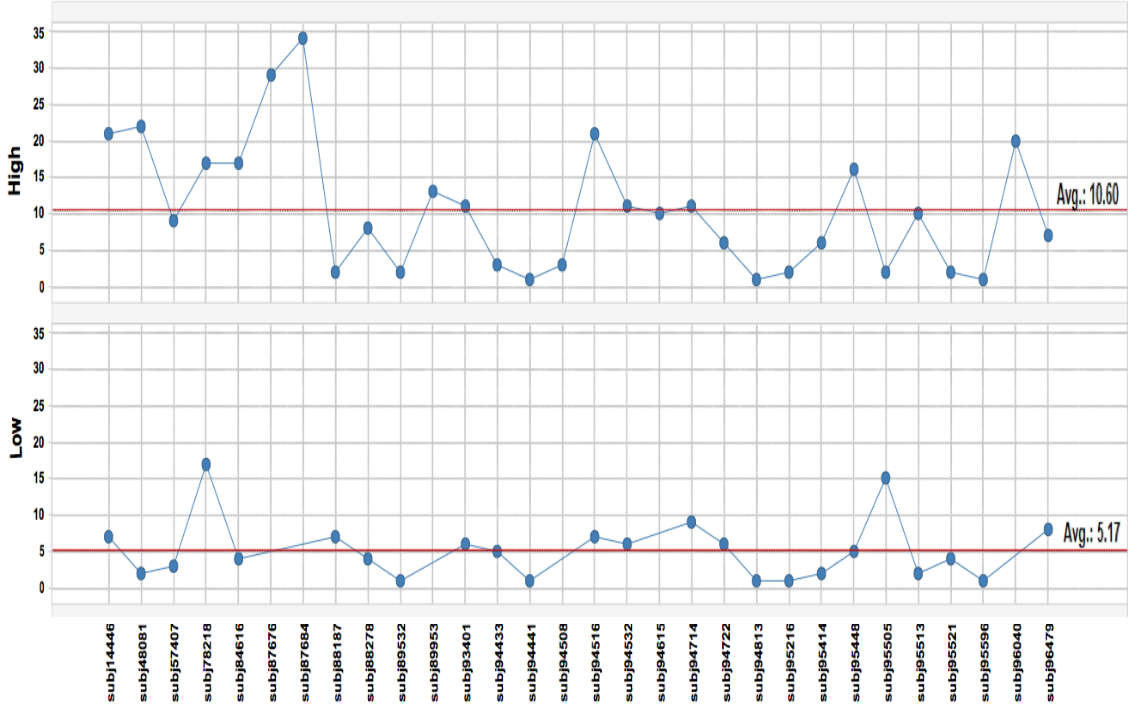
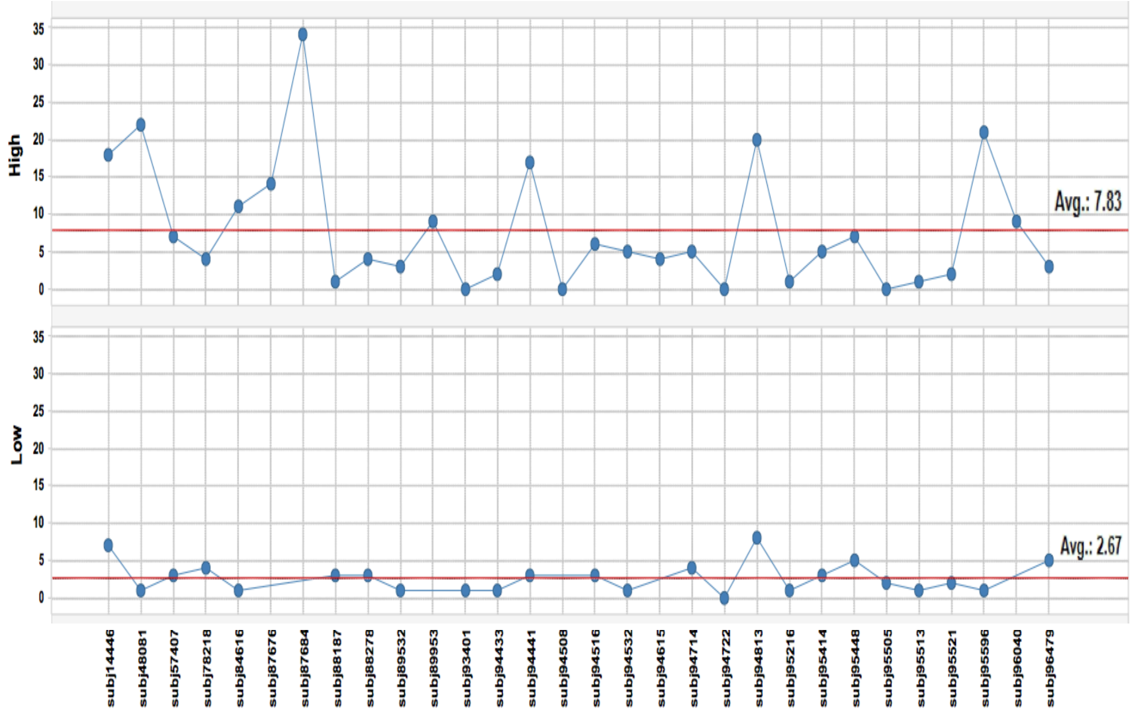


Figure 4.10: Frequency of the most common contact SMS's for each subject by perceived stress level.



days. We obtained more Cell-Tower and WiFi clusters from both parameters due to frequent scanning. Changes of clusters in WiFi represent changes of indoor locations, such as changing environments, areas, departments or either having mini-breaks in specific hours. In contrast, using Google Maps locations show distance and most visited places outdoors.

Table 4.22: Overall number/duration (sec.) of phone application usage by perceived stress level (SL). Descriptive statistics (Mean±SD) provides overall usage of applications from 30 subjects during the entire monitoring period.

Perceived Stress Level	**Frequency System-Apps** Nr.(Mean±SD)	**Frequency Social-Apps** Nr.(Mean±SD)	**Duration System-Apps** Nr.(Mean±SD)	**Duration Social-Apps** Nr.(Mean±SD)
High	5531 (24.0±26.1)	357 (3.5±4.0)	48445 (211.0±157.2)	4621 (45.3±52.5)
Moderate	7823 (25.2±28.5)	508 (4.0±4.2)	57607 (185.2±153.3)	7420 (57.0±73.2)
Low	**13787** (31.0±28.2)	**966** (4.3±4.2)	**88782** (197.2±150.3)	**9582** (42.3±65.2)

4.3.15 Application usage

Another source that provides information relevant to subjects daily activities at work is the usage of the smartphone applications. Recall that we divided the type of applications subjects ran on their devices during the working days and we categorized them into system and social applications (as described in Section 4.3.9). Next, we examine the frequency (number of accesses) and the duration of the applications used and contrast them with the perceived self-reported stress on a daily basis (see Table 4.22). Results show that in stress-less times subjects tend to use longer times the smartphone (both social and system applications). This also seems a good indicator for identifying perceived stress levels.

In summary, from these results we can draw the following conclusions:

- Activity levels changed with perceived stress and with weekdays.
- There is an opposite behaviour of activity levels in male and female in terms of job-demand and energy.
- There is more social interaction with higher stress levels except for older people that show an opposite behaviour.
- There is more social interaction during the afternoons.

- There is an increase level in social interaction by women towards the end of the week.
- There is a very different social interaction among employees of different companies. Curiously the company with higher stress levels also have higher percentages of social interaction.
- There are shorter outgoing calls and longer incoming calls during high stress levels.
- People use much more their smartphones during lower perceived stress levels.

4.4 Chapter Summary

In this chapter, we presented the data the were collected in both studies selected for this research. We demonstrated the features that has been extracted from sensors and the methods used to analyse the data.

In the following chapter, we present the methods proposed to infer behaviour changes and to handle the scarce data collected from bipolar disorder patients.

Chapter 5

SCARCE DATA AND CLASSIFICATION OF BIPOLAR DISORDER

> *"Our technologies become more complex while we become more simple. They learn about us while we come to know less and less about them. No one person can understand everything going on in an iPhone, much less pervasive systems."*
>
> – **Douglas Rushkoff**

This chapter summaries the thesis' the proposed approach in classification of motor activity levels in different bipolar disorder states. We begin with the importance of monitoring physical activity in bipolar disorder in they real-life activities. Further, we use semi-supervised learning method to address the problem of scarce data and missing information. We frame the challenges facing the building of accurate models for predicting disease progression. The chapter provides the proposed approaches (i.e., Self-training, Intermediate models) to improve the knowledge of their state. Finally, it closes with summary of current research directions.

The contributions of this chapter are as follows: [1]

[1] This chapter is manly based on the following research work:
I. Maxhuni, A., Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O. and Morales, E.F., 2016. Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. Pervasive and Mobile Computing. (Maxhuni et al., 2016a).
II. Osmani, V., **Maxhuni, A.**, Grünerbl, A., Lukowicz, P., Haring, C. and Mayora, O., 2013, December. Monitoring activity of patients with bipolar disorder using smart phones. In Proceedings of International Conference on Advances in Mobile Computing & Multimedia (p. 85). ACM. (Osmani et al., 2013a).

A.1 *We propose using "Semi-supervised learning" methods to cope with scarce data acquired from the bipolar disorder patients in our trials.*

A.2 *We propose analysing correlation strength between patients physical activity levels and their psychological state.*

A.3 *We propose using our novel "Intermediate models" method to build better predictive performance.*

A.4 *We have evaluated the dataset comprising 5 subjects; we measure the motor activity and speech production during the phone conversation to classify the depression level.*

The outline of this chapter is as following: the Section 5.2 describes the problem statement related to disease interventions and importance of classifying psychological state of bipolar patients. Section 5.3 begins with evaluation of physical activity level and psychiatric evaluation level. Further, in the Section 5.4 and Section 5.5 we analyse the information about patients motor-related behaviour and voice production while having conversation on the phone. Finally, we propose using Semi-supervised learning method and Intermediate Models to improve the classification accuracy of bipolar disorder in presence of scarce data.

5.1 Monitoring bipolar disorder patients

The worldwide prevalence of many chronic health conditions is steadily increasing, so the management of diseases represents one of the most important challenges for health systems. The World Health Organisation (WHO) has ranked mental disorders and mental injuries within the top 20 causes of disability among all medical conditions worldwide in persons aged in the range 14 to 55 (WHO, 2001). Like other psychiatric disorder such as schizophrenia and major depression, bipolar disorder (BD) is a severe and chronic psychiatric illness that is associated with high rates of medical morbidity and premature mortality (Bopp et al., 2010). In 2001 bipolar disorder was ranked as the 6th leading disabling illness worldwide (WHO, 2001) and is associated with high cost for healthcare.

As a matter of fact, the mortality is high in people who suffer with bipolar disorder and is estimated two to three times higher in comparison with the mortality of general population (Belmaker, 2004). Relapse in bipolar patient increases over time and the relapse can vary from a few weeks to many months. Therefore, patients with bipolar disorder require lifetime maintenance therapy (Morriss, 2004).

Illness characteristics and neuro-cognitive deficits certainly influence the quality of life and general functioning in bipolar disorder patients. One of its main characteristics is a repeated relapse of two polar episodes, mania and depression. Patients suffering from the disorder may experience episodes of altered mood states ranging from depression with sadness, hopelessness (including suicidal ideation), loss of energy, and psycho-motor retardation, whereas manic episodes are characterized by irritability, excessive energy (hyperactivity), reduction in the need of sleep and psycho-motor agitation or acceleration.

The diagnosis of bipolar disorder is based on clinical evaluations through interviews and evaluations of scores gathered by quantitative psycho-pathological rating scales that were developed in the early 1960s (*e.g.*, HAMD, BRAMS, YMRS) and other more recent variations of them (*e.g.*, BSDS). Although these interviews and questionnaires are well established and defined in a specific manual (Faurholt-Jepsen et al., 2012), they have their drawback, as they are performed on sporadic days, while a change to a potentially dangerous state can be produced in between these sessions. Other approaches include daily self-reports, however, they can be unreliable as they often depend on current mood episode polarity of the patients (Sims et al., 1999).

Currently, drug therapy is the main treatment in BD, but its effectiveness critically depends on the timing of administration and has to be individually modified according to a patients' state of mind. Therapy can be very effective if administered at the beginning of a patient's transition to a different state, however it may be less effective in severe states where the symptoms are present and persisted to a significant degree. The advantages of using smartphone technology to monitor bipolar disorder have recently been documented in the work carried out in MONARCA EU project (Grunerbl et al., 2014, 2015; Osmani, 2015; Osmani et al., 2013b) and have presented the basic concepts of using smartphones for the management of bipolar disorder. Using the sensor embedded in smartphones for inferring significant usage data, such as location patterns from day-to-day activities, social-interaction sensing, level of physical activity, that objectively monitor the state of patients with bipolar disorder might increase the availability and pervasiveness of treatment. Sensor data acquired from smartphones offers huge potential that through machine learning techniques get valuable insights of behaviour of bipolar disorder patients in their real-life.

5.2 Monitoring in mental-health

In recent years, different systems have been developed aiming to monitor, diagnose and provide health services to the individuals, including the mental-care. Considering the popularity of smartphone devices, new challenging possibilities are opening up, those of monitoring subjects outside the laboratory, in unconstrained and uncontrolled environments so as to capture subjects' natural behaviour. Possibilities of sensing outside the lab are numerous, ranging from lifestyle monitoring, behaviour change, detecting stress and burnout in workers, up to applications in medicine, including monitoring of patients with major depression and bipolar disorder.

Smartphone computing can have a substantial impact in monitoring patients with mental disorders due to the following factors:

- symptoms of mental disorders are primarily manifested through changes in patients' behaviour. For example depression is manifested through motor retardation, where such change in behaviour can be captured through analysis of the information from the motion sensors on the mobile phone; and
- psychiatric assessment of mental disorders is typically carried out through the use of a questionnaire. The questionnaire relies on patients recalling events pertaining to their past behaviour, such as amount of physical activity for example reported by the patient. Self-reporting suffers from a number of issues, including:
    1. **Recall bias:** where subjects have difficulties recalling events in the past;
    2. **Subjectivity:** self-reporting may be affected by the current mental state of the subject; and
    3. **High effort:** self-reporting requires high effort in order to gather high quality data, especially in longitudinal studies, where data is gathered either through self-reporting or through a third party observer.

Smartphone computing can address these difficulties through continuous monitoring of user activities, by sampling sensors commonly found on mobile devices and in return providing objective measures of behaviour phenomena and also allow for experience sampling through self-reporting. Continuous monitoring is especially suitable for measuring physical activity, since activity levels of individuals can be measured through the phone's accelerometer and a solid picture of overall physical levels can be inferred. Measuring physical activity levels in this manner alleviates the issues faced when relying on subjects' memory of physical activity events, which is the current practice in psychiatry.

5.3 Physical activity monitoring in bipolar disorder patients

Significant interest in physical activity monitoring for patients with bipolar disorder is increasing. Some of the most important findings regarding physical activity, have shown how physical activity reduces risk for chronic diseases, such as cardiovascular diseases (Winkler et al., 2011), obesity (Kriska et al., 2003), and enhance mental-health with respect to lowering levels of anxiety and depression, elevating mood, improving self- esteem and reducing stress (Bartholomew et al., 2005; Mata et al., 2012; Vancampfort et al., 2013a). Hence, an accurate measurement of physical activity is an important component of research, in order to monitor people's health and to quantify the relation between physical activity and outcomes of chronic diseases.

Most studies utilizing self-monitoring are based on traditional monitoring with paper and pencil diaries and questionnaires (Gwaltney et al., 2008). These methods are often biased assessment of the health outcomes. Due to irritable state of individuals with bipolar disorder during depressive and manic state, using traditional methods patients are prone to neglect or to overestimate performed activities. Thus, if self-assessment on the mobile phone enables easier monitoring and tracking of the patients' progress than traditional methods, then the data collected will be of higher quality. The benefits of using technology include more accurate data and also provide clinicians with the ability to evaluate the patients' progress in a more granular scale and increase the efficacy of the treatments. Moreover, bipolar patients who are trained to use self-help treatments can benefit from greater control over their care and life decisions and can detect early warning signs of serious illness (Morriss, 2004).

5.3.1 Correlation between physical activity levels and episodic state in bipolar disorder

During initial analysis phase, we were interested whether overall physical activity levels show any correlation with the patients' state. Literature suggests that patients in the depressive state show decreased levels of physical activity in comparison to their normal state, while the contrary holds true for manic patients. Note that in this research we did not carry out between subjects comparison, rather we focused on differences within subject. Table 5.1 shows activity levels of patients for the whole duration of monitoring of 3 months and correlation with the psychiatric evaluation scores, using Pearson correlation coefficient r.

As it can be seen from the Table 5.1, there exist a correlation between the patients' state and the overall physical activity levels. The correlation is strongest for patient P0101 *(r=0.672)*, while there is a low negative correlation for P0302 *(r=-0.148)*, which indicates that the overall level of physical activity (as measured by the phone) was decreasing as the patient's state was improving (patient P0302 went from major depressive episode (-3) to a normal state (0)).

While there have been studies that correlate overall physical activity levels with depressive and manic episodes (Judd et al., 2012; Vancampfort et al., 2013b), our research did not yield strong correlation for all patients using overall daily physical activity levels.

Considering these results, we have decided to investigate further in order to understand how the daily behaviour levels correspond to bipolar disorder episodes. In this respect, we

Table 5.1: Correlation between patient state and overall physical activity levels *($p<0.05$, $N=5$)*.

Patient ID	***r***
P0101	*0.672*
P0102	*0.377*
P0201	*0.332*
P0302	*-0.148*
P0702	*0.290*

have divided the day into four intervals, namely Morning (06 AM to 12 PM), Afternoon (12 PM to 06 PM), Evening (06 PM to 12AM) and Night (12 AM to 06 AM). Clearly, different patients will have different behaviour patterns as to what constitutes morning time, however the division of the day was setup in order to investigate whether at specific 6-hour intervals there is a higher correlation of physical activity and patient state.

5.3.2 Daily interval analysis

Once the days were divided in intervals, we investigated trends of physical activity levels in comparison to the patients' psychiatric evaluation. In order to normalize activity levels we have calculated the sum of all activity percentages in hourly basis for each day. This provides the average of activity level for each hour and each day. Separating the activities into hours allowed us to compare normalized average activity levels in different hours of the day. Motivated by the clinical work carried out in studying bipolar disorder patients in (Faurholt-Jepsen et al., 2012), where patients in depressive state have decreased morning activity levels, we examined association between morning Physical Activity (PA) levels and psychiatric scores, as shown in Table 5.2. Mean levels of PA in the morning had a noticeable difference when patients went from a depressive state to a normal state. This increase can be seen across all patients, although it is most noticeable for patient P0201, where the average increase in physical activity went from 16.17% during depressive state to 45.03% during normal state; and patient P0302 where the average PA increase went from 16.17% during depressive state to 45.03% during normal state.

The other two patients, P0102 and P0702 had a noticeable decrease of physical activity as they went from a normal state to a depressive state. As such there was a 55.20% decrease in physical activity levels for patient P0102 that went from normal state to severe depression (score of -3), while for patient P0702 the decrease in physical activity was 36.25% as he experienced a depressive episode with score of -2. For the patient that experienced a manic episode, P0101 we have seen a reverse trend, similar to the research reported in (Grunerbl et al., 2014). The average PA decreased from 5.70% during a manic episode to 1.14% during a mild manic episode as shown in Table 5.2. The reason that

Table 5.2: Relationship between overall physical activity (PA) and psychiatric assessment scores in depressive/manic episode (*n/a - not applicable since the patient did not experience a decrease in the assessment score).

P. ID	Nr. Evaluations	Average of PA in Depressive State (Score)	Average of PA in Improved State (Score)	PA Improvement Factor	Average of PA in Depressive State	Decrease of PA
P0102	5	12.49% (-2)	14.42% (0)	15.45%	6.46% (-3)	55.20%
P0201	5	16,17% (-1)	45.03% (0)	178.47%	38.7% (-3)	14.06%
P0302	5	4.42% (-3)	12.16% (0)	175.11%	n/a*	n/a*
P0702	3	15.42 (-2.5)	21.93% (-0.5)	42.22%	13.98% (-2)	36.25%

P. ID	Nr. Evaluations	Average of PA in Manic State (Score)	Average of PA in Improved State (Score)	PA Improvement Factor	Average of PA in Manic State	Decrease of PA
P0101	5	5.7% (+2)	1.14% (+0.5)	80%	n/a	n/a

recorded activity levels were low for the manic patient can be attributed to the fact that the usage of the phone for this patient was very low; which, incidentally, is one of the symptoms of mania. This was also confirmed from the recordings of phone usage logs (provided by the application), resulting in low amount of accelerometer data that was available for analysis.

5.3.3 Correlation of physical activity during daily intervals with psychiatric assessment scores

Previous section focused on morning activity levels and their relationship with the psychiatric assessment scores. However, we also wanted to investigate whether there is a correlation between physical activity levels during other daily intervals.In this respect we have calculated Pearson correlation coefficient between physical activity levels during each daily interval and psychiatric evaluation scores for all the patients. Results of the correlation are shown in Table 5.3.

Table 5.3: Correlation between patients' state and physical activity level during day intervals (p<0.05, N = 5, N** = 3) (*n/s not statistically significant result (p > 0.05) - *n/d not enough data recorded, due to phone being off)

Patient ID	Morning	Afternoon	Evening	Night
P0101	n/s*	0.315	-0.045	n/d*
P0102	0.581	-0.542	0.619	n/d*
P0201	0.261	0.586	0.243	n/d*
P0302	0.858	-0.842	-0.627	0.604
P0702**	-0.746	0.213	0.452	0.007

One of the interesting findings from analysing activities of these patients is that there is much stronger correlation between the individual daily intervals than there is for the overall activity levels (shown in Table 5.1). These results can be seen from patient P0102 where correlation with overall activity level is *r=0.377* whereas strongest correlation with daily interval is *r = 0.619* (Evening). A similar pattern emerges with other patients also, such as P0201, where the values are *r=0.332* for overall activity levels versus *r=0.586* for daily interval; P0302, with values *r =-0.148* (overall) vs *r=0.858* (interval); and, P0702 with values *r =0.290* (overall) vs *r=-0.746* (interval), where this patient had a strong negative correlation of physical activity levels with psychiatric scores.

One exception to this pattern is patient P0101, where correlation with overall activity levels is much higher (*r=0.672*) than the correlation with daily interval (*r=0.315*). Without a further research, we can only speculate on the reasons for these results. However, from the study group, this patient was the only one to have experienced a manic episode at the onset of the trial, with the state decreasing in severity towards the end of the trial.One speculative explanation may be that the patient's overall activity levels may have correlated well with their state, however due to missing data for the morning and night interval, it is impossible to understand whether those intervals may have affected the overall correlation score.

5.4 Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients

There is growing amount of scientific evidence that motor activity is the most consistent indicator of bipolar disorder. Motor activity includes several areas such as body movement, motor response time, level of psycho-motor activity, and speech related motor activity. Motor activity information can be used to classify episode type in bipolar patients, which is highly relevant, since severe depression and manic states can result in mortality. This chapter introduces a system able to classify the state of patients suffering from bipolar disorder using sensed information from smartphones. Further, we present the evaluation performance of several classifiers, different sets of features and the role of the questionnaires for classifying bipolar disorder episodes. Finally, we present our novel approach for observing of day-to-day phone conversation to classify impaired life functioning in individuals with bipolar disorder.

5.4.1 Monitoring motor activity behaviour in bipolar disorder patients

Motor activity is often used as a term to describe a group of symptoms that may range from mild to very severe, and is common feature of bipolar disorder. Assessing the motor activity of the patients with bipolar disorder has always been an essential part of psychiatric evaluations. Clinical measurement of motor activity is largely subjective and derives from caregivers' observations of specific behaviour. Motor functioning manifests itself in different areas such as speech production, facial expressions, gait, gestures, fine motor behaviour and the overall gross motor activity (Alderfer and Allen, 2002). Furthermore, motor agitation has been shown to be potentially disruptive in patients with bipolar disorder who are experiencing a manic episode, a period when patients have increased activity levels, pressed to incoherent speech, racing thoughts and a decreased need for sleep. Motor activity may also be present during mixed and depressive episodes of bipolar patients, which can be reflected in motor retardation and irritable periods of time (Faurholt-Jepsen et al., 2012). Therefore, monitoring motor activity is relevant for classifying critical state of the disorder. Smartphone is an enabling technology for this purpose due to increasing sensing capabilities.

Sensor data acquired from smartphones offers huge potential that through machine learning techniques get valuable insights of behaviour of bipolar disorder patients in their real life. In contrast to other studies, we show that mood episodes of bipolar patients can be predicted using only information obtained during phone calls.

To our knowledge, no research to date has focused on a naturalistic observation of the day-to-day relationship between motor activities during phone conversation and patients' mood episode in individuals with bipolar disorder. This current approach shows that motor activity features extracted from motion readings and speech articulation from smartphone sensors can be used to classify the course of mood episodes of a bipolar disorder patients. This is important because a non invasive and ubiquitous technology, like smartphones, can be used to obtain reliable information for patients during their phone conversations, in contrast to other studies using smartphone over long periods of time that can produce unreliable information when the phones are carried in purses, left at homes or use for playing or texting.

In following sections, we demonstrate the methodology used and features extracted in classification of motor activity in bipolar patients. The Figure 5.1 demonstrates two categories of features that were extracted, *i.e.*, speech features (prosodic, spectral) and intensity of phone handling during phone conversation (features in time and frequency domain).

5.4.2 Experimental results

This section shows four experimental results in order to validate our model to classify bipolar disorder episodes with the available data:

1. Comparing the performance of different classifiers on the data

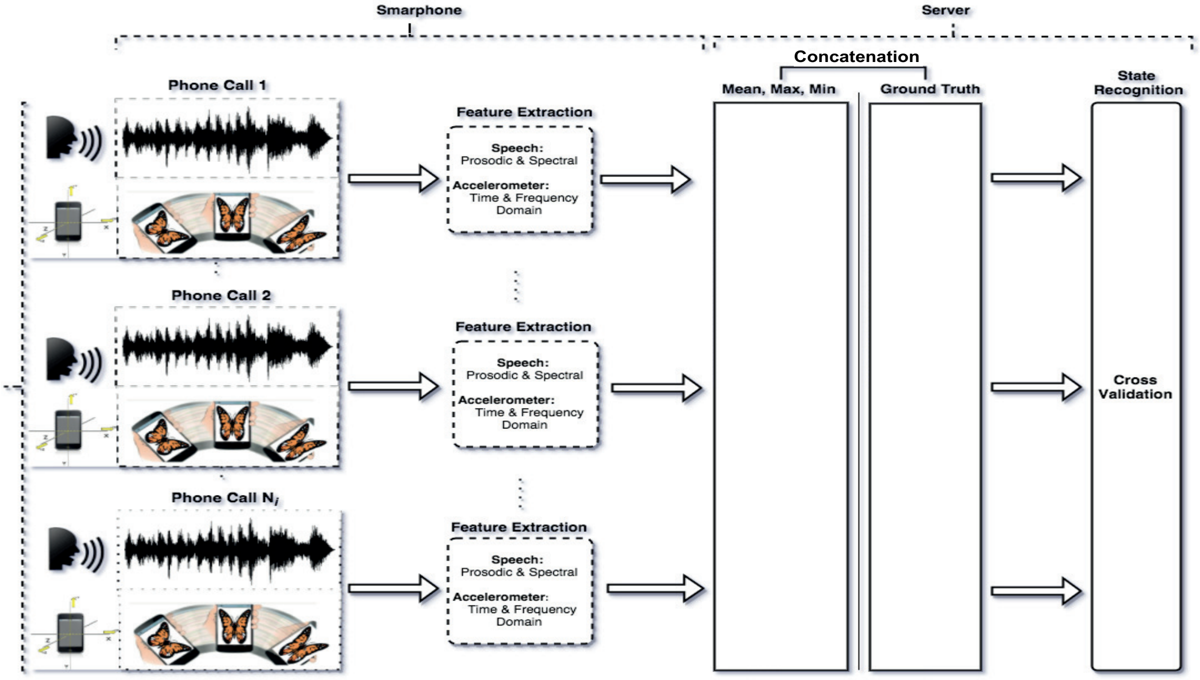

Smarphone
Server
Concatenation
Phone Call 1
Mean, Max, Min
Ground Truth
State Recognition
Feature Extraction
Speech: Prosodic & Spectral
Accelerometer: Time & Frequency Domain
Phone Call 2
Feature Extraction
Speech: Prosodic & Spectral
Accelerometer: Time & Frequency Domain
Cross Validation
Phone Call N$_i$
Feature Extraction
Speech: Prosodic & Spectral
Accelerometer: Time & Frequency Domain

Figure 5.1: Proposed approach for classifying motor activity in bipolar disorder patients.

2. Selecting a set of features appropriate to the given task
3. Assess the effect of the information from the questionnaires on knowledge of depression in patients
4. Use a semi-supervised learning methodology to address the problem on how to use information from unlabeled data to enhance classification accuracy of bipolar disorder episodes from the phone calls information and specify the relationship between labeled and unlabeled data from entire data set.

We learned a model for each patient and also a single model combining all the information from all the patients. We performed 10-fold cross validation for all the experiments and report global accuracy, precision and recall values for each of the episodes.

5.4.3 Experiments with different classifiers

Table 5.4 a) and Table 5.4 b) shows the results from using emotional and spectral audio features with frequency domain features from the accelerometers and with information from the questionnaires. Similar results were obtained with other sets of features.

The tables show the accuracy results for different classifiers for each patient, their average, and the results for a single model with information from all patients (last column

Table 5.4: Accuracy results from different classifiers taken from Weka with their default parameters.

a) Accuracy results from Frequency domain features and all Audio features.

Classifier	**P0201**	**P0302**	**P0702**	**P0902**	**P1002**	**Avg.(SD)**	**All P.**
C4.5	89.93	85.47	**78.81**	87.79	85.79	85.56 (±4.17)	76.50
Random Forest (RF)	87.25	84.62	70.86	89.44	83.76	83.19 (±7.24)	70.33
SVM	**92.28**	75.21	75.50	83.50	**87.82**	82.86 (±7.52)	69.99
Naive Bayes (NB)	71.81	62.39	61.59	62.71	78.17	67.33 (±7.35)	47.59
k-NN (1)	87.90	63.68	59.60	79.54	81.22	74.39 (±12.14)	69.43
AdaBoost.M1	84.56	**87.18**	74.17	**89.77**	86.80	84.50 (±6.06)	49.20
Bagging	89.26	86.32	71.52	89.44	86.29	**85.57** (±7.45)	**79.04**

b) Accuracy results from Frequency domain features and Spectral features.

Classifier	**P0201**	**P0302**	**P0702**	**P0902**	**P1002**	**Avg.(SD)**	**All P.**
C4.5	90.27	83.76	**78.81**	87.79	85.79	**85.28** (±4.35)	76.50
Random Forest (RF)	89.93	82.90	70.86	**90.43**	85.79	83.98 (±7.96)	**79.84**
SVM	**92.95**	75.21	76.16	83.83	**86.80**	82.99 (±7.44)	69.32
Naive Bayes (NB)	72.48	61.97	64.90	62.38	77.16	67.78 (±6.73)	46.83
k-NN (1)	87.58	63.38	61.59	77.89	81.73	74.43 (±11.46)	58.58
AdaBoost.M1	84.56	**87.18**	74.17	89.77	87.82	84.7 (±6.17)	49.20
Bagging	89.26	86.32	72.85	89.77	86.29	84.90 (±6.93)	78.95

Table 5.5: Accuracy results from using different sets of features.

Features	**P0201**	**P0302**	**P0702**	**P0902**	**P1002**	**Avg.(SD)**
Accelerometer:						
– Time Domain (TD)	89.53	73.08	72.19	83.83	85.28	80.78 (±7.73)
– Frequency Domain (FD)	89.93	83.76	75.50	87.71	85.79	84.54(±5.55)
Audio:						
– Emotional+Spectral	**90.60**	71.79	74.17	86.46	83.24	81.25 (±8.03)
– Spectral	**90.60**	73.93	74.17	87.12	83.24	81.81 (±7.55)
– Emotional	**90.60**	72.22	74.17	88.11	82.74	81.57 (±8.18)
– TD+Spectral	89.52	70.94	69.54	83.50	85.28	79.75 (±8.97)
– TD+Emotional	89.52	74.36	70.20	83.17	84.26	80.30 (±7.85)
– TD+(Emotional+Spectral)	89.53	70.09	69.54	82.18	85.28	79.32 (±9.07)
– TD+(Emotional+Spectral) without Questionnaire	50.0	59.40	70.86	51.16	81.22	62.53 (±13.37)
– FD+Spectral	90.26	83.76	**78.81**	87.78	85.79	85.28 (±4.34)
– FD+Emotional	90.27	**85.47**	74.83	87.79	**86.29**	84.93 (±5.93)
– FD+(Emotional+Spectral)	89.93	**85.47**	**78.81**	87.79	85.79	**85.56** (±4.18)
– FD+(Emotional+Spectral) without Questionnaire	79.53	84.19	74.83	**88.45**	85.79	82.56 (±5.40)

named ”All P.”). As can be seen from the table, there is no winning classifiers for all the data sets, although on average decision trees performed better than most other classifiers.

It also performed reasonably well with information from all the patients. For these reasons and in the rest of the experiments we only report results for C4.5.

5.4.4 Different Sets of Features

We tested different sets of features. In particular, using only accelerometer features (time domain vs. frequency domain), only audio features (emotional and spectral), and combining accelerometer features with different audio features. In all these results information from the questionnaires was also included. Table 5.5 shows the results.

As can be seen from the experiments, using only features from the accelerometers have results over 80% on average with the frequency domain features performing slightly better. It is also interesting to notice that the audio features have similar performance, when both types, emotional and spectral, are considered together or when tested in isolation. The best results are obtained when the spectral and emotional features from audio are combined with the frequency domain features from the accelerometers.

For the rest of the experiments, all the results will be presented only with frequency domain features combined with the spectral and emotional features.

5.4.5 Impact from the questionnaire

Assessing the impact on the results from the questionnaires is important for producing a fully autonomous application. This is relevant as self-assessment can be counterproductive for depressed patients as they are reminded every day, with the questionnaires, of their state of depression. We performed tests with the frequency domain features and the audio features with and without information from the questionnaires, and also using only information from the questionnaires (examples with only three features). Table 5.6 shows the results.

As can be seen there is a small decrement in the results obtained without information from the questionnaires, however, the average results are still over 82%, from which it is reasonable to think in the development of a fully automatic monitoring tool. From the table it can be seen that using only information from the questionnaires, produces very competitive results. It is interesting to notice, that in this case, two models (marked with "*") are simply a majority class classifier, which of course are very poor classified for individual class values.

5.4.6 Semi-supervised learning and motor activity classification

As described in Table 4.2, there are more than 900 phone calls without an associated episode. In this subsection, we decided to use a semi-supervised algorithm to see if we

Table 5.6: Accuracy results using information from questionnaires. Results with an ”*” indicate that the model is simply the majority class.

Patient	With Questionnaires	Without Questionnaires	Only Questionnaires
P0201	89.93	79.53	90.60
P0302	85.47	84.19	76.92
P0702	78.81	74.83	74.17 (*)
P0902	87.78	88.45	89.77
P1002	85.79	85.79	82.74 (*)
Avg.:	85.55	82.55	82.84
All P.:	76.50	60.78	59.76

Table 5.7: Accuracy results from a semi-supervised learning approach.

Patient	Supervised	Semi-Supervised
P0201	81.54	83.78
P0302	85.47	81.53
P0702	72.84	71.45
P0902	87.45	88.77
P1002	85.76	83.78
Avg.:	82.61	81.86
All Patients:	65.55	62.71

can improve on the performance of previous results using all the available data. We followed a simple approach where we divided the data into ten folds; the training data was used to classify the unlabeled data. This classification included a weight associated with the classified value. We then used all the classified data with the original training set to produce an extended training set. We created a model with this set and test it on the testing set and then we averaged the results over the 10 folds. The results are presented in Table 5.7.

As can be seen from the results, adding information from other calls is not making much difference in the final results. There is a large number of alternative semi-supervised algorithms that can be considered in the future to improve over these results.

5.4.7 Precision and recall

Although the overall accuracy results may look promising, it is important to analyse the individual precision and recall measurements to see how effective the constructive models are for each episode. Table 5.8 shows the results for each patient for Mild Depression (MiD), Moderate Depression (MoD), Severe Depression (SeD), Mild Manic (MiM), and Normal (N) state.

Table 5.8: Precision and recall results for some of the states of patients.

	Precision		**Recall**	
Patient (State)	**+Questionnaire**	**-Questionnaire**	**+Questionnaire**	**-Questionnaire**
P0201 **(Mild Depression)**	1.000	0.851	0.947	0.912
(Normal)	0.890	0.826	0.919	0.799
(Severe Depression)	0.640	0.455	0.667	0.795
P0302 **(Severe Depression)**	0.863	0.855	0.889	0.874
(Normal)	0.842	0.823	0.808	0.798
P0702 **(Moderate Depression)**	0.851	0.843	0.866	0.813
(Mild Depression)	0.595	0.512	0.564	0.564
P0902 **(Normal)**	0.892	0.899	0.876	0.882
(Moderate Depression)	0.862	0.869	0.880	0.887
P1002 **(Mild Manic)**	0.899	0.904	0.932	0.926
(Moderate Depression)	0.613	0.904	0.514	0.543
All Patients:				
– **(Severe Depression)**	0.725	0.649	0.447	0.415
– **(Moderate Depression)**	0.699	0.760	0.641	0.681
– **(Mild Depression)**	0.790	0.684	0.640	0.588
– **(Normal)**	0.790	0.836	0.633	0.633
– **(Mild Manic)**	0.824	0.809	0.613	0.654
Average:	0.766	0.765	0.606	0.608

It is interesting to note that most of the cases and both measures we have results above 80% with information from questionnaires and very similar results without information from questionnaires. We believe that these results give evidence that smartphone technologies can be effectively used as aid in the diagnosis of bipolar disorder episodes.

5.4.8 Predictive classes vs. Expert evaluation

The last set of experiments was designed to show how the inductive models to classify the prospective onset of episodes of patients for all the available phone calls. We show only figures for the best results obtained with patient P0201 (Figure 5.2) and for the worst results obtained with patient P0702 (Figure 5.3). Both figures show at the top the evaluation scores from the psychiatrist, in the middle the classified states from the model, and the bottom the weight or confidence in the class classified by the model.

As can be seen from the figures, the induced models follow closely the assessment of the experts (which is not surprising as the models were trained with this information), and make reasonable classifications in the intermediate states between psychiatric assessments. The figures also show, in red, the classification errors produced by the models. In particular, it can be seen that the models make few errors, which can be further reduced, if the classification with a weight less than a threshold value, *e.g.*, 0.8, are discarded.

Figure 5.2: Results from the induced model of patient P0201 and the assessments from the psychiatrist.

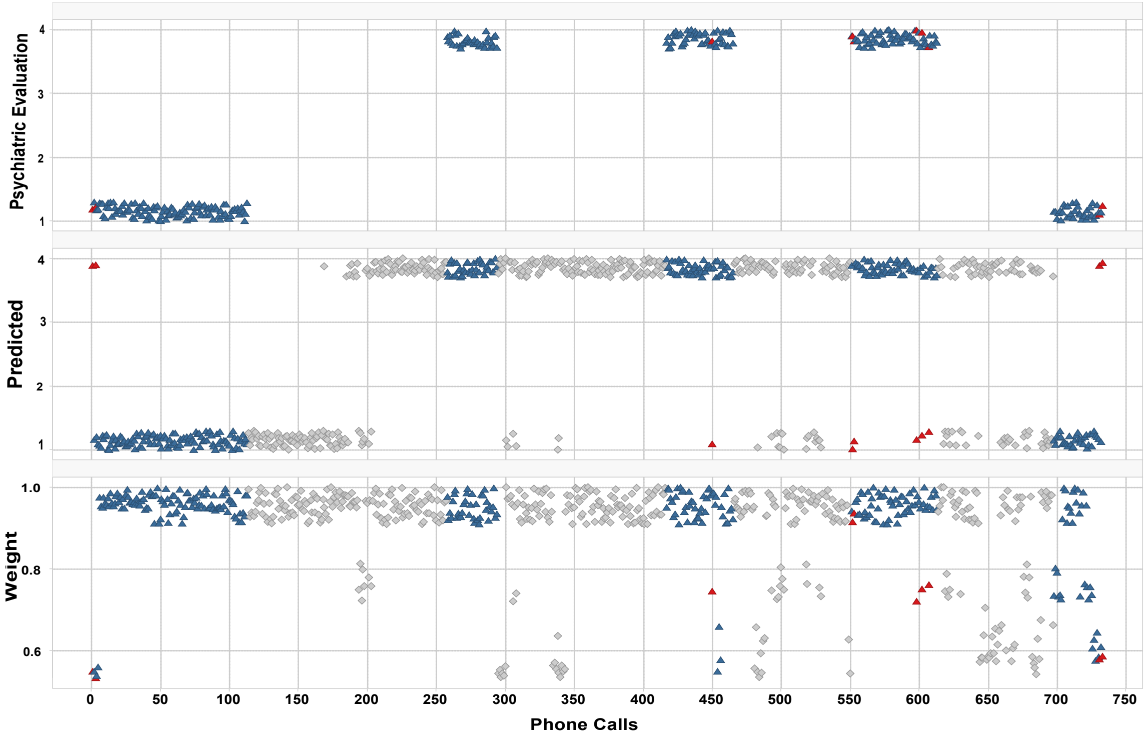
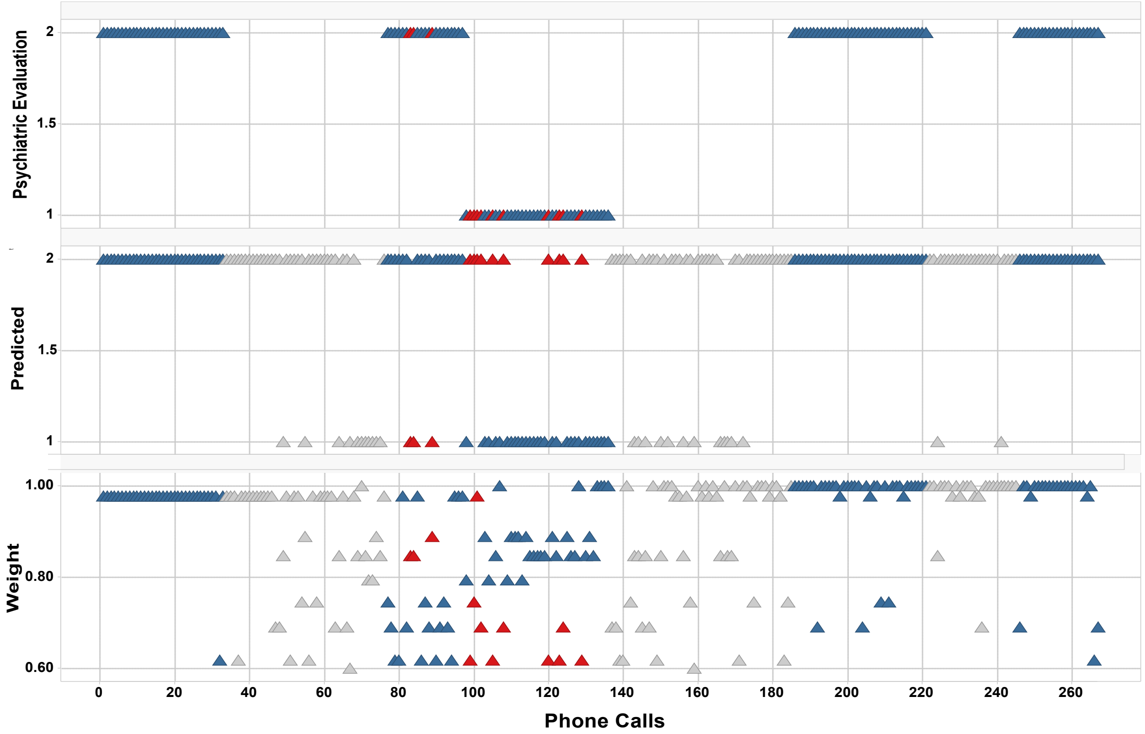


Figure 5.3: Results from the induced model of patient P0702 and the assessments from the psychiatrist.



5.5 Using motor activity and voice features with Intermediate Models in bipolar disorder

In the previous sections 5.4 we demonstrate the importance of analysing motor activity-related behaviour to classify the episodic state of the patients. We compared the standard supervised methods with semi-supervised learning methods. As a ground truth we used the psychiatric evaluation evaluated from the psychiatrist during their regular interviews. We demonstrated the problem of unlabeled instances between interviews and usage of semi-supervised learning method to address this problem.

In this research, we propose using the data derived from self-reported wellbeing questionnaires that are collected in daily basis. We propose using a novel intermediate models and the key advantage of using this approach is to improve the performance of supervised classifier. We build three intermediate models using the items recorded, *i.e.*, physical, activity, and the psychological condition to build the final model for classification of episodic state in bipolar disorder.

In the context of our research work, the following research questions are put forth:

- Is it possible to improve classification accuracy by incorporating intermediate *hidden* variables related to the patients' wellbeing, before building the final model for classification episodic state of the patients?

The present work tries to answer the research question by comparing measurements derived from questionnaires and motor activity-related behaviour during phone conversations.

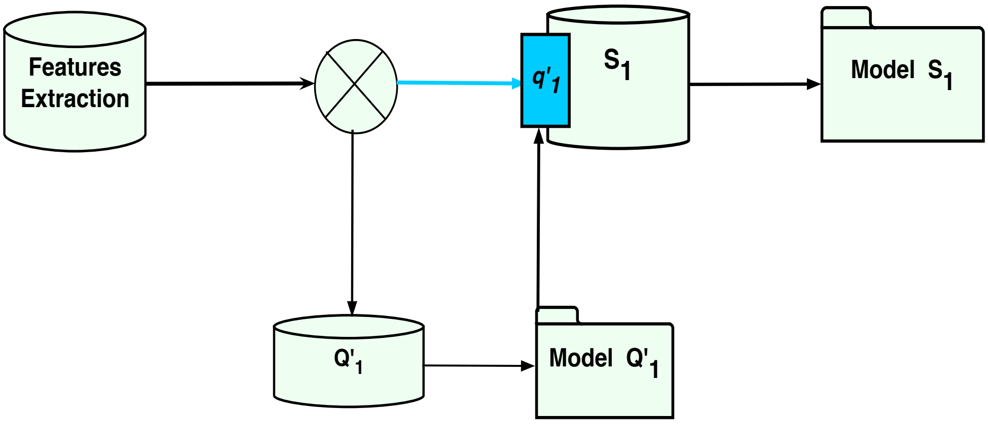
We performed an experimental analysis using real world data. The research includes 2 aspects:

- Using semi-supervised learning to complete the models for subjects with missing data.
- Using *Intermediate Models* to predict mood variables, which are incorporated in the final model with the aim at improving the accuracy of the predictions.

5.5.1 Experiments

Similar as in previous research work, we focus on analysing accelerometer raw data during phone conversation, where we are sure that the subjects are holding their smartphones. This type of measurement has the advantage of their availability and unobtrusiveness. We believe that analysing data collected from accelerometer readings during the phone conversations provide adequate information for classifying the trajectory of the episodic

Figure 5.4: Intermediate Models. Based on the accelerometer data from the smartphones, 30 frequency domain features are extracted. These are used to build the intermediate models for the mood variables, Q_1; and the model for stress, S_1. In the prediction stage both models are combine via a weighted linear combination to predict the stress level.



state changes. The second type of data that was analysed for this research includes the subjective information related to patients' physical, activity and psychological wellbeing.

5.5.2 Intermediate Models

The information provided by the patients through the questionnaires is very useful, however, it is a tedious task for the patients. In this research we propose to predict the wellbeing-related variables associated to the questionnaires using the data from the smartphone to alleviate the patients from this burden. We then use the predicted mood variables with the rest of the data from the smartphones to classify the episodic state of the patients. We call the models that predict psychological and wellbeing conditions variables from the questionnaire *Intermediate Models* as they are used as input for the final predictive model. Although the use of additional variables, such as latent variables (Li et al., 2009), have been previously used in the literature, we are not aware of research that aims at building an intermediate model that can then be used as input for the final model. Figure 5.4 illustrates the procedure for building the intermediate models.

In this research, we used three variables derived from physical, activity and psychological condition to build 3 intermediate models. We train each classifier separately using each the self-reported questionnaires derived from the daily self-assessment. In the prediction stage, the intermediate models use the information from the smartphones to predict a weighted set of wellbeing conditions based on the accuracy of each model. Then all the data from the smartphones and the mood variables are used as input for the final episodic state model.

5.5.3 Experimental results

Our experiments have the following objectives

- Compare the performance of different classifiers on the data.
- Assess the effect of Intermediate Models to enhance the knowledge of self-reported psychological condition in bipolar disorder patients.
- Use SSL to address the problem on how to use information from unlabeled data to enhance classification accuracy.

For all the experiments, we used Weka's (Hall et al., 2009) classifiers with their default parameters. We build a model for each subject and performed a 10-fold cross validation for all the experiments; we report the global accuracy and precision values.

5.5.4 Experiments with different classifiers

In previous research, we demonstrated that the information obtained from the frequency domain features of the accelerometers lead to higher classification accuracy combined with all audio features. In Figures 5.5 [2], we use the extracted features from frequency domain with all audio and spectral features. The result are compared using supervised and semi-supervised learning using the approach with intermediate models. As can be seen from the tables, the C4.5 are the winning classifier for all the data sets. Using semi-supervised methods have been shown on average decision trees performed better than the most other classifiers. In the following experiments we only report result from C4.5.

5.5.5 Different sets of features and Intermediate Models

Different set of features using the Intermediate Models has been tested. Different set of features derive from accelerometer features in frequency domain, all audio features (emotional and spectral), and combining frequency domain features with emotional or spectral features. The results' information are shown in the Tables (Table 5.9 and Table 5.10).

Using only features from the accelerometer in frequency domain have results over 81% on average. However, using only the features derived from audio the performance lower than using accelerometer features. The best results are obtained when the spectral and emotion features from audio are combined with the frequency domain features from the accelerometers.

In Figure 5.5 we demonstrate the results after using a semi-supervised learning algorithm with the aim to improve on the performance of other previous results using all

[2] More details about motor accuracy are shown in Tables A.1, A.2, A.3 and A.4

Figure 5.5: Accuracy results from accelerometer frequency domain features and all audio features.

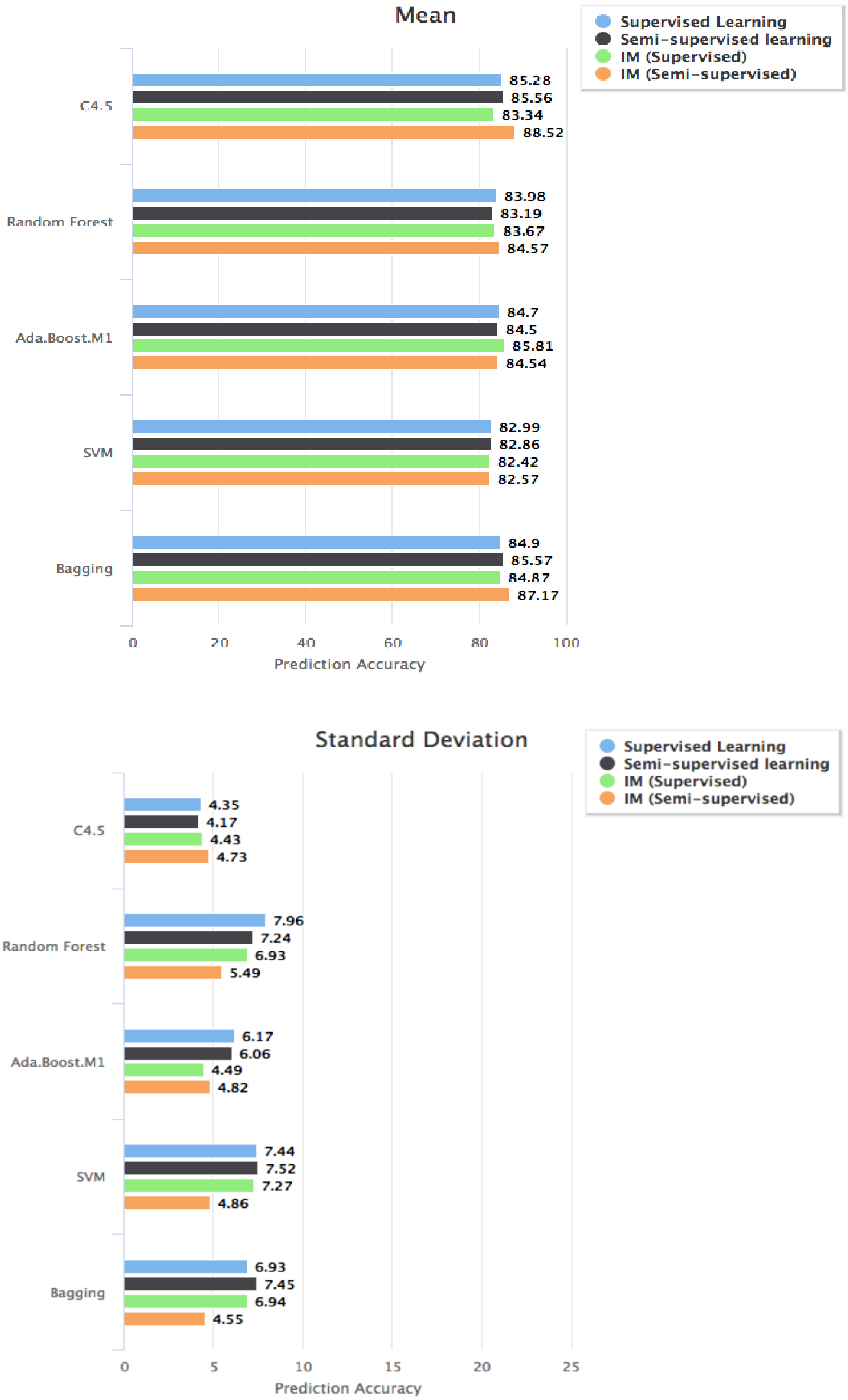


Table 5.9: IM: Accuracy results from using different sets of features.

Features	**P0201**	**P0302**	**P0702**	**P0902**	**P1002**	**Avg. (±SD)**
Accelerometer:						
– Time Domain (TD)	59.12	61.11	69.54	56.16	80.71	65.33 (±9.93)
– Frequency Domain (FD)	81.54	84.19	69.54	86.80	**85.79**	81.57 (±7.01)
Audio:						
– Emotional+Spectral	56.71	55.56	67.96	53.82	78.36	62.48 (±10.47)
– Spectral	55.70	67.16	67.96	53.82	79.85	64.90 (±10.55)
– Emotional	59.73	64.93	**73.79**	56.63	79.10	66.84 (±9.45)
– TD+Spectral	59.80	58.97	70.86	60.13	79.70	65.89 (±9.13)
– TD+Emotional	62.16	57.69	70.86	77.83	79.70	69.65 (±9.60)
– TD+(Emotional+Spectral)	60.81	58.55	70.86	60.61	78.17	65.80 (±8.41)
– FD+Spectral	**86.91**	84.62	70.86	**96.47**	83.76	**84.52 (±9.16)**
– FD+Emotional	82.55	84.62	70.86	95.22	83.76	83.40 (±8.65)
– FD+(Emotional+Spectral)	81.54	**85.04**	70.86	92.72	84.77	82.99 (±7.93)

Table 5.10: Intermediate models and semi-supervised learning - Accuracy results from using different sets of features.

Features	**P0201**	**P0302**	**P0702**	**P0902**	**P1002**	**Avg. (±SD)**
Accelerometer:						
– Time Domain (TD)	72.12	76.91	77.15	56.16	84.00	73.27 (±10.46)
– Frequency Domain (FD)	89.73	84.19	77.15	86.80	85.33	84.64 (±4.67)
Audio:						
– Emotional+Spectral	76.35	71.13	80.33	65.48	79.74	74.61 (±6.28)
– Spectral	71.77	82.43	84.70	73.60	81.05	78.71 (±5.69)
– Emotional	64.08	75.52	**85.25**	56.63	82.35	72.77 (±12.16)
– TD+Spectral	76.76	66.74	85.02	60.13	76.44	73.02 (±9.69)
– TD+Emotional	73.96	67.44	89.03	77.83	76.89	77.03 (±7.84)
– TD+(Emotional+Spectral)	83.85	68.13	80.90	60.61	74.67	73.63 (±9.46)
– FD+Spectral	**94.88**	**90.99**	85.02	**96.47**	**87.11**	**90.89 (±4.89)**
– FD+Emotional	89.84	**90.99**	82.77	95.22	**87.11**	89.19 (±4.63)
– FD+(Emotional+Spectral)	**90.88**	90.99	80.90	92.72	**87.11**	88.52 (±4.73)

the available data. It also interesting to notice that using semi-supervised methods with models built using intermediate models have achieved better accuracy in comparison with other methods used so far. In contrary to previous work in previous Section 5.4, using the intermediate models has been shown to improve the accuracy (as shown in Table 5.9 and Table 5.10) were we add information from unlabeled phone calls and increase the performance accuracy (yielded accuracy of ≈ 90%).

5.6 Chapter Summary

In this chapter we have presented how to classify the course of mood episodes of bipolar disorder patients from information extracted from smartphones during phone conversation. We used information from patients during 12 weeks on unconstrained conditions. We considered a wide range of features, both from accelerometer information and from audio information during the phone calls and analyse their behaviour for different users and mood episodes. We also make a comparison of different classifiers and different sets of features.

5.6.1 Semi-supervised learning in classification of bipolar disorder

In this research 5.4, the information obtained from the frequency domain features of the accelerometers lead to higher classification accuracy than the information extracted from audio. Also, the frequency domain features produced better classification results than the time domain features. When we combined the audio features with the accelerometer features, there was only a small improvement when the emotional and spectral features were included. Adding information from the questionnaires improved the overall results and also showed good results when considered on their own. However, without information from the questionnaires we obtained reasonable results ($>80\%$ for accuracy, precision and recall), suitable for the development of automatic tools that could aid psychiatrists in the monitoring of their patients.

5.6.2 Intermediate models in classification of bipolar disorder motor activity

In this research we presented a new novel method, namely Intermediate Method used to classify the course of mood episodes of bipolar disorder patients from the accelerometer and voice features extracted from smartphones during phone conversations. Involving the self-reported wellbeing for building the intermediate models has been shown to improve the accuracy for classifying bipolar disorder episodic states. Further, we make a compassion of different classifiers and different set of features. Similarly, as in previous section, using the information obtained from the frequency features of the accelerometers lead to higher classification accuracy than the information extracted from audio signals. Combining the data from audio features with the accelerometer features, there was an improvement when the spectral were included.

The proposed methods using the Intermediate Models and Semi-Supervised learning methods has been shown to improve the overall results. Although relying (only) on psychological evaluation information (we obtained reasonable precision from $\approx 73\%$ to $\approx 90\%$), using the information from self-reported questionnaires on the smartphone suggest for the development personalized models with small labeled dataset would be suitable for the automatic behaviour changes recognitions that could aid psychiatrist in the nearest future in the monitoring of their patients as they go in their daily life.

Chapter 6

SCARCE DATA AND IMPROVEMENT OF STRESS PREDICTION

> *"If we can reduce the cost and improve the quality of medical technology through advances in nanotechnology, we can more widely address the medical conditions that are prevalent and reduce the level of human suffering."*
>
> – **Ralph Merkle**

The key message in the previous chapter is that current sensing systems are promising the near future in healthcare services and together with ML techniques are improving the diagnostic accuracy in mental-health. In this chapter, we begin with a brief introduction of the study setup and the features extracted to building a classification model. Further, we use several ML methods to predict the perceived work-related stress on the data acquired from employees in their real-working environments. Finally, we frame the challenges facing the building of accurate models for stress detection.

The contributions of this chapter are as follows:[1]

[1] This chapter is mainly based on the following research work:
I. Maxhuni, A., L. Hernandez, E. Sucar, V. Osmani, E. Morales, and O. Mayora, *"Stress Modeling and Prediction in Presence of Scarce Data"*, Elsevier Journal of Biomedical Informatics, 2016, Journal Article.
II. Maxhuni, A., P. Hernandez-Leal, E. M. Manzanares, E. Sucar, A. Muñoz-Melendez, and O. Mayora, *"Using Intermediate Models and Knowledge Learning to Improve Stress Prediction"*, FI-eHealth, Puebla, Mexico, EAI, May, 2016, Conference Paper.
III. Hernandez-Leal, P., **Maxhuni, A.**, Sucar, L. E., Osmani, V., Morales, E. F., and Mayora, O. (2015, December). Stress Modeling Using Transfer Learning in Presence of Scarce Data. In Ambient Intelligence for Health (pp. 224-236). Springer International Publishing.

A.1 *We propose using "Semi-supervised learning" methods to cope with scarce data from the subjects in our research.*

A.2 *"Transfer learning" method is proposed to transfer information from other models to our target model which contains insufficient data to produce an accurate one.*

A.3 *We propose using "Ensemble learning" methods to build multiple models to obtain better predictive performance than could be obtained from any single model.*

A.4 *We have evaluated the datasets comprising 30 subjects; we measure the robustness of our proposed methods to address the problem scarce data and improve the accuracy for classification of perceived stress.*

The outline of this chapter is as following: the current Section in 6.1 describes the problem statement related to stress assessment. This section defines the conditions, such as feature extractions, classification problems, classifications methods of the research carried out in this chapter. Section 6.2 proposes using machine learning methods (i.e., Semi-supervised learning, Transfer learning, Ensemble learning) for stress modeling and prediction in presence of scarce data. Finally, in the Section 6.3 we investigate the information about user's motor activity-related behaviour while having conversation on the phone toward less obtrusive method for stress detection.

6.1 Stress assessment

Nowadays, social competition is becoming increasingly stronger, which together with the rapid economic transformation have changed the dynamics of workplace environments. Due to these changes, enterprise employees are experiencing a period of intense job-insecurity, increased work-loads, and long working hours. All these factors are known to engender work-related stress of different degrees, affecting the physiological and psychological functioning of the employees. According to recent reports from the European Agency for Safety and Health at Work - EUOSHA (Milczarek et al., 2009), stress was found to be the second most common work-related health problem across 27 Member states of the European Union (EU). Overall, 22% of EU employees reported work-related stress.

Furthermore, it is also demonstrated that long-term exposure to stress can lead to many serious health problems, causing physical illness through its physiological effects (*e.g.*, fatigue, decreased sleep quality), behaviour changes (*e.g.*, addiction, attention deficit), and social isolation issues (*e.g.*, anger) (Bongers et al., 1993; Glanz et al., 2008; Korabik et al., 1993; Maslach et al., 2001; Paoli, 2003; Sultan-Taïeb et al., 2013). As a

IV. Maxhuni, A., L. Hernandez, E. Sucar, V. Osmani, E. Morales, and O. Mayora, *"Stress Assessment Using Smartphones"*, 2016, Journal Article. (in review) consequence, these negative effects have been shown to decrease wellbeing at workplace and employees' work effectiveness. Moreover, long-term exposure to stress typically leads to job-burnout, a state that leads to mental and physical exhaustion (Maslach et al., 2001). For the reasons previously mentioned it is important to measure stress as a way of monitoring individual's wellbeing. However, unlike other mental and physical problems, stress is not easy to measure (Occupational Safety and Stress, 1999). Thus, its assessment represents a current open problem.

Measuring physiological dynamics has become a challenging issue, from both research and clinical practice. To date, physiological measurements and self-reported questionnaires are the most common methods used to infer work-related stress. However, only very limited research has been directed in detecting psychological factors deriving from behavioural dynamics that connotes psychological functions at workplaces. Therefore, monitoring the affect changes of employees and other personality traits (*e.g.*, behavioural aspects) should be of great interest for both healthcare institutions and organisations.

A number of studies have investigated detecting stress and emotions based on facial expressions (Valstar et al., 2011). Other mood and stress detectors have used individual physiological parameters. These include heart rate and the galvanic skin response (GSR) (Bakker et al., 2011; Muaremi et al., 2013). Lastly, other studies have analysed voice acquired from individuals to detect stress in laboratory or clinical settings (He et al., 2009). However, their limitation is that laboratory settings are often an inadequate environment compared to the complexity of real-day environment monitoring at diverse scales (*i.e.*, physically and socially). In this regard, another aspect that has to be considered when it comes to long-term monitoring, is that sensors have to be as least intrusive as possible trying to minimize the impact on workers' routines and their natural behaviour.

Smartphones are becoming more powerful (in terms of sensors capabilities) and every year the number of these devices is increasing. For these reasons, smartphones are excellent candidates to be used for monitoring everyday activities including activities in working environments. Thus, the challenge is to use the sensor capabilities of the smartphones to detect stress-related behaviour of a person in an unobtrusive manner. Then, this could be communicated to the person in order to take pre-emptive actions and alleviate high stress levels (Sanches et al., 2010).

Several factors can affect employees' stress at work, however our approach focuses on behaviour changes that can be directly measured using smartphones: location changes, physical activities, social interactions and phone application usage. In this section we demonstrate our objective aiming at detecting behaviour changes using only information obtained from smartphones and investigate their correlation with perceived stress levels.

The following research questions were put forth:

- Is there a correlation between the subjects' behavioural characteristics, extracted from smartphone sensor data, and their self-reported stress levels?
- Is it possible to improve prediction accuracy of work-related stress based on smartphone sensor data by combining limited labeled data and unlabeled data?

6.1.1 Correlation between objective and Self-reported emotions data

We conducted two correlation analyses to investigate the association between four factors: perceived stress, negative-mood, positive-mood, and overall mood score. Emotions were divided in two categories: negative-mood (tense, angry, anxious and sad) and positive-mood (friendly, energetic, cheerful and being good at current activity). As presented in the Chapter 4, mood items were rated on a 5-point scale established by "low or not at all" (1) to "high or very much so" (5), similar to research work in (Lutgendorf et al., 1999) using POMS model of mood assessing. An overall score derived from both types of emotions was obtained by subtracting negative mood scores from positive scores.

6.1.2 Pearson correlation in stress events

A *two-tailed Pearson* correlation and multiple linear regression analysis were performed to examine the relationships among perceived stress and wellbeing (moods) scores with objective measurements. First, we performed the correlation tests between objective and subjective variables. The Pearson correlation coefficient ρ was used, we take as statistically significant when ρ *<0.05 (*)* and ρ *<0.01 (**)*. In Table 6.1[2] we present the correlations between objective measurements (rows) and subjective measurements derived from self-reported stress, negative-mood score, positive-mood score, and overall-mood score (columns) and we can make some observations:

- For stress level,

– Physical activity level:	**r= *-0.153***, $^{**}\rho$ **<0.01**, *N=1465*
– Number of system Apps:	**r= *-0.129***, $^{**}\rho$ **<0.01**, *N=1292*
– WiFi location:	**r= *-0.087***, $^{**}\rho$ **<0.01**, *N=1456*
– Cellular location:	**r= *-0.070***, $^{**}\rho$ **<0.01**, *N=1456*
– Number and duration of Outgoing calls:	**r= *-0.098***, $^{**}\rho$ **<0.01**, *N=1120*
– Number and length of SMS responses:	**r= 0.090**, $^{**}\rho$ **<0.01**, *N=505*

  obtained statistically significant correlations.
- In particular, for missing calls we expected to have correlation with different factors. We assume that during a stress-full day, participants are more prone to reject

[2] More details related to correlation are presented in the Table A.8

the phone-conversations due to responsibilities and task that they have to achieve. However, it was shown to have a weak correlations with the stress factor.

- Negative emotions show high correlation with accelerometer, number of system apps, social interaction information and social-activeness (number of incoming and outgoing phones calls, and outgoing SMS's).
- Social interaction information and the use of social applications showed high correlation with positive mood scores.
- Information from social applications and location obtained low correlation with negative emotions. This is interesting because these same two variables obtained high correlation with positive emotions. Similarly, the number of incoming calls show low correlation with positive emotions but is highly correlated with negative emotions.

Table 6.1: Pearson correlations between objective variables and Perceived Stress Level, Negative Mood Score, Positive Mood Score, and Overall Mood Score.

Objective Variables	**Stress Level**	**Negative Mood**	**Positive Mood**	**Total Mood Score**
Physical Activity Level	*-0.153***	*-0.112***	*0.071***	*0.116***
Cellular Locations	*-0.070 **	*-0.070**	0.033	*0.065**
Google-Maps Locations	0.051	0.017	*0.079**	0.033
Wifi Locations	*0.087***	0.039	*-0.120***	*-0.093***
Social-Interaction	0.032	*0.059**	*-0.142***	*-0.119***
Number-Outgoing-Calls	*-0.980***	*-0.112***	*0.083***	*0.121***
Number-Incoming-Calls	-0.005	*-0.090***	-0.019	0.05
Missed-Incoming-Call	-0.006	-0.023	-0.012	0.009
Duration-Outgoing-Call	*-0.098***	*-0.097***	*0.101***	*0.123***
Duration-Incoming-Call	0.037	-0.034	*0.091**	*0.074**
Number-SMS-Outgoing	*0.090***	*-0.071**	0.004	0.05
Number-SMS-Incoming	0.006	-0.012	-0.044	-0.016
Length-SMS-Outgoing	*-0.154***	*-0.153***	*0.106**	*0.156***
Length-SMS-Incoming	0.013	-0.028	*0.088**	0.069
Duration-Apps-System	0.008	-0.021	-0.024	0.001
Duration-Apps-Social	0.067	0.067	*-0.218***	*-0.161***
Number-Apps-System	*-0.129***	*-0.181***	*0.194***	*0.228***
Number-Apps-Social	-0.060	-0.040	-0.004	0.024

– Significant at the level: *ρ <**0.05**; **ρ <**0.01**.

6.1.2.1 Multiple regression analysis

In order to obtain the best possible model for prediction of stress and total mood score we decided to use multiple linear regression. We found that regression result was significant for stress (r^2*=0.3912, F(18,64)=2.28,* ρ *<0.008*) and with total mood-scores (r^2*=0.4419, F(18,64)=2.81,* ρ *<0.001*) using all features (as shown in Table 6.2, which depict the

Table 6.2: Significant results from the multiple regression using objective measurements with respect to Stress and Total Mood Score.

	Stress			**Total Mood Score**		
Objective Variables	β	**t**	ρ	β	**t**	ρ
Physical-Activity Levels	-.0111	-5.88	***0.001***	-.0111	-5.88	***0.001***
Cellular Location	-.2333	-2.29	***0.022***	.0376	2.10	***0.036***
Google-Maps Location	.0685	1.65	0.100	.0077	1.06	0.289
WiFi Location	.0057	3.34	***0.001***	-.0041	-3.58	***0.001***
Social Interaction (SI)	.0001	1.13	0.258	-.0008	-4.28	***0.001***
Number-Outgoing-Calls	-.0374	-3.31	***0.001***	.0081	4.07	***0.001***
Number-Incoming-Calls	-.0033	-0.17	0.866	.0058	1.68	0.093
Missed-Incoming-Call	-.0015	-0.19	0.847	.0004	0.29	0.769
Duration-Outgoing-Call	-.0125	-2.73	***0.006***	.0026	3.43	***0.001***
Duration-Incoming-Call	.0048	1.01	0.313	.0016	2.02	***0.044***
Number-SMS-Outgoing	.0188	3.05	***0.002***	.0018	1.68	0.092
Number-SMS-Incoming	.0003	0.19	0.850	-.0001	-0.54	0.590
Length-SMS-Outgoing	-.0015	-3.49	***0.001***	.0003	3.55	***0.001***
Length-SMS-Incoming	.0001	0.34	0.737	.0001	1.72	0.086
Duration-Application-System	.0001	0.31	0.759	.0001	0.03	0.976
Duration-Application-Social	.0001	1.43	0.153	-.0001	-3.47	***0.001***
Number-Application-System	-.0061	-4.69	***0.001***	.0020	8.42	***0.001***
Number-Application-Social	-.0189	-1.27	0.203	.0014	0.51	0.610

Significant at the level: ρ **<0.05**; ρ **<0.01**.

name of each feature, the regression coefficient, β, the distribution value, t, the and ρ-value for each used feature). This results show that selected features are having an effect on predicting stress (ρ <*0.008*) and total mood score (ρ <*0.001*). Similarly, several objective variables (with italic typeface in Table 6.1) show significant correlation with perceived stress and total mood score of the subjects. It is interesting to note that these objective variables (physical activity, cellular and Wifi location, number and duration of outgoing calls, number and length of outgoing SMSs, and number of applications) also show significant linear correlation using Pearson.

To summarize the correlation results:

- Stress level is highly correlated with physical activity, WiFi location, number and duration of outgoing calls and SMS, and with social apps. These values are consistent with what was obtained with multiple linear regression.
- In contrast, negative mood is highly correlated with the number of incoming calls and is not correlated with WiFi location.
- Similarly, positive mood is highly correlated with social interaction and duration of social apps but it is not correlated with the number of outgoing SMS.

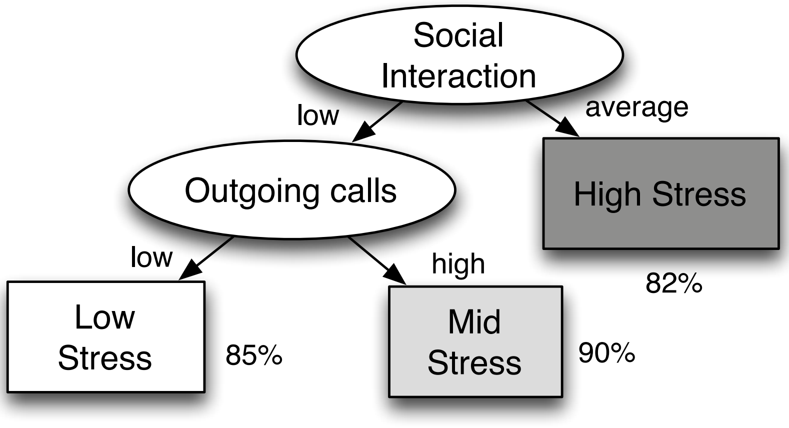


Figure 6.1: An example of a decision tree, each oval represent a decision node which contain arrows to other decision nodes. Squares are leaves (terminal nodes) that give the classification value, in this case they represent Low, Mid or High level of stress.

6.1.3 Stress prediction as classification problem

In the previous section we analysed the relation between the measured objective variables with perceived stress. We presented results showing many features correlated with stress levels. Thus, our next step is to make a model capable of predicting the stress level given the objective variables.

Predicting perceived stress of the user can be seen as a classification problem. In this case, the attributes correspond to each feature related to the objective variables and the class to predict is the self-reported stress level (low, moderate, high). Since we are interested in analysing behaviour changes or patterns that may appear in daily activities, we used decision trees which can be easily understood. Our approach was to build a decision tree for each subject of the study, with the idea of analysing individual behaviours and models.

As we mentioned in previous chapter, an important benefit of decision trees is that they can be easily understood, for example obtaining rules to be further analysed. In Figure 6.1 we present a decision tree that classifies the stress level of a subject in the research work. The subject shows low levels of stress when having an average level of social interaction, or when the social interaction and number of outgoing calls is low. On the contrary, if this subject had low level of social interaction but a high number of outgoing calls then it is more probable to have a mid level of stress.

We performed classification of the stress variable using the C4.5 algorithm (Quinlan, 1993) and 10-fold cross validation for each user. Table 6.3 presents the classification accuracy of stress level for the 30 subjects. In average the accuracy obtained was 67.57%.

However, we noted that dataset contained 20% of missing data. This is an important portion which can be exploited with a SSL technique.

Table 6.3: Stress Prediction using decision trees before and after applying a Semi-supervised learning approach. Overall classes represent overall number of labeled instances derived from self-reported stress in supervised learning and after performing semi-supervised learning methods.

Subjects (30)	**Supervised**	**Semi-Supervised**	**Increase**
Accuracy (%)	67.57 ± 15.60	**71.73 ± 15.25**	4.20 ± 9.52
Overall Classes (%)	1465/1832 (79.97)	**(1722/1832) (94.00)**	14.03
Precision(%)	65.4	**68.9**	3.5
Recall(%)	68.9	**73.0**	4.1
F-Score (%)	66.0	**70.0**	4.0

6.1.4 Semi-supervised learning (SSL)

In most real-world datasets it is common to have missing data. The most basic approach is to ignore those instances. However, that information even when is not complete can be helpful and should not be discarded. Semi-supervised learning (Longstaff et al., 2010; Zhu, 2006) has been suggested as a method aiming to address this issues in machine learning. The main objective of semi-supervised learning is to learn from both labeled and unlabeled data, *i.e.*, by exploiting unlabeled samples to improve the learning performance.

For this research we consider one of the most common methods of SSL that uses a single classifier called Self-Training (Zhu, 2006). It works by selecting the most confident unlabeled points, together with their predicted labels and then adding those to the training set. In each iteration the newly high-confidence (>80%) labeled instances are added to the original labeled data. Note that the classifier uses its own predictions to teach itself. The classifier is re-trained and the procedure repeated (see Algorithm 2).

In Table 6.3 we present the results in terms of accuracy after applying the SSL approach on all subjects in this research. Using the Self-Training method, we were able to improve the accuracy on predicting stress to *71.73% (+4.20%)*. In Table 6.3 we demonstrate that using Self-Training we were able to reduce the number of missing classes from 20% to 6%. We have also analysed accuracy results by gender. Results show that the *Male* achieved better accuracy 72%(*Precision: 73.5%; Recall: 78.5%*) for supervised approach and 76.4% (*Precision: 73.5%; Recall: 78.5%*) for SSL, in contrast to *Female* with 59.8%(*Precision: 59.0%;Recall: 60.0%*) for supervised and 64.8% (*Precision: 62.0%; Recall: 65.0%*) when using SSL approach.

In this section we have shown that simple models can be generated to predict stress levels with around 70% of accuracy. Unsurprisingly, most of the models used the relevant features identified in the previous section. It is also shown that a slight improvement in the predictive performance can be achieved with a simple semi-supervised learning algorithm. It is left as future work to use other more powerful classifiers and semi-supervised techniques.

6.2 Stress modeling using transfer learning in presence of scarce data

The objective of this research is to model stress levels from different behavioural variables obtained from smartphones and in particular with the limitation that the labeled data for a person is scarce. This scarcity of data is a common problem while monitoring humans *in situ* and requires constant annotation of their current wellbeing, as the data derived from self-reports are considered as a ground truth. From the collected data we extracted several features such as physical activity level, location, social interaction and social-activity. In order to deal with scarce data, common to many real-world applications, we apply two machine learning techniques, namely, semi-supervised learning, to be reduce unlabeled data, and transfer learning (Pan and Yang, 2010) to use previously learned models to improve the model of a person with scarce data.

The proposed approach learns a model for each subject participated in a study. This approach is useful not only to predict the stress levels but also to perform comparisons among different subjects in order to obtain groups of people (clusters) that behave similarly. Moreover, when a model is built for a new subject it usually contains insufficient information to have an accurate model. For this reason we use a transfer learning approach that uses data from similar subjects in order to improve the target model, which results in better prediction results.

Our research addresses 4 aspects:

1. Using semi-supervised learning to complete the models for subjects with missing data.

2. Clustering the subjects based on the similarity of the learned decision trees.

3. Applying transfer learning to improve the model of a new user with scarce data.

4. Using ensemble methods to improve the accuracy of the models.

To the best of our knowledge, few works have dealt with scarce data even when this is a common challenge in health research, most often founded in studies where participants

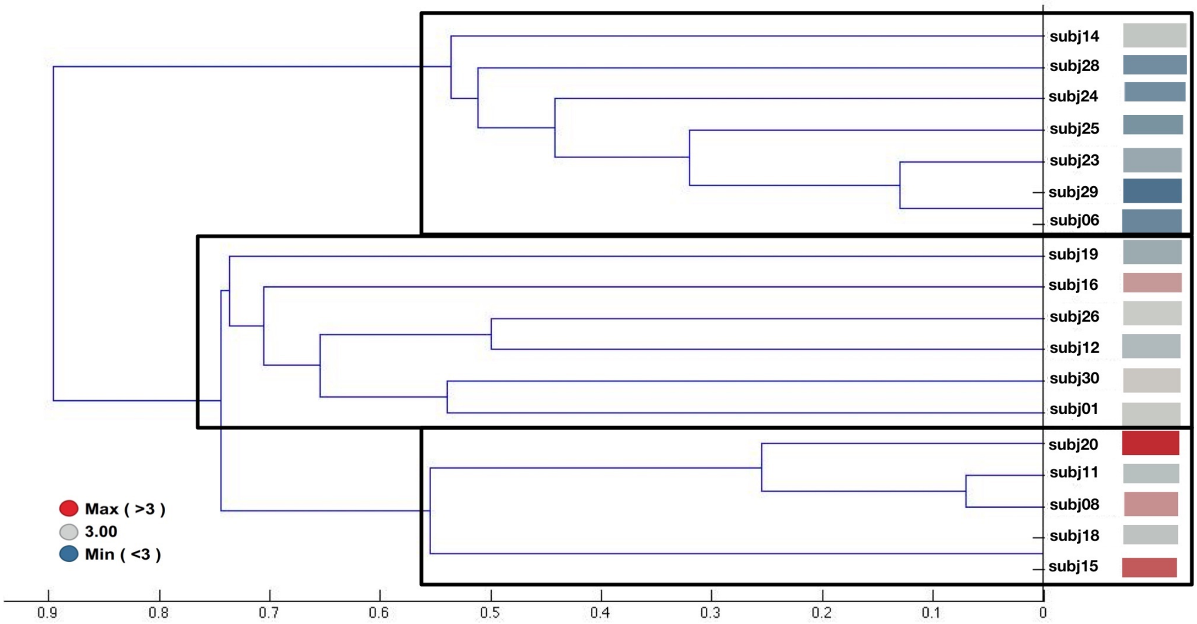


Figure 6.2: Dendrogram obtained by computing similarities between models of each subject (using only 18 subjects). Three major clusters can be noted, colour boxes correspond to average stress for different subjects (best seen in colour).

use self-report instruments.

6.2.1 Modeling Stress

Predicting perceived stress of a person can be modeled as a classification problem. We used decision trees (Quinlan, 1993) to model subject's stress since this representation can be easily understood by a human, and this could help to have a better understanding of what causes stress. Also, using this representation we can compare different subjects, which is important for transfer learning. Our approach is to build a decision tree, a model to predict stress, for each subject of the study. To learn decision trees we used the C4.5 algorithm using as attributes the objective variables presented in Chapter 4 and the class to predict is the self-reported stress level (*Low, Mid, High*).

Our first objective is to analyse how subjects are related to each other in terms of how similar are their models. From the set of 30 subjects, we removed those that had a significant number of missing values (mainly in the questionnaires for self-evaluation of their stress level). Thus, having a remaining set of 18 subjects.

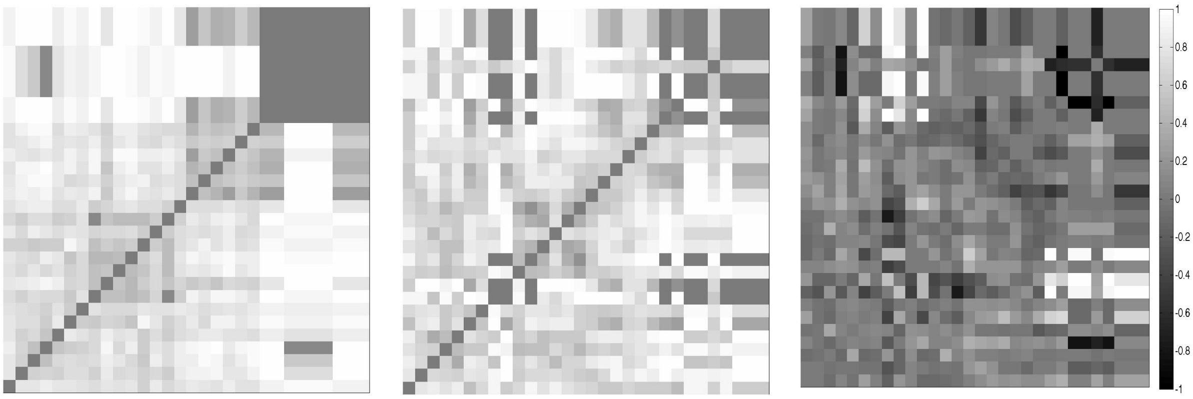
A decision tree was learned for each subject and using the distance in Equation (2.3) we compared all pairs of models to obtain a similarity matrix. From that matrix we performed hierarchical clustering using the unweighted pair group method with arithmetic mean (UPGMA) algorithm which yields the dendrogram depicted in Figure 6.2, where a coloured box indicates the average self-reported stress for that subject. From the figure, we can observe 3 clusters with 7, 6 and 4 subjects. The largest cluster (with 7 subjects) roughly corresponds to subjects which reported low levels of stress in average (denoted by the blue boxes). The second major cluster (with 6 subjects) corresponds to subjects who reported a mid level of stress (gray boxes). A third cluster with only 4 subjects shows subjects with high and mid level of stress.

6.2.2 Missing data and Semi-supervised learning

Since the initial data had a large portion of missing values (≈20% of overall dataset), semi-supervised learning was used to fill those. In this research, we use self-training (ST) Zhu, 2006 with C4.5 as classifier. We have trained a model for each subject and we have also established a single model combining all the attributes from all the subjects. We performed 10-fold cross validation in all the experiments using Weka Hall et al., 2009 with the default parameters of C4.5 classifier. The new classified data with high confidence (≥80%) is added to the training set, the classifier is re-trained and the procedure repeated. Using ST we were able to reduce the unlabeled data (improving the labeled dataset in ≈14%). This resulted in improving the average accuracy (4.20%), precision (3.5%), recall (4.1%) and F-score (4.0%) as shown in the Table 6.3.

After applying the semi-supervised learning phase, there is enough data to compute comparisons with the 30 subjects in the study. The process described in the previous section was repeated to obtain a similarity matrix, depicted in Fig.6.3 (a), where the more similar a subject is to another the darker that square is (subjects are ordered by clusters). To evaluate our proposed transfer learning approach, we generated another dataset which has a reduced amount of instances. We randomly removed 50% of the data from all subjects. The similarity matrix of this reduced dataset is depicted Figure 6.3 (b). Finally, in Figure 6.3 (c) we depict the matrix resulting from the difference of (a) and (b), where a grey box means no difference.

In summary, we have three similarity matrices: i) initial dataset (18 subjects) ii) after applying semi-supervised technique dataset (30 subjects) and iii) after removing 50% of data (30 subjects). All of them have different missing data. For each matrix we computed its average value, with the following results. The initial data showed a more disperse set of distances with an average of 0.65 ± 0.18 (higher value, means subjects are more different to each other). After the semi-supervised algorithm was applied the average distance was 0.55 ± 0.16 even when the number of subjects increased (30 subjects). Finally, when the data was reduced the average distance decreased to 0.49 ± 0.15, which may not happen in all cases.



(a) (b) (c)

Figure 6.3: Similarity matrices of 30 users using (a) all data (after semi-supervised learning) and (b) with 50% of instances removed –darker cells indicate high similarity. (c) depicts the difference between (a) and (b); a white cell indicates a + difference, black a − negative difference, and grey no difference.

$$\Delta i, j(original, modified) = |e_{i,j}^{original} - e_{i,j}^{modified}| \tag{6.1}$$

Since we are interested in knowing how the similarity among models is affected by adding or removing data, we evaluated the percentage of entries (models) where $\Delta_{i,j} > \epsilon$ with $\epsilon = 0.1, \ldots, 0.9$ between two matrices. After applying the semi-supervised approach, only 1% of entries changed more than 0.8 (1.0 is the maximum possible change). After applying the semi-supervised approach the similarity matrices were only slightly altered with an average value of 0.12 ± 0.14, meaning there were no drastic changes in similarities. In contrast, when we reduced the data by 50% and compare the similarity matrices their difference in average was 0.19 ± 0.20, which is expected since the data was significantly reduced. Moreover, only 5% of the entries were altered more than 0.9 (i.e., the similarity matrix changed completely).

These results show that 1) the semi-supervised approach does not alter drastically the learned models and 2) the used similarity measure is robust even when data is added or remove from the model. This is an important result which will be useful in the next section since we start with the reduced data and show that using transfer learning can improve the accuracy of the learned models.

6.2.3 Transfer Learning

The previous section showed how to use semi-supervised learning to cope with missing data by using the information obtained from one subject. A different way to solve this problem is to use information from another known models (another subjects in the study). In this way, we need to *transfer* information from other models to our target model which contains insufficient data to produce an accurate one.

In order to perform transfer learning we need information of other subjects, in particular our approach assumes a set of previously learned models (decision trees) along with their respective data (used to learn the decision trees). When, a new subject appears, it is expected to be associated with scarce data, which can result in having a model with poor predictive accuracy. TL uses information from other subjects to improve the model.

First we learn a model t_i for the new subject i using only the available data. This model is compared with the rest of the T models of the other users using Equation 2.3. In order to select which data should be transferred four different approaches were evaluated. The first two are simple approaches transferring all data from the most similar subject. The third one is based on sampling data weighted by its distance and the last one is based on ensembles that weight their prediction based on its distance to the target model. In detail,

1. Naive approach. Select the most similar model,k, to t_i:

   $$k = argmin_{t_j \in T} d(t_i, t_j)$$

   and transfer all its data to i. A new model is learned using the original and the transferred data.

2. Threshold approach. If most similar subject to t_i is closer than a threshold β then transfer its data.

   $$k = argmin_{t_j \in T} d(t_i, t_j) \text{ and } d(t_i, t_j) < \beta$$

   A new model is learned using the original and the transferred data.

3. Sampling weighted approach. Select the K most similar (source) models closer to t_i:

   $$K = \bigcup_{m|\text{most similar to } t_i}$$

   Then, for each source model perform sampling weighted by its distance to t_i. Sampled data is transferred and used with the existing data, to learn a new model.

4. Ensemble weighted approach. Use the K most similar (source) models closer to t_i and the model learned with scarce data to classify the target data. The voting scheme (to select the actual prediction from the ensemble) is weighted by the distance from each model to the target one.

We applied the four proposed transfer learning approaches on the data which has a percentage of data removed and we use as upper bound the results obtained with the complete data.

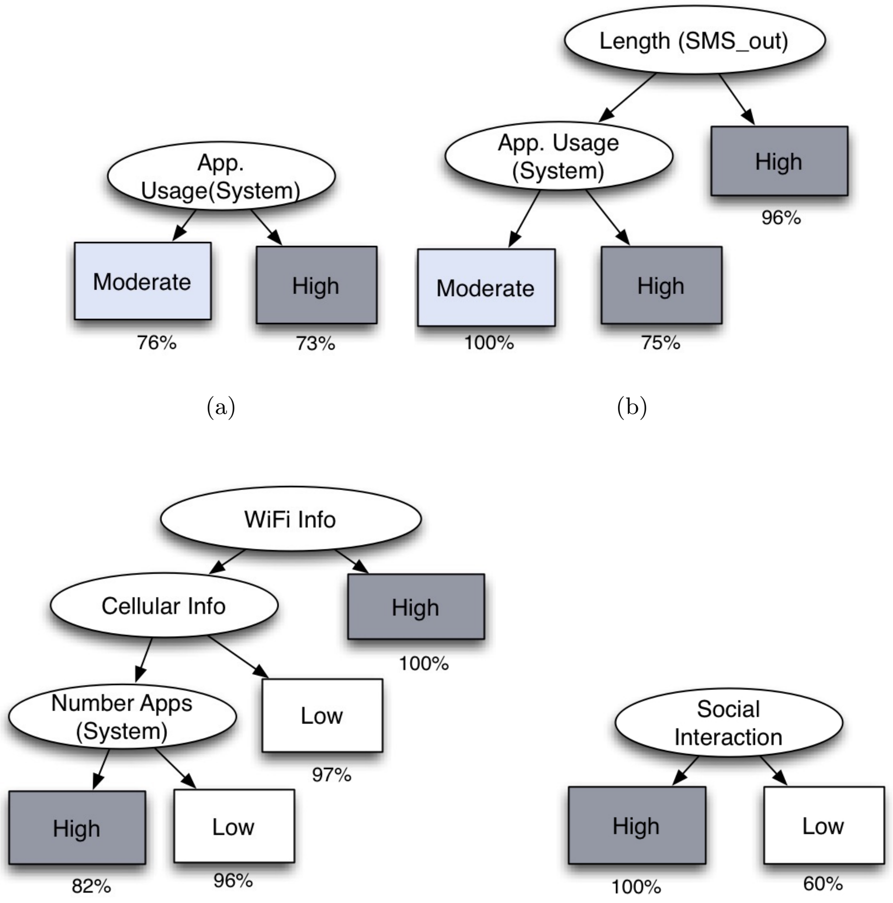
One of the important aspects in transfer learning is deciding which data to transfer. In our case we are interested in how similar source models are to our current target model (with scarce data). We computed the distance to the nearest model, the farthest model and average for every subject in the study. From the results we obtained an average distance of 0.42 (using Equation 2.3) to the nearest subject, in contrast, the average to all models was 0.74 ± 0.17. We also noted that there are cases where a subject has several nearest models with the same distance. There are 18 subjects that have a unique nearest subject. These subjects were selected for the proposed transfer learning approach (see Table 6.4).

Table 6.4: Classification accuracy using the naive transfer learning approach, Δ transfer shows the difference between no transfer and transfer columns, $d(near)$ shows the distance to the nearest model. All data shows the accuracy using all original data (upper bound). Using the naive approach does not yield the best accuracy in average.

S.ID	**No Trans.**	**Naive Trans.**	$d(near)$	**Δ Trans.**	**All data**
S09	57.69	**73.08**	0.36	15.38	76.92
S30	42.86	**53.57**	0.36	10.71	78.57
S11	65.45	**74.55**	0.62	9.09	72.72
S10	44.89	**51.02**	0.27	6.13	71.42
S28	57.35	**63.24**	0.18	5.88	77.94
S16	61.11	**62.96**	0.48	1.85	74.07
S24	**67.14**	**67.14**	0.36	0.00	71.42
S12	**55.93**	54.24	0.32	-1.69	62.71
S25	**85.71**	83.67	0.39	-2.04	89.79
S14	**51.56**	48.44	0.49	-3.13	82.81
S23	**53.33**	50.00	0.53	-3.33	58.33
S05	**70.69**	65.52	0.36	-5.17	86.20
S19	**60.00**	53.33	0.54	-6.67	90.00
S08	**57.41**	50.00	0.46	-7.41	55.55
S18	**70.27**	62.16	0.32	-8.11	75.67
S04	**81.25**	71.88	0.42	-9.38	84.37
S01	**72.86**	61.43	0.58	-11.43	78.57
S29	**62.07**	44.83	0.60	-17.24	79.31
Avg.±Std. Dev.	**62.09 ± 11.32**	60.61 ± 10.71	0.42 ± 0.12	-1.47 ± 8.42	75.91 ± 9.70

First, we evaluated the naive transfer learning approach. Accuracy for the transfer learning approach is obtained by learning a classifier using the reduced data and the transferred data, then testing that model on the data without removed instances. As an upper value of the possible accuracy we learned a model with the complete data and the evaluation was performed on that same dataset. Table 6.4 summarises the results using the naive approach showing the accuracy results with and without our proposed transfer learning approach and the accuracy using the complete data.

Using the naive approach did not improve the accuracy for all subjects. This happens because we are ignoring when transfer can be more useful: the distance to the nearest subject. The idea is to use transfer only when the distance is small (i.e., when the model is close to another) defined by a threshold β. To exemplify this behaviour see Figure 6.4 (a) and (b) where we depict trees which have a $d = 0.36$. In this case trees are similar in their decision nodes. In contrast, Figures 6.4 (c) and (d) show trees which have a $d = 0.60$. Note, that in this case the trees show different decision nodes.



(a) (b)

(c) (d)

Figure 6.4: Learned models of different subjects: $S30$ (a) and its most similar $S17$ (b). $S29$ (c) and its most similar model $S05$ (d).

Our second approach, threshold based, takes into account this distance with respect to the closest model. We performed experiments varying the threshold, β, with values between $[0, 1]$. From the results we observed that trivial approaches: not using transfer or using transfer on all subjects do not obtain the best results (62.09 and 60.61 accuracy for $\beta = 0$ and $\beta = 1$, respectively). However, selecting the appropriate threshold of transfer increases the accuracy (63.37 with a threshold of 0.37). Table 6.5 summarises the results of using the threshold transfer approach ($\beta = 0.37$). In particular, it shows that accuracy improves from 58.35 to 61.24 when models that are closer than the threshold are used. On the other hand, when $d \geq \beta$ it is better not to use transfer learning since the models are far from each other and this causes a negative transfer effect.

Table 6.5: Classification accuracy, Δ transfer shows the difference between no transfer and transfer columns. All data shows the accuracy using all original data (upper bound). The number of initial and transferred instances is shown. The top part of the table shows the results when the distance to the closest subject is small (< 0.37), while the bottom when it is large (> 0.37).

Subject ID	No Transfer	Threshold Trans.	Δ Transfer	All data	Total Inst.	Trans. Inst.	$d(near)$
S28	57.35	**63.24**	5.88	77.94	61	26	0.18
S10	44.89	**51.02**	6.13	71.42	57	31	0.27
S12	**55.93**	54.24	-1.69	62.71	49	31	0.32
S18	**70.27**	62.16	-8.11	75.67	49	18	0.32
S24	**67.14**	**67.14**	0.00	71.42	67	31	0.36
S05	**70.69**	65.52	-5.17	86.20	66	37	0.36
S30	42.86	**53.57**	10.71	78.57	66	29	0.36
S09	57.69	**73.08**	15.38	76.92	53	35	0.36
Avg.± Std.dev.	58.35 ± 10.0	**61.25 ± 7.1**	2.89 ± 7.5	75.11 ± 6.4	58.5 ± 7.1	29.75 ± 5.4	0.31 ± 0.06
S25	**85.71**	83.67	-2.04	89.79	55	31	0.39
S04	**81.25**	71.88	-9.38	84.37	63	31	0.42
S08	**57.41**	50.00	-7.41	55.55	62	35	0.46
S16	61.11	**62.96**	1.85	74.07	59	29	0.48
S14	**51.56**	48.44	-3.13	82.81	63	31	0.49
S23	**53.33**	50.00	-3.33	58.33	67	35	0.53
S19	**60.00**	53.33	-6.67	90.00	59	26	0.54
S01	**72.86**	61.43	-11.43	78.57	73	32	0.58
S29	**62.07**	44.83	-17.24	79.31	62	33	0.60
S11	65.45	**74.55**	9.09	72.72	59	29	0.62
Avg.± Std.dev.	**65.08 ± 11.4**	60.11 ± 13.0	-4.9 ± 7.3	76.55 ± 11.8	62.2 ± 4.9	31.2 ± 2.7	0.51 ± 0.08

Our third transfer learning approach is based on sampling from similar models. Thus, our approach is to select the k closest models to our subject and sample its associated data to obtain data to be transferred. We tried different values for the number of similar models and we used a weighted approach to determine how many instances should be sampled. This is based on the distance to the target model, bounded to half of number of total instances in the source trees. For example, if the distance between trees is 0.0 (i.e., totally similar) and there are 100 instances in the source, 50 instances will be sampled from that source and transferred.

We performed different experiments varying the number of similar subjects to be sampled from 1 to 7, results showed that, transferring information from only one subject (the most similar one) obtained the best scores in average 63.3 ± 10.92 (avg. accuracy ± std. dev.). In contrast, increasing the number of close trees decreased the accuracy to 55.26 ± 13.3 (using the 7 closest similar subjects).

6.2.4 Ensemble method

Finally, our last approach is based on ensembles and we tried two different approaches to improve accuracy. First we need to select two parameters, the number of trees used in the ensemble (counting also the target tree) and the way to combine their results. For selecting the number of trees in the ensemble we tried ensembles with size $\{3, 4, \ldots, 15\}$. To decide how to join the results of those trees we tried two approaches. The *simple voting* approach sums the results from different trees uniformly. This approach was tested with different number of close trees. However, results did not increase, in fact the average accuracy obtained was 49.99 ± 29.15.

Thus, we tried a second approach that weights their predictions based on the distance to the target tree (recall that distance between trees is in range of $[0, 1]$). We evaluated different number of trees in the ensemble from 3 to 15. However, the best scores were obtained using 4 trees in the ensemble (3 most similar source trees and the target tree) obtaining 72.7 ± 20.2. Increasing the number of trees consistently decreased the accuracy (63.3 ± 22.9 with 15 trees).

6.2.5 Summary of analysis

We proposed four different transfer learning approaches to cope with scarce data. Table 6.6 summarises the results of the proposed approaches compared without transfer and with all the original data (used as upper bound). Results show that threshold, sample weighted and ensemble weighted approaches obtained better scores than without a transfer approach. The threshold and sampling approaches obtained similar scores and the ensemble approach obtained the best scores increasing the accuracy almost by 10% in average.

As conclusions from the experiments we note that:

- Transfer from few, but similar, subjects was better than using more subjects which are not close to the target model.

- Transfer using another models (ensemble approach) was better than transferring instances.

Table 6.6: Classification accuracies using the proposed approaches and using all original data (upper bound).

		Transfer learning approaches				
Subject ID	**No transfer**	**Naive**	**Threshold**	**Sampling weighted**	**Ensemble weighted**	**All data**
S01	72.86	61.43	72.86	64.28	**87.14**	78.57
S04	**81.25**	71.88	81.25	65.62	73.44	84.37
S05	**70.69**	65.52	65.52	75.86	68.97	86.20
S08	57.41	50.00	57.41	57.40	**85.19**	55.55
S09	57.69	**73.08**	**73.08**	65.38	38.46	76.92
S10	44.89	51.02	51.02	55.10	**63.27**	71.42
S11	65.45	74.55	65.45	**76.36**	65.45	72.72
S12	55.93	54.24	54.24	55.93	**62.71**	62.71
S14	51.56	48.44	51.56	53.12	**90.00**	82.81
S16	61.11	62.96	61.11	62.96	**90.74**	74.07
S18	70.27	62.16	62.16	70.27	**81.08**	75.67
S19	60.00	53.33	60.00	70.00	**90.00**	90.00
S23	53.33	50.00	**53.33**	38.33	38.33	58.33
S24	67.14	67.14	67.14	**70.00**	**70.00**	71.42
S25	**85.71**	83.67	**85.71**	83.67	**85.71**	89.79
S28	57.35	63.24	63.24	60.29	**95.59**	77.94
S29	62.07	44.83	62.07	**67.24**	36.21	79.31
S30	42.86	53.57	53.57	48.21	**66.07**	78.57
Avg.±Std.dev.	62.09±11.0	60.61±10.4	63.37±9.5	63.33±10.6	**71.58±18.2**	75.91±9.4

6.3 Using motor activity-related behavioural features toward unobtrusive stress recognition

In the previous sections 6.1 and 6.2 we demonstrate the importance of analysing behaviour patterns as an objective signal that may have impact on cognitive function. This section introduces motor activity-related behavioural features that can be extracted from a smartphones during phone conversation, with the view towards unobtrusive stress detection. We used quantitative analytic methodology of motor behaviour pattern classification for work-context, individual employees in our longitudinally collected data. The Fourier analysis of the motor activity intensity was measured during phone conversation and showed that the relation between the high frequency range was lower in high perceived level compared to subjects with lower perceived level.

We evaluate the performance of novel method, namely Intermediate Models that has been used in previous research (in Section 5.5) to infer motor activity in bipolar disorder patients. The key advantage of the proposed *intermediate models* approaches is to improve the performance of supervised classifier. We build six intermediate models using the self-reported mood states to build the final model in predicting stress.

6.3.1 Stress modeling using Intermediate Models

In the previous Section 6.1 we reported current approaches for inferring stress that rely mostly on self-reported questionnaires, such as the work in (Näätänen and Kiuru, 2003). This results in problem for an effective measurement, since subjects are often affected by a personal confidence. For example employees might have more predisposition to report information in their favour or for the organisation than reporting their true health-conditions. To overcome these situations smartphones are becoming useful to perform research due to their availability, rich set of embedded sensors and their capacity to be unobtrusive for the subjects. However, still remains an open problem how stress can be effectively detected with the help of systems that retain an increased degree of unobtrusiveness.

Motor activity-related behaviour (*i.e.*, body hyperactivity, trembling, uncontrollable movement, hand movement (Morgan III et al., 2015; Smith and Seidel, 1982)) has shown association with perceived stress. Currently, the clinicians assess measurement of motor activity in laboratory settings. Studies measuring level of motor activity in psychological stress have typically used traditional monitoring with paper and pencil diaries, and questionnaires (Prasad et al., 2004). Monitoring motor activity during sleep may be measured by actigraphs (Mezick et al., 2009) (using piezoelectric accelerometer). However, little is known if data captured from an actigraph could provide motor activity characteristics in perceived stress level in working environments.

Smartphones are a good candidate for monitoring motor activity behaviour patterns in daily activities. Information from smart phones enables easier monitoring and tracking of people than traditional methods, as most people already carry a smartphone so no additional sensors are required. Another benefit of using this technology is that other information (such as phone calls, location, use of social networks) can be obtained and included. In this research, we collect data from accelerometers during phone calls to infer motor activity changes in working employees.

To our knowledge, no research has explored until now the potential of motor activity related behaviour features in working environments with the aim to detect stress. This is important since, motor activity features could be acquired through the use of smartphone's accelerometer during phone-conversation, in a totally unobtrusive manner.

In the context of our research work, the following research questions are put forth:

- ◆ Is there a relationship between motor activity features that can be automatically extracted from a accelerometer sensor embedded on smart phones and the self-reported stress levels?
- ◆ Is it possible to improve stress detection by incorporating intermediate *hidden* variables related to the subjects' mood, before building the final model for predicting

stress?

The present work tries to answer both these research questions by comparing standard stress measurement questionnaires and motor activity behaviour during phone conversations.

We performed an experimental analysis using real world data. The research includes 2 aspects:

- Using semi-supervised learning to complete the models for subjects with missing data.
- Using Intermediate Models to predict mood variables, which are incorporated in the final model with the aim at improving the accuracy of the predictions.

6.3.2 Intermediate models

Self-reported questionnaires acquired from participants are useful in understanding perceived mood and stress. However, it is a tedious task for the user. In this research we propose to predict the mood variables associated to the questionnaires using the data from the smartphone to alleviate the user from this burden. We then use the predicted mood variables with the rest of the data from the smartphones to predict the stress levels. We call the models that predict the mood variables from the questionnaire *Intermediate Models* as there are used as input for the final predictive model. Although the use of additional variables, such as latent variables, have been previously used in the literature, we are not aware of research that aims at building an intermediate model that can then be used as input for the final model.

We used six variables derived from NA and PA (3 per each mood affect) to build 6 intermediate models. Furthermore, we train each classifier separately using each the self-reported questionnaires derived from the 'Positive Mood Affect (PA)' and the 'Negative Mood Affect (NA)'. In the prediction stage, the intermediate models use the information from the smartphones to predict a weighted set of mood variables based on the accuracy of each model. Then all the data from the smartphones and the mood variables are used as input for the final stress model.

6.3.3 Semi-supervised learning

Similar as in previous work in bipolar disorder 5.5, also in this research we consider one of the most common methods of SSL that uses a single classifier called Self-Training (Zhu, 2006).

Our experiments have the following objectives:

- Compare the performance of different classifiers on the data.

- Assess the effect of Intermediate Models to enhance the knowledge of perceived stress in employees.
- Use SSL to address the problem on how to use information from unlabeled data to enhance classification accuracy.

For all the experiments, we used Weka's (Hall et al., 2009) classifiers with their default parameters. We build a model for each subject and performed a 10-fold cross validation for all the experiments; we report the average accuracy, precision, recall and f-score values for all participants. In Figure 6.5 show the results using different classifiers. In the first experiment we compare the performance of the classifiers based only on the labeled data (Supervised) with the inclusion of unlabeled data using semi-supervised learning (SSL). In the second experiment we analyse the impact of using the intermediate models, without and with SSL.

6.3.4 Comparison of results using proposed approaches

As described in Figure 6.5 [3] there are more than 2033 (27.6%) of phone conversation without an associated stress level. To address this issue, we used SSL (Self-Training approach) to see if we can enhance on the performance of previous result using all the available data. As can be seen from the results presented in the Figure 6.5, adding information from other phone conversation is improving the accuracy results for circa 4% and around 10% improvements in terms of Precision, Recall and F-Measures.

Using subjects self-reported mood, we propose building an intermediate model approach aiming at improving the classification accuracy. For this research we train the classifier separately using each items from 'Positive Mood Affect -PA' and 'Negative Mood Affect -NA' from the questionnaire.

In our dataset, more than 2033 (27.6%) of the phone conversation did not have an associated stress level (the user did not answer the questionnaire). To address this issue, we used the SSL Self-Training Method described above. We followed a simple approach where we divided the data into ten folds, where the training data was used to classify the unlabeled data, as threshold for the confidence we used $\geq 80\%$ for the highest classified value. Then we used all the classified data with the original training set to produce an extended training set. As can be seen from the results adding information from other phone conversation is improving the accuracy results in terms of Accuracy Precision, Recall and F-Measures for all the classifiers, in some cases as for C4.5 the improvement is significant (nearly 10%).

By incorporating the intermediate models, a further improvement is obtained in both

[3] More details from accuracy comparison between methods are shown in Table A.10 and Table A.11

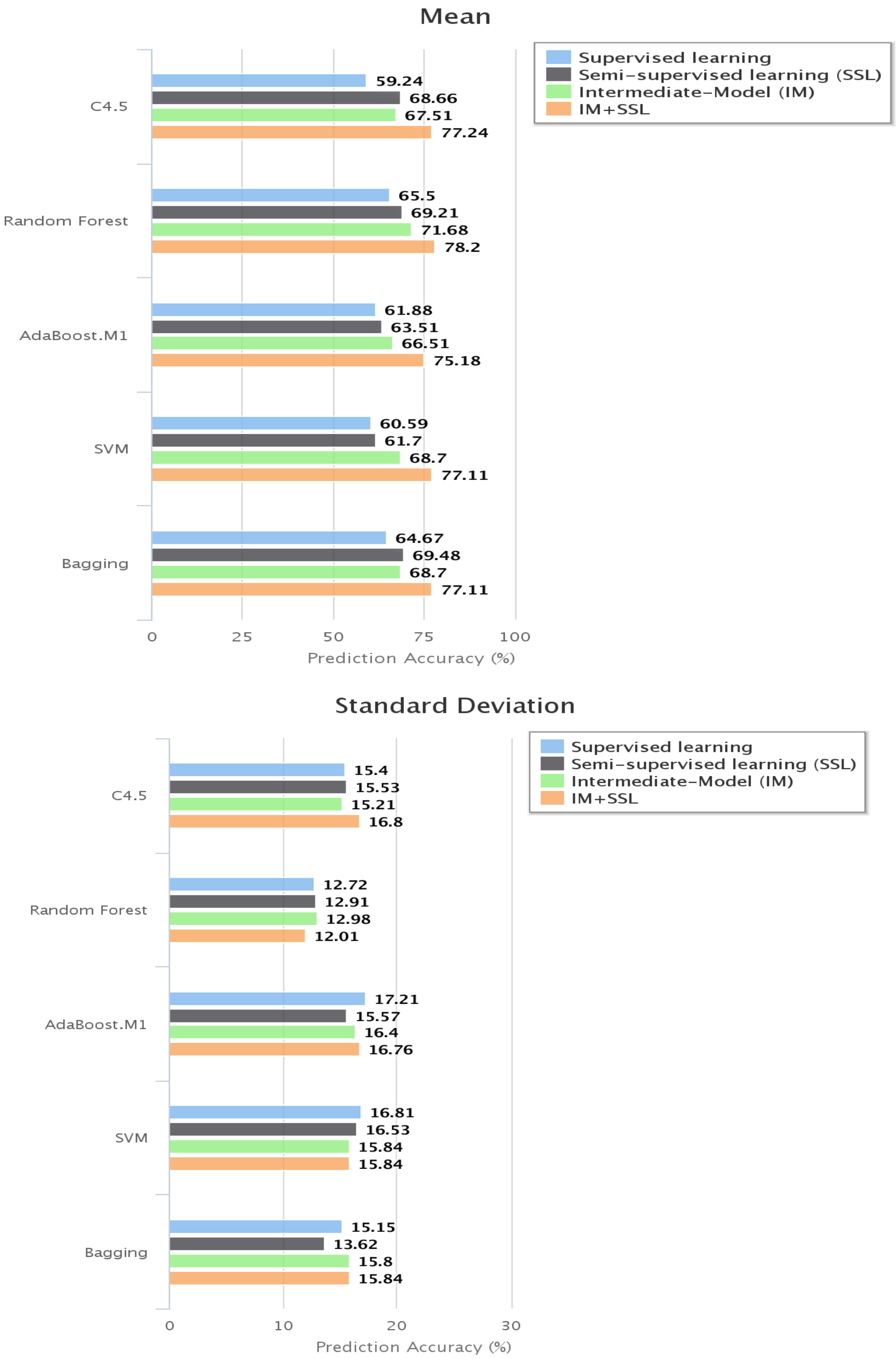


Figure 6.5: Comparison in terms of accuracy using supervised learning, semi-supervised learning (SSL), intermediate models (IM) and semi-supervised & intermediate models (SSL+IM) with different classifiers for predicting perceived stress.

case, without and with SSL. As it can be observed in Table A.11, the best results are obtained by combining SSL and the intermediate models, and in particular with the random forest classifiers.

6.4 Discussion

Using smartphones for monitoring behaviour patterns of individuals in their working environments has the potential to provide valuable insights of their health. This research aims to do that by combining data from different sources, such as objective data measurements and subjective self-reported data. The challenges that we faced in the study arise in the integration of multiple objective and subjective data streams, the definition of the questionnaires and the large number of missing values since data was collected in a real-life environment from heterogeneous sources.

A common issue when dealing with health applications is the challenge of recruiting sufficient number of participants (Xiang et al., 2013). We have faced the same challenge in our study and furthermore we have faced issues with subject compliance leading to a decrease in the amount of self-reported data, but also sensor data (for example, forgetting to charge the battery). With respect to the limitations, it is important to note that we assume that subjects in our study have an inherent degree of similarity in their behaviour for the transfer learning method to perform well.

When we consider a higher number of subjects, we also plan to use demographics and self-reported information related to personality to measure inter-subject similarity and hence we expect a better performance of the transfer learning method. Another limitation is the dissimilarity measure used to compare models. For example, it does not take into account the splitting values inside the attributes and it is affected by the tree size (height) (Miglio and Soffritti, 2004). Therefore, other approaches might be explored Chipman et al., 2001; Fowlkes and Mallows, 1983; Miglio, 1996; Shannon and Banks, 1999.

Finally, one last limitation is that the participants were recruited through two different organisations (i.e., logistic, software development) in the private sector. Thus, there will be some limitation in transfer learning to other organisations or sectors. However, the employees that participated in our study had heterogeneous characteristics with regard to gender, age, marital status, and educational level, which will be an advantage in transfer learning.

6.5 Chapter Summary

In this research work, we have presented an extensive analysis based on real data from 30 users in two organisations related to stress using information derived from smartphones. We contrasted objective variables, acquired from smartphones, such as physical activity, location, social interaction and social-activity with respect to perceived stress levels, considering several demographics (gender, age, education and marital status). Correlation analysis was used to analyse the possibility of using smartphones derived data aiming at predicting perceived stress levels at working environments. We addressed the problem of missing information and scarce data to improve the prediction accuracy using self-training as standard supervise-learning approach and transfer learning approach to find the similarity of perceived stress. We presented improved results using our novel intermediate models on top of the proposed approaches, resulting in improved performance in accuracy. Finally, we propose analysing specific human behaviour during the phone conversation. Motor activity features have been extracted to classify the behaviour changes of the subjects, which behaviour could be a result of daily perceived stress.

6.5.1 Correlation findings in stress

A summary of the most important findings in the Section 6.1.1 have been presented below:

- There is correlation between objective data such as: location information (WiFi and Google Location data), social interaction, and information from phone calls and SMS with subjective data that represents mood of the user (i.e., level of stress).
- Overall physical activity during lower perceived stress times throughout the entire monitoring period was associated with higher activity. In contrast, a high perceived stress showed lower physical activity.
- With respect to gender, men showed a more stable social interaction across the weekdays. In contrast, women then to increase their interaction near the weekend.
- Our results suggests that the more social the subject is the more stressed he gets, this can be explained because the subject is probably talking with colleagues about work which increases its stress. On the other side there is negative correlation between duration of calls and stress, the reason could be that the subject is stressed so she has no time to spend on calls.
- Based on smartphone data it is possible to predict stress using decision trees. However, missing data is an aspect to take into account. In this work using semi-supervised learning techniques we increased the accuracy from 67.57% to 71.73% for predicting stress.

And, some of the conclusions of this work are summarized:

- There is clearly a high to moderate perceived stress in most of the employees. This confirms some of the findings on other reported studies about stress. The possible consequences of stress motivated our work for finding unobtrusive ways to detect it, via smartphones, and analyse in more deep the most relevant aspects related with changes in the behaviour of employees under different stress conditions. We believe that this is an important step towards a better understanding of behaviour of employees under stress and to design remedy actions.
- It appears that women tend to present higher percentage levels of perceived stress. This does not necessarily mean that they are more stressed, but at least that they perceive it more. Whether this has to do this with higher sensitivity levels in women than men, a biased finding due to our small sample size or to a more profound reason related to gender, this requires further and deeper studies.
- Perceived stress varies among companies and this could be related to their working conditions. Identifying working conditions on companies with low levels of stress could help to establish better working policies to reduce stress among employees.
- There appears to be different behaviours in some job-related aspects in relation to stress between men and women. Although again this needs deeper and thorough study, if it is the case it could help to improve some working conditions based on gender.
- The use of smartphones has become part of the daily activities of people and our experiments showed that there are clear changes in their use (phone calls, SMSs, apps) under different stress conditions.
- There is a clear correlation between how people behave at work (physical activity, WiFi location, number and duration of outgoing calls and SMS, and with social apps) and stress levels. This could be easily monitored with current smartphones, as shown in this research, to detect possible stress levels and help to implement corrective measures.

6.5.2 Findings using Transfer-learning in stress prediction

In the Section 6.2, we have demonstrated the importance of obtaining sufficient data in order to predict effectively behaviour changes of the user relevant to stress. We proposed building a reliable user-specific model from a considerable amount of data. This data is divided into two parts: the objective data which is obtained automatically from the device and subjective data which is generated by the person.

Data collected in this study have around 21% of missing labels, thus, in this research work proposes techniques to address the problem of having limited data. One of those approaches is semi-supervised learning which uses the learned model to complete missing values and reduce the amount of unlabeled data. Another related approach is called transfer learning which uses information from another sources to improve the quality of a new model. Further, we have proposed four different methods based on transfer learning to deal with the scarcity of data. The proposed approaches are based on obtaining a distance among models and using similar (close) models to improve the predictive accuracy. In this work we transfer instances (sampling based approach) from another close model or using close models from other subjects (ensemble approach). As a result, we have shown that the weighted ensemble approach increases the accuracy by almost 10% compared with the no-transfer approach through the experimental evaluation with real-word data obtained from employees of two different companies.

A future exploration avenue is to use of multi-label classifiers, where a set of classes (in this case all the variables associated with the questionnaires) can be predicted at the same time and where dependencies between these classes can be incorporated to improve the classification performance.

6.5.3 Motor activity findings in prediction of Stress@Work

Finally, in the Section 6.3 we presented a research work of how to predict perceived stress of employees by analysing motor activity behavioural data during phone conversations. We extracted several frequency domain features to analyse the motor activity-related behaviour from different users. The results demonstrated that subjects have distinctly different profiles of motor activity and that the results differ according to perceived stress analysed. We assume that this methodology may have great potential for behaviour analysis and more acceptable for the monitored subjects due to level of obtrusiveness.

Similarly, as in previous sections, we dealt with large number of unlabeled instances. To address these issues, we proposed using semi-supervised learning techniques, which have shown to improve the prediction level and increasing number of labeled instances. Additionally, we also applied a novel approach to incorporate unobserved variables via intermediate models. We evaluated experimentally the impact of using SSL, intermediate models and both combined, using different base classifiers. The proposed approach for creating intermediate models has been shown to increase the prediction of the stress level of the users using the data derived from motor activity; from 61.5% using the standard supervised methods to $\geq$78% after applying intermediate models and SSL.

As a future line of this work would be applying transfer learning and multi-label supervised-learning approaches and identify similar pattern of users in different stages of perceived level.

Chapter 7

CONCLUSIONS

This chapter summarizes the main achievements of this research work, discusses the outcomes of this dissertation, acknowledges the limitations and future research ideas. We review the literature in Chapter 3 seeking the current research challenges addressing the problem of acquiring a large amount of labeled training data in real-world monitoring scenarios, requiring a human effort and time to label data. The learning from literature drew the path way which we took to build a machine learning solutions that enables addressing scarce data and unlabeled information. We validate our fundamental question of how to extract knowledge out of unlabeled data in order to infer a human behaviour and improve classification performance compared to conventional machine learning methods (i.e., dropping cases entirely when they have missing ground truth).

For this thesis, our proposed approaches have considered how to address the issues of unlabeled and scarce data in the mental-health and human behaviour fields. We propose solutions to the challenges in both areas by introducing our novel Intermediate Models, following the use of Semi-supervised learning and Transfer Learning approaches that can learn effectively within this regime. Our work has considered how these approaches relates to a challenge of human behavioural classification from smartphone collected data. We have focused in this direction as we believe that the challenges to perform scalable classification is currently one of the most critical bottleneck of the monitoring devices using sensing modalities.

7.1 Contributions

This PhD thesis begins to solve several open problems in machine learning and have been applied in two healthcare domains for monitoring human wellbeing. As we discussed earlier, collecting training labeled data are expensive, as a human annotation must take the effort to label data, thus, it is frequently the case that labeled training data are sparse.

We have also emphasized in earlier chapters that unlabeled data are often plentiful and is all around us in the different forms, for example, phone recordings, web queries, metadata, sensory, locations and others logs.

In this thesis, we propose a solution to several of the large challenges in the area machine learning by introducing our novel *Intermediate Model* for improving the accuracy performance of final model, and the setting of *Semi-supervised learning*, *Transfer Learning* that can learn effectively within this regime. One of the key question in this research work is how to extract the knowledge and efficient value out of these unlabeled resources in a wide range of learning environments. By leveraging unlabeled data, we have demonstrated that we go beyond the limited models that can be learned from small portion of training sets. This research work suggest that it is highly advantageous to have SSL and TL integrated in monitoring systems that both benefits can take advantage when new unlabeled data becomes available.

All three methods make very different assumption about the underlying data, however. In the Chapter 5, we have demonstrated our results using *Self-training* method in the data collected from the bipolar patients. Using Self-training enabled us new perspective to tackle missing labeled instances between psychiatric evaluation and collected sensory data. We have demonstrated the evidence that in future monitoring in-remote mental-disorders is no longer dependent on continues human observer or even continuous self-reports from the patients. The results in Chapter 5 have provided an evidence that with few labeled instances available during the learning helped us to guide the learning models and evaluating the performance ST algorithm. In the Chapter 5, we presented performance accuracy difference between supervised and semi-supervised learning methods. The supervised methods performed slightly better ($\approx$0.75%) to semi-supervised learning. However, there were more than 900 phone-calls that where without associated episode which were included into the building of final model using semi-supervised learning, respectively Self-training methods. On the other hand, prediction of perceived stress using Self-learning approach, in Chapter 6 we demonstrate the improvements of overall accuracy from 67.57% to 71.73%. We were able to reduce the number of missing classes from $\approx$20% to $\approx$6% and improve the knowledge of days without associated stress level.

The TL approaches also differ in their basic mode of learning relationships between the participants in order to transfer knowledge deduced from the source labeled data to the target unlabeled data. In Chapter 6 we have demonstrated the use of Transfer learning posing as well new challenges in machine learning, such as mapping between the different feature vector spaces. Using both approaches SSL and TL (as shown in Chapter 6), both methods are shown to resolve space complexity through unlabeled data without reducing learner accuracy. In the Chapter 6, combining both approaches, we have validated the proposed machine learning methods to augment a small amount of labeled data with large amount of unlabeled data to improve classification performance.

Further, we have demonstrated *Intermediate Model* approach with a novel assumption of improving the scope of standard supervised learning, semi-supervised learning, and transfer learning by incorporating new information and allowing unlabeled data to be of value in the learning process for building the final model. In the Chapter 5, we have presented the results using standard supervised learning and semi-supervised learning. The results obtained from additional information added from *Intermediate Models* has been shown to improve the overall performance accuracy (from ≈73% to ≈90%). Similar, in the Chapter 6 the proposed approach for creating intermediate models has been shown to increase the prediction of the stress; from 61.5% using the standard supervised methods to 71.68% after applying intermediate models and ≈78% after being combined with SSL.

To sum-up, with our studies we have evaluated the impact of these techniques in two real studies to classify the state-mood of bipolar disorder patients and the perceived stress of employees at work using the acquired data from smartphones. We have used in both domains real data from subjects for several monitoring weeks on unconstrained conditions. And in both cases the incorporation of additional information, automatically extracted from original dataset, into the learning process, has been shown to increase the performance of the induced models. For our scarce data problem, we can conclude that using the proposed *Intermediate Models* to enrich learning and performance of models as the best approach for our research work, because it has been shown to provide an attractive balance of both accuracy and conceptual simplicity. Thus, we encourage researcher to conduct similar methodological assessments to find the most suitable method of increasing unlabeled instance for their specific datasets and measures.

Although the existing noise in the features extracted, the results achieved from IM with SSL and TL methods have greatly improved performance over supervised learning. The work in the Chapters 5 and Chapter 6 shows that using proposed approaches leads to effective performance with small amount of labeled data. Combining these methods helps to resolve fundamental Ubiquitous Computing problem on the way towards self-sufficient autonomous systems that supervise their own learning. The findings of this research work provides guidelines to researchers and machine learning developer who design a monitoring systems for different domains.

The main contributions of this dissertation to the field of Ubiquitous Computing are summarized below:

- We presented the first work to manage scarce data to monitor mental-health and human behaviour using collected longitudinal smartphone data.
- We proposed using *Self-Training* algorithm as a standard semi-supervised learning

method whose goal is to improve any existing supervised classifier when unlabeled data is available and increase the accuracy prediction.

- We proposed a *Transfer-learning* approach that obtains information from another source model to improve the predictive accuracy of the target learned model.
- Finally, we presented our novel *Intermediate Models* that are used as an input for the final predictive model.

We also made contributions to the understanding of human behavior, such as:

- In healthcare, the scarce data and missing information in existing systems for monitoring human behaviour are often dropped from the researchers in the field. In contrary, we proposed machine learning models that use this scarce data which has been shown to improve the knowledge of monitored subjects and at the same time improving the performance accuracy.
- In bipolar disorder, we proposed extracting and analysing motor activity behaviour in patients from two sources, such as motor intensity and voice features during the phone conversation. To the best of our knowledge, our work is the first in the field combing both features to predict the episodic state in bipolar disorder.
- Finally, in work-related stress, despite the methods proposed to build accurate models, we proposed new methods for extracting contextual data from smartphone raw data and interpreting similarity or de-similarity of subjects behaviour during the monitoring days.

All the contributions made by this research work push against the boundaries of how researchers should design a system in the future. We have taken the first step toward handling scarce information aiming at improving predictive models. With the proposed approaches, we were able to provide better predictive models in understanding individuals' behaviour, as well as observing similarities across group behaviour.

The approaches proposed for handling scarce data have instilled in us a belief, that following these approaches may also contribute in addressing open problems that scarce information brings to the fore. We believe that implementation of proposed approaches and the operation of these systems need a broader perspective. Thus, we hope that the example of recognizing behaviour-related pattern in subjects participated in our studies represents only the beginning of how future systems can be improved. The contribution of this thesis opens up numerous opportunities to design effective intervention for aiding individuals wellbeing as well as improving healthcare services.

Our work in general shows how scarce information was handled enables smartphone classification to be more robust and efficient. We hope this dissertation have provided a motivation to researchers for seeking for better solutions to address scarce information that can further impact classification systems in different domains.

Finally, using the features extracted from speech and acceleration signals during the phone conversation, we were able to classify bipolar disorder episodic states and perceived stress level from extracted features and less obtrusive than current standards in monitoring motor-activities. These methods can be also combined with other stream of sensory data during phone-conversations that may help us further understand individuals behaviour.

7.2 Limitations

Thesis demonstrates the importance of employing machine-learning techniques to handle scarce data collected from smartphone sensors to monitor behavior and mental-health. There are, however, several issues associated with the use of proposed approaches and a limited number of participants. Following limitations has been identified:

7.2.1 Limited number of participants

In this research work, we have a limited number of subjects that participated in the studies. It is obvious that having larger number of subjects and acquiring continuous data that involve long-term continuous observations of subject would increase of statistical significance and precision accuracy.

In addition, data was collected from a specific population. In bipolar disorder, only 5 of the patients were involved in a phone conversation in different stages of disease. The remaining patients *(N=7)* were either missing sensory information or involvement a phone conversation was only in one stage of disease. Furthermore, trial period is another limitation of this thesis, thus, collecting long-term continuous data of patients may increase the knowledge of depressive or manic symptoms. In addition, bipolar disorder patients were included at the study at the beginning of their course of treatment, which limits investigating course of illness.

With regard to the stress predictions, there were 30 subjects from two different organisations and from a specific location. This may limit the findings since perceived stress differ from other group of population or other working environments (non-related to IT or logistics). However, our methods has been shown to be feasible, which potential could be carried over to other group of populations. In this thesis, we have demonstrated approaches that could address issues of scarcity information, however, we did not provide any methodology for an efficacious intervention.

7.2.2 Feedback solutions

Data acquired from the systems was secured into the servers during the trials to protect their privacy and they were analysed off-line. At this stage of our research, we were interested to evaluate proper features and algorithms. However, next stages of our research, features extraction and algorithm performance should be performed directly on the smartphones or pre-processed in a server side and provide a feedback to the participants smartphone.

7.2.3 When to use proposed approaches?

In this thesis, *Intermediate Models* has been suggested to improve the performance accuracy of the final model. Combining TL and SSL methods can assist in building a self-learning system that reduces user burden for labeling their wellbeing in daily basis.

In principle, using SSL to improve a classifier $C : L \rightarrow U$ while involving large amounts of unlabeled data compared to having small amount of labeled instances. However, SSL methods may fail to improve the classification performance or either fail completely when there are no sufficient labeled classes. The reason for that is that unlabeled instances with lower weights are included into the labeled data, thus, leads to decrease of classification performance or amplifies noise in labeled data. Therefore, having all the classes before building the training models.

In this research, TL approach has been applied in dataset collected from normal subjects at their working environments, with the assumption that participants may perceive similar stress. Using these methods, has been shown to improve the classification performance for the subjects with scarce data. However, these methods are not recommended applying in mental-disorders with different cognitive impairments. For instance, in bipolar disorder patients, mood alternate between elevated and depressed over time and no patient have similar episodic state to the others. Thus, applying transfer learning methods may fail in state prediction.

7.3 Future research work

With the proposed methods, we aimed at handling scarce data to improve detection of behaviour patterns in monitored participants. We truly believe that *Intermediate Model* approach combined with semi-supervised learning and transfer-learning methods could play crucial role in future effort for creating accurate predictive models including the healthcare monitoring, especially for remote-monitoring of individuals.

However, there are several research directions that we are planning to follow in the near future. Although the advances put forth in this research work, some issues still remain. In the following, we briefly summarize a few of these future challenges below:

7.3.1 Feature selections

In this research work, we have considered a large number of features, many of which were reported as useful in the literature, however, other features could be considered as well as using feature selection algorithms. The key question in machine learning is how to produce the instances by a vector of features and reduce major computational difficulties that may lead to poor prediction accuracy (Beniwal and Arora, 2012). Thus, in the monitoring systems where real-time processing is required, applying this step in order to improve the efficiency and effectiveness is needed. In Chapter 3, we have reviewed literature of feature selection as an important step and the way it is used to remove redundancy and noise from collected raw data. Many research work in a field of Ubiquitous Computing consider this step as compulsory and selecting features with higher rank scores should be distinctive features before feeding them to the classifier, as it shown in the review of Mehmood et al., 2012.

However, in this research work, our analysis also discerns which features contribute most to behaviour changes detection. In bipolar disorder, motor activity and speech features tended to be the strongest predictors of patients episodic state. However, in future work all the features which have no influences on the class information will be removed as irrelevant features. In predicting stress at work, we have used several features categories, such as *location*, *physical activity*, *motor activity*, *social-interaction* and other features. In the Section 6.1.2.1 we have demonstrated most important features using Multiple-regression analysis (Efroymson, 1960) to analyse how each variable category has effect into the correlation, thus, in accuracy performance to predict stress at work. Therefore, we will consider applying existing feature selection methods (*i.e.*, *PCA*-principal component analysis (Malhi and Gao, 2004), *ICA*-independent component analysis (Fortuna and Capson, 2004), and *KPCA*-kernel principal component analysis (Cao et al., 2003)).

In this line, we also plan to continue to explore our work on transfer learning along the following directions:

- ◆ We plan to apply dimensionality reduction and feature selection methods using transfer learning approaches. However, there are several research issues that are needed to be addressed, like, i) how to determine the number of the reduced dimensionality, and ii) how to develop an efficient algorithm for automatic self-learning transfer learning from scarce data similar to recent study in (Raina et al., 2007).
- ◆ Most of research work in transfer learning assumed that data from different domains must be independent distributed. However, in real-life settings, such as prediction content of users social networks, generally data are found often relational, which in turn presents a major challenge to transfer learning (Kumaraswamy et al., 2015). In future work, we plan to apply the dimensionality reduction in a relational learning

manner, and in this way we make sure that the data in source and targeted subjects can be relational instead of being independent distributed.

- We also plan to research the negative transfer learning issue. As shown in the Chapter 6, when the source and target tasks are dissimilar, all the knowledge extracted from a source task did not help improve the performance of the targeted task. Therefore, avoiding negative transfer and ensure that the safe transfer of knowledge to targeted domain is crucial in transfer learning.

7.3.2 Future challenges using semi-supervised learning

As discussed throughout this thesis, we demonstrated the use of semi-supervised learning methods to learn from unlabeled data. The performance of semi-supervised algorithms may suffer if the wrong algorithm is chosen, thus, a secure semi-supervised learning algorithms have to ensure their performance which is at least as well as supervised learning.

In this research work, we have analysed the performance of semi-supervised learning algorithm (namely Self-training algorithm) for two specific domains. In order to ensure the performance of Self-training, in this research work we used only decision trees algorithms and all reported results that were obtained from both domains used default parameters of classifiers. In future work, more sophisticated semi-supervised algorithms (*i.e.*, *Co-training*) with other algorithms using different parameters of classifiers, can be used to take advantage of the available unlabeled data. Using Co-training is slightly similar to Self-training approach, however, a critical difference from Self-training is that Co-training uses two classifiers instead of one and operates on a different view of the same instance. The strength of Co-training is that a classifier trained on the first view assigns predicted labels and are given to the classifier that operates on the second view or other way around. The main idea is that a classifier trained on the first view assigns predicted labels, which are given to the classifier operating on the second view, and contrariwise (Blum and Mitchell, 1998).

Using the *Co-training* it is expected that better results can be obtained with a careful tuning of parameters of the classifiers. In addition, the classifier may be improve the accuracy performance by adding intermediate model weights in different stages of model building of co-training classification. We showed with some experiments that our Self-training approaches combined with IM approach performs better than or comparably to existing algorithms which are supervised in nature.

Furthermore, it is advisable to find more theoretically justified form of SSL by choosing automatically among different classification semi-supervised algorithms. The key challenge is to determine a logic prior over classifiers of using different types of SSL learning in order to define a proper likelihood function. Finally, future challenges using SSL are when researchers are able to exploit unlabeled data without being experienced in machine learning or adapting the development of SSL into their studies.

7.3.3 Future challenges using multi-label classification

In this thesis, we proposed using IM that are generated from one variable of the self-reported questionnaire at a time. IM proposed in this research work assumes that each questionnaire variables can be obtained independently from the values of the other questions. In order to compare the results from proposed approach, in our future work, we would like to explore the use of multi-label classifiers (Tsoumakas and Katakis, 2006), where a set of classes (*i.e.*, all the variables associated with the questionnaires) can be predicted at the same time, and where some dependencies between them can be incorporated.

The main advantage of this method is that many binary classifiers can be readily used to build a multi-label learning models. However, using this method ignores the underlying mutual correlation among different label, however, in practice could have significant contributions to the classification performance (Zhu et al., 2005). Another disadvantage using multi-labeled classifier for analysis of data of individuals monitored in healthcare using the self-reported questionnaires rates (*e.g.*, rating their emotional status $\{1, .., 5\}$) limits defining labeling of instances related to their wellbeing (*e.g.*, low, moderate, high) into two binary levels $\{0, 1\}$ and inter-label correlations between labeled variables.

We would also like to combine multi-label learning approach with Semi-supervised algorithms to exploit unlabeled data information and develop more robust predictive models. Semi-supervised multi-label learning is proposed in (Liu et al., 2006), were labeled *(l)* instances (x_1, y_1),···,(x_l, y_l), and unlabeled *(u)* instances x_{l+1},···, x_{l+u}, where each x_i = (x_{i1},···, x_{im})^T is an m-dimensional feature vector and each y_i = (y_{i1},···, y_{ik})^T is a k-dimensional label vector. Here, the approach assumes that the label of each instance for each category is binary: $\mathrm{y}_{ij} \in \{0, 1\}$. And $n = l + u$ are the total number of instances, X = (x_1,···, x_n)^T and Y = (y_1,···, y_n)^T = (c_1,···, c_k).

Finally, we would also like applying TL and Ensemble approaches in the models build from multi-label learning and exploit the performance of approach.

7.3.4 Future challenges using Transfer-learning

Following the advances in machine learning framework, we believe that the automatic self-training models is the future of monitoring human wellbeing. Knowledge transfer across individuals that provide different distributions is known problem in machine learning that has not been investigated in details. In the Chapter 6, we have demonstrated using TL in employees with low rate of labeled instances to understand their daily behaviour patterns.

Despite the improvements, there are several open ideas that we are planing to address in future work.

One of the aspect that could improve our work using TL is to analyse in depth other decision trees or other classification algorithms with different parameters that could help us in obtaining better clusters of individuals who behave similarly. Based on the clustering assumption, we would be able to design an effective weighting scheme and achieve better model weights. We assume that tunning decision and applying feature selection could help us building better prediction models for new users with few data. In the future work, we also want to test different levels of granularity for the time dimension to see whether appear during different time intervals. A future line of research is to construct prototype models using information from more individuals, during longer periods of time, and with variations across different wellbeing states.

It is encouraging that combining our simple algorithms, such as Self-training and Transfer-learning methods, as shown in the Chapters 5 and 6, we produce good results across a individuals behaviours. With this thesis, we hope to initiate further research in this area.

7.3.5 User feedback

One of the key role of mental-health services should be to provide meaningful aspects of individual mental-health status, such as changes or improvements of users wellbeing. Involving user in these services could make their lives better. As discussed in Chapter 3, providing feedback information to users may help change bad behaviour patterns and can be used to encourage for improving behaviours. In this research, we have been mainly concerned with building accurate machine learning models to infer human behaviour pattern even in scarce data. However, an obvious consequence of a good inductive models are to develop an application to alert doctors about possible state or other warning signs of their patients. This could be useful to follow up on the effectiveness of medication treatments and it is critical to perform preventive measures on patients in different severe states.

Another aspect that could improve the system providing the feedback-loop between physicians and patients in real-time. This link between the physicians and patients it has been suggested as an essential in an emergency situation in healthcare (Anliker et al., 2004; Bergelson and Naydenov, 2007; Suh et al., 2011), including the intervention for severe mental illnesses (Depp et al., 2010).

In future work, we plan applying our proposed methods to learn automatically individual or groups models to provide real-time feedback information to users with their current state. Feedback loop methods and interaction between the physicians and patients within one closed system would improve the prediction accuracy by adding more knowledge to the system. Finally, building an advanced generic model for all patients and improve healthcare intervention from remote on a daily basis.

7.4 Final summary

To summarize this dissertation, this research work makes contribution to the field of ubiquitous computing and the methods proposed advances the state of the art in healthcare monitoring to address scarce data. The proposed methods used in this thesis contribute to many active areas of research, including problem formulation and the application of these ideas to real-world problems in pervasive health computing, and other challenging domains.

All the effort required for obtaining large amount of labeled data, is clearly becoming important to research for new machine learning algorithms, such as semi-supervised learning, transfer learning approaches that can improve monitoring in real-world learning settings. On the other hand, using these methods increases security in making restrictive assumption about the use of unlabeled data. The work in this research work establishes a major step in this direction, and the future work proposed here may help to grasp the potential of unlabeled datasets.

Systems used in trials have been found to be capable for capturing human behviour patterns in an automatic and unobtrusive manner. We believe that data collected from the systems and the features extracted, provide useful information about individual's behaviour changes and their health status. Using the approaches proposed in this thesis, it is possible to provide a feedback or alert users about their imminent bipolar episode or high stress events. Such a system would provide healthcare professionals with additional information derived from individuals behaviour. It is also important to emphasize that using the Frameworks in Monarca and Turnout-BurnOut may be applicable to other groups or disease with very little changes required.

Finally, we remain with a hope that methods proposed will become a fruitful for both machine learning theory and practical applications in healthcare domain.

REFERENCES

Aharony, Nadav et al. (2011). "Social fMRI: Investigating and shaping social mechanisms in the real world". In: *Pervasive and Mobile Computing* 7.6, pp. 643–659.

AIS (2015). *Effects of Stress - American Institute of Stress*. `http://www.stress.org/topic-effects/.`. Accessed: 2015-AUG-29.

Al-Mardini, Mamoun et al. (2014). "Classifying obstructive sleep apnea using smartphones". In: *Journal of Biomedical Informatics* 52, pp. 251–259.

Alderfer, Benjamin S and Michael H Allen (2002). "Treatment of agitation in bipolar disorder across the life cycle." In: *The Journal of clinical psychiatry* 64, pp. 3–9.

Altman, Naomi S (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". In: *The American Statistician* 46.3, pp. 175–185.

Amini, Massih et al. (2009). "A transductive bound for the voted classifier with an application to semi-supervised learning". In: *Advances in Neural Information Processing Systems*, pp. 65–72.

Anliker, Urs et al. (2004). "AMON: a wearable multiparameter medical monitoring and alert system". In: *IEEE Transactions on information technology in Biomedicine* 8.4, pp. 415–427.

Bakker, J et al. (2011). "What's Your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data". In: *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pp. 573–580.

Balcan, Maria-Florina et al. (2004). "Co-training and expansion: Towards bridging theory and practice". In: *Advances in neural information processing systems*, pp. 89–96.

Bartholomew, John B et al. (2005). "Effects of acute exercise on mood and well-being in patients with major depressive disorder". In: *Medicine and Science in Sports and Exercise* 37.12, p. 2032.

Batterham, Philip J et al. (2009). "Modifiable risk factors predicting major depressive disorder at four year follow-up: a decision tree approach". In: *BMC psychiatry* 9.1, p. 75.

Bellman, Richard (1957). "Dynamic Programming Princeton University Press". In: *Princeton, NJ.*

Belmaker, RH (2004). "Bipolar disorder". In: *New England Journal of Medicine* 351.5, pp. 476–486.

Beniwal, Sunita and Jitender Arora (2012). “Classification and feature selection techniques in data mining”. In: *International Journal of Engineering Research & Technology (IJERT)* 1.6.

Bennett, Kristin P et al. (2002). “Exploiting unlabeled data in ensemble methods”. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, pp. 289–296.

Bergelson, Michael and Narcis M Naydenov (2007). *System and method for real-time remote monitoring of implantable medical devices.* US Patent 7,218,967.

Bernaards, CM et al. (2006). “Can strenuous leisure time physical activity prevent psychological complaints in a working population?” In: *Occupational and environmental medicine* 63.1, pp. 10–16.

Birant, Derya and Alp Kut (2007). “ST-DBSCAN: An algorithm for clustering spatial–temporal data”. In: *Data & Knowledge Engineering* 60.1, pp. 208–221.

Blum, Avrim and Tom Mitchell (1998). “Combining labeled and unlabeled data with co-training”. In: *Proceedings of the eleventh annual conference on Computational learning theory.* ACM, pp. 92–100.

Boersma, Paul (2002). “Praat, a system for doing phonetics by computer”. In: *Glot international* 5.9/10, pp. 341–345.

Bogomolov, Andrey et al. (2014). “Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits”. In: *Proceedings of the ACM International Conference on Multimedia.* ACM, pp. 477–486.

Bongers, Paulien M et al. (1993). “Psychosocial factors at work and musculoskeletal disease”. In: *Scandinavian journal of work, environment & health*, pp. 297–312.

Booch, G. et al. (1999). *The Unified Modeling Language User Guide.* The Addison-Wesley Object Technology Series. Addison-Wesley.

Bopp, Jedediah M et al. (2010). “The longitudinal course of bipolar disorder as revealed through weekly text messaging: a feasibility study”. In: *Bipolar disorders* 12.3, pp. 327–334.

Borriello, Gaetano et al. (2007). “Guest Editors’ Introduction Pervasive Computing in Healthcare”. In: *IEEE Pervasive Computing* 6.1, pp. 0017–19.

Brefeld, Ulf and Tobias Scheffer (2006). “Semi-supervised learning for structured output variables”. In: *Proceedings of the 23rd international conference on Machine learning.* ACM, pp. 145–152.

Breiman, Leo (1996). “Bagging predictors”. In: *Machine learning* 24.2, pp. 123–140.

Breiman, Leo et al. (1984). *Classification and regression trees.* CRC press.

Brodley, Carla E. and Mark A. Friedl (1999). “Identifying mislabeled training data”. In: *Journal of Artificial Intelligence Research*, pp. 131–167.

Cao, LJ et al. (2003). “A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine”. In: *Neurocomputing* 55.1, pp. 321–336.

Caruana, Rich (1998). “Multitask learning”. In: *Learning to learn.* Springer, pp. 95–133.

Chapelle, Olivier et al. (2006a). "Branch and bound for semi-supervised support vector machines". In: *Advances in neural information processing systems*, pp. 217–224.

Chapelle, Olivier et al. (2006b). "Semi-supervised learning". In:

Chipman, Hugh A et al. (2001). "Managing Multiple Models". In: *Eighth International Workshop on Artificial Intelligence and Statistics*. Key West, Florida, USA, pp. 11–18.

Clifton, DA et al. (2015). "Health informatics via machine learning for the clinical management of patients". In: *Yearbook of medical informatics* 10.1, p. 38.

Cochocki, A and Rolf Unbehauen (1993). *Neural networks for optimization and signal processing*. John Wiley & Sons, Inc.

Cohen, Sheldon and Thomas A Wills (1985). "Stress, social support, and the buffering hypothesis." In: *Psychological bulletin* 98.2, p. 310.

Consolvo, Sunny et al. (2008). "Activity sensing in the wild: a field trial of ubifit garden". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 1797–1806.

Constandache, Ionut et al. (2010). "Towards mobile phone localization without war-driving". In: *Infocom, 2010 proceedings ieee*. IEEE, pp. 1–9.

Conway, Terry L et al. (1981). "Occupational stress and variation in cigarette, coffee, and alcohol consumption". In: *Journal of Health and Social Behavior*, pp. 155–165.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

Cover, Thomas M and Joy A Thomas (2012). *Elements of information theory*. John Wiley & Sons.

Cuervo, Eduardo et al. (2010). "MAUI: making smartphones last longer with code offload". In: *Proceedings of the 8th international conference on Mobile systems, applications, and services*. ACM, pp. 49–62.

Dai, Wenyuan et al. (2007a). "Boosting for transfer learning". In: *Proceedings of the 24th international conference on Machine learning*. ACM, pp. 193–200.

Dai, Wenyuan et al. (2007b). "Transferring naive bayes classifiers for text classification". In: *Proceedings of the national conference on artificial intelligence*. Vol. 22. 1. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, p. 540.

Dai, Wenyuan et al. (2008a). "Self-taught clustering". In: *Proceedings of the 25th international conference on Machine learning*. ACM, pp. 200–207.

Dai, Wenyuan et al. (2008b). "Translated learning: Transfer learning across different feature spaces". In: *Advances in neural information processing systems*, pp. 353–360.

Dantzig, Saskia et al. (2013). "Toward a persuasive mobile application to reduce sedentary behavior". In: *Personal and ubiquitous computing* 17.6, pp. 1237–1246.

Dasgupta, Sanjoy et al. (2002). "PAC generalization bounds for co-training". In: *Advances in neural information processing systems* 1, pp. 375–382.

Dempster, Arthur P et al. (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.

Denis, Francois et al. (2002). "Text classification from positive and unlabeled examples". In: *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU'02*, pp. 1927–1934.

Depp, Colin A et al. (2010). "Mobile interventions for severe mental illness: design and preliminary data from three approaches". In: *The Journal of nervous and mental disease* 198.10, p. 715.

Devijver, Pierre A and Josef Kittler (1982). *Pattern recognition: A statistical approach.* Vol. 761. Prentice-Hall London.

Dietterich, Thomas G (2000). "Ensemble methods in machine learning". In: *International workshop on multiple classifier systems.* Cagliari, Italy, pp. 1–15.

Duda, Richard O et al. (2012). *Pattern classification.* John Wiley & Sons.

Efroymson, MA (1960). "Multiple regression analysis". In: *Mathematical methods for digital computers* 1, pp. 191–203.

Ertin, Emre et al. (2011). "AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field". In: *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems.* ACM, pp. 274–287.

Eyben, Florian et al. (2009). "OpenEAR - introducing the Munich open-source emotion and affect recognition toolkit". In: *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on.* IEEE, pp. 1–6.

Faurholt-Jepsen, Maria et al. (2012). "Differences in psychomotor activity in patients suffering from unipolar and bipolar affective disorder in the remitted or mild/moderate depressive state". In: *Journal of affective disorders* 141.2, pp. 457–463.

Ferdous, Raihana et al. (2015). "Does verbal interaction among colleagues affect perceived stress levels?" In: *Proceedings of IEEE International Conference on Communications (ICC).*

Fleshner, F (2005). "Physical activity and stress resistance: sympathetic nervous system adaptations prevent stress-induced immunosuppression". In: *Exercise and sport sciences reviews* 33.3, pp. 120–126.

Fortuna, Jeff and David Capson (2004). "Improved support vector classification using PCA and ICA feature space modification". In: *Pattern recognition* 37.6, pp. 1117–1129.

Fowlkes, Edward B and Colin L Mallows (1983). "A method for comparing two hierarchical clusterings". In: *Journal of the American statistical association* 78.383, pp. 553–569.

Freund, Yoav and Robert E Schapire (1995). "A desicion-theoretic generalization of on-line learning and an application to boosting". In: *European conference on computational learning theory.* Springer, pp. 23–37.

Freund, Yoav, Robert E Schapire, et al. (1996). "Experiments with a new boosting algorithm". In: *Icml.* Vol. 96, pp. 148–156.

Friedman, Nir et al. (1997). “Bayesian network classifiers”. In: *Machine learning* 29.2-3, pp. 131–163.

Fung, Glenn and Olvi L Mangasarian (2001). “Semi-supervised support vector machines for unlabeled data classification”. In: *Optimization methods and software* 15.1, pp. 29–44.

Garcia-Ceja, Enrique et al. (2014). “Detecting walking in synchrony through smartphone accelerometer and wi-fi traces”. In: *European Conference on Ambient Intelligence*. Springer, pp. 33–46.

Glanz, Karen et al. (2008). *Health behavior and health education: theory, research, and practice.* John Wiley & Sons.

Goldman, Sally and Yan Zhou (2000). “Enhancing supervised learning with unlabeled data”. In: *ICML*, pp. 327–334.

Grandvalet, Yves, Christophe Ambroise, et al. (2001). “Semi-supervised marginboost”. In: *Advances in neural information processing systems*, pp. 553–560.

Grunerbl, Agnes et al. (2014). “Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients”. In: *Proceedings of the 5th Augmented Human International Conference*. ACM, p. 38.

Grunerbl, Agnes et al. (2015). *Smartphone-based recognition of States and state changes in bipolar disorder patients.*

Guidoux, Romain et al. (2014). “A smartphone-driven methodology for estimating physical activities and energy expenditure in free living conditions”. In: *Journal of Biomedical Informatics* 52, pp. 271–278.

Gwaltney, Chad J et al. (2008). “Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review”. In: *Value in Health* 11.2, pp. 322–333.

Hall, Mark et al. (2009). “The WEKA data mining software: an update”. In: *ACM SIGKDD explorations newsletter* 11.1, pp. 10–18.

Hansen, René and Mik P Christensen (2011). *Assessing the mood of bipolar patients using speech analysis.*

Harpale, Abhay and Yiming Yang (2010). “Active learning for multi-task adaptive filtering”. In:

Harris, Fredric J (1978). “On the use of windows for harmonic analysis with the discrete Fourier transform”. In: *Proceedings of the IEEE* 66.1, pp. 51–83.

He, Ling et al. (2009). “Stress detection using speech spectrograms and sigma-pi neuron units”. In: *Natural Computation, 2009. ICNC’09. Fifth International Conference on.* Vol. 2. Tianjin, China, pp. 260–264.

Hedelin, Per and Dieter Huber (1990). “Pitch period determination of aperiodic speech signals”. In: *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on.* IEEE, pp. 361–364.

Hernandez-Leal, Pablo et al. (2015). "Stress Modelling Using Transfer Learning in Presence of Scarce Data". In: *Ambient Intelligence for Health*. Springer, pp. 224–236.

Honorio, Jean and Dimitris Samaras (2010). "Multi-task learning of Gaussian graphical models". In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 447–454.

Jaakkola, Tommi et al. (1999). "Maximum entropy discrimination". In:

Jeong, Do-Un et al. (2007). "Classification of posture and movement using a 3-axis accelerometer". In: *Convergence Information Technology, 2007. International Conference on*. IEEE, pp. 837–844.

Joachims, Thorsten (1999). "Transductive inference for text classification using support vector machines". In: *ICML*. Vol. 99, pp. 200–209.

Johnson, Richard Arnold, Dean W Wichern, et al. (1992). *Applied multivariate statistical analysis*. Vol. 4. Prentice hall Englewood Cliffs, NJ.

Johnson, Rie and Tong Zhang (2007). "On the Effectiveness of Laplacian Normalization for Graph Semi-supervised Learning." In: *Journal of Machine Learning Research* 8.4.

Judd, Lewis L et al. (2012). "Prevalence and clinical significance of subsyndromal manic symptoms, including irritability and psychomotor agitation, during bipolar major depressive episodes". In: *Journal of affective disorders* 138.3, pp. 440–448.

Kääriäinen, Matti (2005). "Generalization error bounds using unlabeled data". In: *International Conference on Computational Learning Theory*. Springer, pp. 127–142.

Korabik, Karen et al. (1993). *Stress, coping, and social support among women managers*. University of British Columbia Academic Women's Association.

Korhonen, Ilkka (2004). "Guest editorial introduction to the special section on pervasive healthcare". In: *IEEE transactions on information technology in biomedicine* 8.3, p. 229.

Kriska, Andrea M et al. (2003). "Physical activity, obesity, and the incidence of type 2 diabetes in a high-risk population". In: *American Journal of Epidemiology* 158.7, pp. 669–675.

Kumaraswamy, Raksha et al. (2015). "Transfer Learning via Relational Type Matching". In: *Data Mining (ICDM), 2015 IEEE International Conference on*. IEEE, pp. 811–816.

Lane, Nicholas D et al. (2008). "Urban sensing systems: opportunistic or participatory?" In: *Proceedings of the 9th workshop on Mobile computing systems and applications*. ACM, pp. 11–16.

Lane, Nicholas D et al. (2011). "Bewell: A smartphone application to monitor, model and promote wellbeing". In: *5th International ICST Conference on Pervasive Computing Technologies for Healthcare*, pp. 23–26.

Lawrence, Neil D and Michael I Jordan (2004). "Semi-supervised learning via Gaussian processes". In: *Advances in neural information processing systems*, pp. 753–760.

Lee, Te-Won (1998). "Independent component analysis". In: *Independent Component Analysis*. Springer, pp. 27–66.

Li, Li-Jia et al. (2009). “Towards total scene understanding: Classification, annotation and segmentation in an automatic framework”. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, pp. 2036–2043.

Liao, Xuejun et al. (2005). “Logistic regression with an auxiliary data source”. In: *Proceedings of the 22nd international conference on Machine learning*. ACM, pp. 505–512.

Ling, Xiao et al. (2008). “Can chinese web pages be classified with english data source?” In: *Proceedings of the 17th international conference on World Wide Web*. ACM, pp. 969–978.

Liu, Qiuhua et al. (2008). “Detection of unexploded ordnance via efficient semisupervised and active learning”. In: *IEEE Transactions on Geoscience and Remote Sensing* 46.9, pp. 2558–2567.

Liu, Yi et al. (2006). “Semi-supervised multi-label learning by constrained non-negative matrix factorization”. In: *Proceedings of the national conference on artificial intelligence*. Vol. 21. 1. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, p. 421.

Longstaff, Brent et al. (2010). “Improving activity classification for health applications on mobile devices using active and semi-supervised learning”. In: *2010 4th International Conference on Pervasive Computing Technologies for Healthcare*. IEEE, pp. 1–7.

Lu, Hong et al. (2009). “SoundSense: scalable sound sensing for people-centric applications on mobile phones”. In: *Proceedings of the 7th international conference on Mobile systems, applications, and services*. ACM, pp. 165–178.

Lu, Hong et al. (2010). “The Jigsaw continuous sensing engine for mobile phone applications”. In: *Proceedings of the 8th ACM conference on embedded networked sensor systems*. ACM, pp. 71–84.

Lu, Hong et al. (2012). “StressSense: Detecting Stress in Unconstrained Acoustic Environments Using Smartphones”. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. UbiComp ’12. New York, NY, USA: ACM, pp. 351–360. ISBN: 978-1-4503-1224-0. DOI: 10.1145/2370216.2370270. URL: http://doi.acm.org/10.1145/2370216.2370270.

Luis, Roger et al. (2010). “Inductive transfer for learning Bayesian networks”. In: *Machine Learning* 79.1, pp. 227–255.

Luo, Ping et al. (2008). “Transfer learning from multiple source domains via consensus regularization”. In: *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, pp. 103–112.

Lutgendorf, Susan K et al. (1999). “Life stress, mood disturbance, and elevated interleukin-6 in healthy older women”. In: *The Journals of Gerontology Series A: Biological sciences and medical sciences* 54.9, pp. M434–M439.

Lutz, Rafer S et al. (2010). “Exercise caution when stressed: stages of change and the stress–exercise participation relationship”. In: *Psychology of Sport and Exercise* 11.6, pp. 560–567.

Ma, Tinghuai et al. (2010). “Enlarge the Training Data for Activity Recognition”. In: *Information and Computing (ICIC), 2010 Third International Conference on*. Vol. 4. IEEE, pp. 329–332.

Malhi, Arnaz and Robert X Gao (2004). “PCA-based feature selection scheme for machine defect classification”. In: *IEEE Transactions on Instrumentation and Measurement* 53.6, pp. 1517–1525.

Mansour, Yishay et al. (2009). “Domain adaptation with multiple sources”. In: *Advances in neural information processing systems*, pp. 1041–1048.

Marcu, Gabriela et al. (2011). “A framework for overcoming challenges in designing persuasive monitoring and feedback systems for mental illness”. In: *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011 5th International Conference on*. IEEE, pp. 1–8.

Maslach, Christina et al. (2001). “Job burnout”. In: *Annual review of psychology* 52.1, pp. 397–422.

Mata, Jutta et al. (2012). “Walk on the bright side: physical activity and affect in major depressive disorder.” In: *Journal of abnormal psychology* 121.2, p. 297.

Matic, Aleksandar et al. (2012). “Multi-modal mobile sensing of social interactions”. In: *Pervasive computing technologies for healthcare (PervasiveHealth), 2012 6th international conference on*. IEEE, pp. 105–114.

Matic, Aleksandar et al. (2013). “Virtual uniforms: using sound frequencies for grouping individuals”. In: *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. ACM, pp. 159–162.

Matthews, Mark et al. (2014). “Tracking mental well-being: Balancing rich sensing and patient needs”. In: *Computer* 47.4, pp. 36–43.

Maurer, Uwe et al. (2006). “Activity recognition and monitoring using multiple sensors on different body positions”. In: *Wearable and Implantable Body Sensor Networks, 2006. BSN 2006. International Workshop on*. IEEE, 4–pp.

Maxhuni, Alban et al. (2011). “Correlation between self-reported mood states and objectively measured social interactions at work: A pilot study”. In: *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*. IEEE, pp. 308–311.

Maxhuni, Alban et al. (2016a). “Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients”. In: *Pervasive and Mobile Computing*.

Maxhuni, Alban et al. (2016b). “Stress modelling and prediction in presence of scarce data”. In: *Journal of Biomedical Informatics* 63, pp. 344–356.

Maxhuni, Alban et al. (2016c). “Using Intermediate Models and Knowledge Learning to Improve Stress Prediction”. In:

McNair, Douglas M et al. (1971). *Profile of mood states*. Univ.

Mehmood, Tahir et al. (2012). “A review of variable selection methods in partial least squares regression”. In: *Chemometrics and Intelligent Laboratory Systems* 118, pp. 62–69.

Mezick, Elizabeth J et al. (2009). “Intra-individual variability in sleep duration and fragmentation: associations with stress”. In: *Psychoneuroendocrinology* 34.9, pp. 1346–1354.

Miglio, R (1996). “Metodi di partizione ricorsiva nell’analisi discriminante”. PhD thesis. Dipartimento di Scienze Statistiche, Bologna.

Miglio, Rossella and Gabriele Soffritti (2004). “The comparison between classification trees through proximity measures”. In: *Computational Statistics and Data Analysis* 45.3, pp. 577–593.

Mihailidis, Alex and Jakob E Bardram (2006). *Pervasive computing in healthcare.* CRC Press.

Milczarek, Malgorzata et al. (2009). *OSH [Occupational safety and health] in figures: stress at work-facts and figures.* Office for Official Publications of the European Communities.

Miller, David J and Hasan S Uyar (1997). “A mixture of experts classifier with learning based on both labelled and unlabelled data”. In: *Advances in neural information processing systems*, pp. 571–577.

Miluzzo, Emiliano et al. (2010). “EyePhone: activating mobile phones with your eyes”. In: *Proceedings of the second ACM SIGCOMM workshop on Networking, systems, and applications on mobile handhelds.* ACM, pp. 15–20.

Mitchell, Thomas M (1997). *Machine Learning, 1st edition.* New York: McGraw-Hill Higher Education.

Moore, E et al. (2003). “Analysis of prosodic variation in speech for clinical depression”. In: *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE.* Vol. 3. IEEE, pp. 2925–2928.

Moore, Elliot et al. (2008). “Critical analysis of the impact of glottal features in the classification of clinical depression in speech”. In: *Biomedical Engineering, IEEE Transactions on* 55.1, pp. 96–107.

Morgan III, Charles A et al. (2015). “Symptoms of dissociation in humans experiencing acute, uncontrollable stress: a prospective investigation”. In: *American Journal of Psychiatry.*

Morriss, Richard (2004). “The early warning symptom intervention for patients with bipolar affective disorder”. In: *Advances in Psychiatric Treatment* 10.1, pp. 18–26.

Mortel, Thea F Van de et al. (2008). “Faking it: social desirability response bias in self-report research”. In:

Muaremi, Amir et al. (2013). “Towards measuring stress with smartphones and wearable devices during workday and sleep”. In: *BioNanoScience* 3.2, pp. 172–183.

Mukhopadhyay, Sudipta and GC Ray (1998). “A new interpretation of nonlinear energy operator and its efficacy in spike detection”. In: *Biomedical Engineering, IEEE Transactions on* 45.2, pp. 180–187.

Mukhtar, Hamid and Djamel Belaid (2013). "Using Adaptive Feedback for Promoting Awareness about Physical Activeness in Adults". In: *Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC).* IEEE, pp. 638–643.

Mun, Min et al. (2009). "PEIR, the personal environmental impact report, as a platform for participatory sensing systems research". In: *Proceedings of the 7th international conference on Mobile systems, applications, and services.* ACM, pp. 55–68.

Muthén, LK and BO Muthén (2007). "Mplus". In: *Statistical analysis with latent variables.* 3.

Näätänen, Petri and Veijo Kiuru (2003). *Bergen burnout indicator 15.* Edita.

Naranjo, C et al. (2011). "Major depression is associated with impaired processing of emotion in music as well as in facial and vocal stimuli". In: *Journal of affective disorders* 128.3, pp. 243–251.

Nations, U (2013). *World Population Prospects: The 2012 Revision, Highlights and Advance Tables.*

Nigam, Kamal and Rayid Ghani (2000). "Analyzing the effectiveness and applicability of co-training". In: *Proceedings of the ninth international conference on Information and knowledge management.* ACM, pp. 86–93.

Nigam, Kamal et al. (2000). "Text classification from labeled and unlabeled documents using EM". In: *Machine Learning* 39.2, pp. 103–134.

Occupational Safety, American Institute for and Stress (1999). *Stress at Work.* `http://www.cdc.gov/niosh/docs/99-101/`. Accessed: 2015-AUG-29.

Osmani, Venet (2015). "Smartphones in Mental Health: Detecting Depressive and Manic Episodes". In: *Pervasive Computing, IEEE* 14.3, pp. 10–13. ISSN: 1536-1268. DOI: `10.1109/MPRV.2015.54`.

Osmani, Venet et al. (2013a). "Monitoring activity of patients with bipolar disorder using smart phones". In: *Proceedings of International Conference on Advances in Mobile Computing & Multimedia.* ACM, p. 85.

Osmani, Venet et al. (2013b). "Monitoring Activity of Patients with Bipolar Disorder Using Smart Phones". In: *Proceedings of International Conference on Advances in Mobile Computing and Multimedia.* MoMM '13. Vienna, Austria: ACM, 85:85–85:92. ISBN: 978-1-4503-2106-8. DOI: `10.1145/2536853.2536882`. URL: `http://doi.acm.org/10.1145/2536853.2536882`.

Pan, Sinno Jialin and Qiang Yang (2010). "A survey on transfer learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10, pp. 1345–1359.

Pan, Sinno Jialin et al. (2008). "Transfer learning for wifi-based indoor localization". In: *AAAI Workshop on Trading Agent Design and Analysis.* Chicago, IL, USA, pp. 43–48.

Paoli P. Parent-Thirion, A. (2003). "Working Conditions in the acceding and candidate countries." In: *European Foundation for the Improvement of Living and Working Conditions,* *Office for Official Publications of the European Communities.* 6. URL: `www.eurofound.europa.eu/publications/htmlfiles/ef0306.htm`.

Park, Sungmee and Sundaresan Jayaraman (2003). "Enhancing the quality of life through wearable technology". In: *IEEE Engineering in medicine and biology magazine* 22.3, pp. 41–48.

Pearl, Judea (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference.* San Francisco: Morgan Kaufmann.

Penedo, Frank J and Jason R Dahn (2005). "Exercise and well-being: a review of mental and physical health benefits associated with physical activity". In: *Current opinion in psychiatry* 18.2, pp. 189–193.

Pentland, Alex (2004). "Healthwear: medical technology becomes wearable". In: *Computer* 37.5, pp. 42–49.

Pérez-Espinosa, Humberto et al. (2012). "Acoustic feature selection and classification of emotions in speech using a 3d continuous emotion model". In: *Biomedical Signal Processing and Control* 7.1, pp. 79–87.

Portio, Research (2011). "Portio Research. (2011)". In: *Portio research mobile factbook 2011. Chippenham, UK.*

Postolache, O et al. (2007). "Vital signs monitoring system based on emfi sensors and wavelet analysis". In: *Instrumentation and Measurement Technology Conference Proceedings, 2007. IMTC 2007. IEEE.* IEEE, pp. 1–4.

Prasad, Manishi et al. (2004). "A review of self-report instruments measuring health-related work productivity". In: *Pharmacoeconomics* 22.4, pp. 225–244.

Proper, Karin I et al. (2003). "Effect of individual counseling on physical activity fitness and health: a randomized controlled trial in a workplace setting". In: *American journal of preventive medicine* 24.3, pp. 218–226.

Quer, Giorgio et al. (2013). "Bliss Buzzer, a system to monitor health and stress with real-time feedback". In: *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems.* ACM, p. 83.

Quinlan, John Ross (1993). *C4. 5: programs for machine learning.* Morgan Kaufmann.

Raento, Mika et al. (2005). "ContextPhone: A Prototyping Platform for Context-Aware Mobile Applications." In: 4.2, pp. 51–59.

Raina, Rajat et al. (2007). "Self-taught learning: transfer learning from unlabeled data". In: *Proceedings of the 24th international conference on Machine learning.* ACM, pp. 759–766.

Ranzato, Marc'Aurelio and Martin Szummer (2008). "Semi-supervised learning of compact document representations with deep networks". In: *Proceedings of the 25th international conference on Machine learning.* ACM, pp. 792–799.

Rashidi, Parisa and Diane J Cook (2010). "Multi home transfer learning for resident activity discovery and recognition". In: *Proceedings of International Workshop on Knowledge Discovery from Sensor Data (SensorKDD)*, pp. 53–63.

Rish, Irina (2001). “An empirical study of the naive Bayes classifier”. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. 22. IBM New York, pp. 41–46.

Robusto, CC (1957). “The cosine-haversine formula”. In: *American Mathematical Monthly*, pp. 38–40.

Rokach, Lior (2010). “Ensemble-based classifiers”. In: *Artificial Intelligence Review* 33.1-2, pp. 1–39.

Rosenberg, Chuck et al. (2005). “Semi-supervised self-training of object detection models”. In:

Rosset, Saharon et al. (2004). “A method for inferring label sampling mechanisms in semi-supervised learning”. In: *Advances in neural information processing systems*, pp. 1161–1168.

Roy, Suman Deb et al. (2012). “Socialtransfer: cross-domain transfer learning from social streams for media applications”. In: *Proceedings of the 20th ACM international conference on Multimedia*. Nara, Japan, pp. 649–658.

Runtastic, GmbH (2015). *Runtastic GmbH Home Page. Available online: @ONLINE (accessed on 22 July 2015).* URL: `https://www.runtastic.com`.

Sanches, Pedro et al. (2010). “Mind the body!: designing a mobile stress management application encouraging personal reflection”. In: pp. 47–56.

Scherer, Klaus R (1986). “Vocal affect expression: a review and a model for future research.” In: *Psychological bulletin* 99.2, p. 143.

Shannon, W.D. and D. Banks (1999). “Combining classification trees using MLE”. In: *Statistics in medicine* 18.6, pp. 727–740.

Shi, Xiaoxiao et al. (2008). “Actively transfer domain knowledge”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 342–357.

Sims, J et al. (1999). “The vagaries of self-reports of physical activity: a problem revisited and addressed in a study of exercise promotion in the over 65s in general practice”. In: *Family Practice* 16.2, pp. 152–157.

Singh-Manoux, Archana et al. (2005). “Does subjective social status predict health and change in health status better than objective status?” In: *Psychosomatic Medicine* 67.6, pp. 855–861.

Smith, Jonathan C and Jeffrey M Seidel (1982). “The factor structure of self-reported physical stress reactions”. In: *Biofeedback and Self-regulation* 7.1, pp. 35–47.

Spielberger, Charles D et al. (2003). *Occupational stress: Job pressures and lack of support.*

Srivastava, Lara (2005). “Mobile phones and the evolution of social behaviour”. In: *Behaviour & Information Technology* 24.2, pp. 111–129.

Stevovic, Jovan et al. (2013). “Adding Individual Patient Case Data to the Melanoma Targeted Therapy Advisor”. In: *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*. PervasiveHealth ’13. Venice, Italy: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), pp. 85–

88. ISBN: 978-1-936968-80-0. DOI: 10.4108/icst.pervasivehealth.2013.252076. URL: http://dx.doi.org/10.4108/icst.pervasivehealth.2013.252076.

Suh, Myung-kyung et al. (2011). “A remote patient monitoring system for congestive heart failure”. In: *Journal of medical systems* 35.5, pp. 1165–1179.

Sultan-Taïeb, Héléne et al. (2013). “The annual costs of cardiovascular diseases and mental disorders attributable to job strain in France”. In: *BMC public health* 13.1, p. 748.

Thrun, Sebastian and Lorien Pratt (1998). “Learning to learn: Introduction and overview”. In: *Learning to learn*. Springer, pp. 3–17.

Triguero, Isaac et al. (2015). “Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study”. In: *Knowledge and Information Systems* 42.2, pp. 245–284.

Tsoumakas, Grigorios and Ioannis Katakis (2006). “Multi-label classification: An overview”. In: *Dept. of Informatics, Aristotle University of Thessaloniki, Greece*.

Turner, Kimberly and Nikunj C Oza (1999). “Decimated input ensembles for improved generalization”. In: *Neural Networks, 1999. IJCNN’99. International Joint Conference on*. Vol. 5. IEEE, pp. 3069–3074.

Valstar, Michel François et al. (2011). “The first facial expression recognition and analysis challenge”. In: *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 921–926.

Vancampfort, Davy et al. (2013a). “A review of physical activity correlates in patients with bipolar disorder”. In: *Journal of affective disorders* 145.3, pp. 285–291.

— (2013b). “A review of physical activity correlates in patients with bipolar disorder”. In: *Journal of affective disorders* 145.3, pp. 285–291.

Vanello, Nicola et al. (2012). “Speech analysis for mood state characterization in bipolar patients”. In: *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, pp. 2104–2107.

Vapnik, Vladimir et al. (1997). “Support vector method for function approximation, regression estimation, and signal processing”. In: *Advances in neural information processing systems*, pp. 281–287.

Ventä, L et al. (2008). ““My phone is a part of my soul”–How People Bond with Their Mobile Phones”. In: *Mobile Ubiquitous Computing, Systems, Services and Technologies, 2008. UBICOMM’08. The Second International Conference on*. IEEE, pp. 311–317.

Vinciarelli, Alessandro et al. (2009). “Social signal processing: Survey of an emerging domain”. In: *Image and Vision Computing* 27.12, pp. 1743–1759.

Wang, Yi et al. (2009). “A framework of energy efficient mobile sensing for automatic user state recognition”. In: *Proceedings of the 7th international conference on Mobile systems, applications, and services*. ACM, pp. 179–192.

Ward Jr, Joe H (1963). “Hierarchical grouping to optimize an objective function”. In: *Journal of the American statistical association* 58.301, pp. 236–244.

Weiser, Mark (1991). “The computer for the 21st century”. In: *Scientific american* 265.3, pp. 94–104.

Weiss, Karl et al. (2016). “A survey of transfer learning”. In: *Journal of Big Data* 3.1, pp. 1–40.

Welch, William J (1982). “Branch-and-bound search for experimental designs based on D optimality and other criteria”. In: *Technometrics* 24.1, pp. 41–48.

Weston, Jason et al. (2006). “Inference with the universum”. In: *Proceedings of the 23rd international conference on Machine learning.* ACM, pp. 1009–1016.

WHO (2001). *World Health Organization. The WHO World Health Report 2001: New Understanding-New Hope, World Health Organization. Geneva. Switzerland.*

Winkler, Sebastian et al. (2011). “A new telemonitoring system intended for chronic heart failure patients using mobile telephone technology—feasibility study”. In: *International Journal of Cardiology* 153.1, pp. 55–58.

Witten, Ian H and Eibe Frank (2005). *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann.

Xiang, Shuo et al. (2013). “Multi-source learning with block-wise missing data for Alzheimer’s disease prediction”. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.* Chicago, IL, USA, pp. 185–193.

Yang, Jun et al. (2007). “Cross-domain video concept detection using adaptive svms”. In: *Proceedings of the 15th ACM international conference on Multimedia.* ACM, pp. 188–197.

Yang, Qiang and Xindong Wu (2006). “10 challenging problems in data mining research”. In: *International Journal of Information Technology & Decision Making* 5.04, pp. 597–604.

Yu, Chun-Nam John and Thorsten Joachims (2009). “Learning structural svms with latent variables”. In: *Proceedings of the 26th annual international conference on machine learning.* ACM, pp. 1169–1176.

Zha, Zheng-Jun et al. (2009). “Robust Distance Metric Learning with Auxiliary Knowledge.” In: *IJCAI*, pp. 1327–1332.

Zhao, Peilin and Steven C Hoi (2010). “OTL: A framework of online transfer learning”. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 1231–1238.

Zhou, Yan and Sally Goldman (2004). “Democratic co-learning”. In: *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on.* IEEE, pp. 594–602.

Zhou, Zhi-Hua and Ming Li (2005). “Tri-training: Exploiting unlabeled data using three classifiers”. In: *IEEE Transactions on knowledge and Data Engineering* 17.11, pp. 1529–1541.

Zhou, Zhi-Hua and Jun-Ming Xu (2007). “On the relation between multi-instance learning and semi-supervised learning”. In: *Proceedings of the 24th international conference on Machine learning.* ACM, pp. 1167–1174.

Zhu, Shenghuo et al. (2005). “Multi-labelled classification using maximum entropy method”. In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 274–281.

Zhu, Xiaojin (2006). *Semi-Supervised Learning Literature Survey.*

List of Acronyms

ADS	**General Depression Scale**
AdaBoost.M1	**Adaptive Boosting Classifier**
AI	**Ambient Intelligence**
AP	**Access Point**
BRAMS	**Bech Ragaelsen Mania Scale**
BSDS	**Bipolar Spectrum Diagnostic Scale**
BD	**Bipolar Disorder**
C4.5	**ID3 algorithm**
CD	**Curse of Dimensionality**
DBSCAN	**Density-Based Spatial Clustering of Applications with Noise**
DT	**Decision Tree Classifier**
EL	**Ensemble Learning**
EU	**European Union**
FD	**Frequency Domain**
GPS	**Global Positioning System**
GSR	**Galvanic Skin Response**
GT	**Ground Truth**
HAMD	**Hamilton Rating Scale for Depression**
HRV	**Heart Rate Variability**

IM	**Intermediate Model**
k-NN	**k-Nearest Neighbors**
MAE	**Mean Absolute Error**
MdAE	**Median Absolute Error**
MFCC	**Mel-Frequency Cepstral Coefficient**
NB	**Naive Bayes**
NFC	**Near Field Communication**
PA/pACL	**Physical Activity/Physical Activity Level**
PCA	**Principal Component Analysis**
PE	**Psychiatric Evaluation**
PL	**Perceived Level**
RF	**Random Forest**
RMS	**Root Mean Square**
SL	**Supervised Learning**
SSL	**Semi-supervised learning**
ST	**Self-Training**
SVM	**Support Vector Machine**
TD	**Time Domain**
TL	**Transfer Learning**
U.S	**United States of America**
YMRS	**Young Mania Rating Scale**
ZCR	**Zero-Crossing-Rate**

Appendix A

APPENDIX

A.1 REFEREED PUBLICATIONS AS A Ph.D. CANDIDATE

My publications as a Ph.D. candidate are listed below, including a journal publication that is currently under review. The published work includes ideas that are indirectly related to the central topic of this thesis in *Ubiquitous Computing*.

A.1.1 Journal Publications

- ◆ **Maxhuni, A.**, Hernandez-Leal, P., Sucar, L.E., Osmani, V., Morales, E.F. and Mayora, O., 2016. Smartphone Assessment using Smartphones. ACM Transactions on Interactive Intelligent Systems (TIIS). (in review).
- ◆ **Maxhuni, A.**, Hernandez-Leal, P., Sucar, L.E., Osmani, V., Morales, E.F. and Mayora, O., 2016. Stress modeling and prediction in presence of scarce data. Journal of Biomedical Informatics, 63, pp.344-356. (Maxhuni et al., 2016b).
- ◆ **Maxhuni, A.**, Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O. and Morales, E.F., 2016. Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. Pervasive and Mobile Computing. (Maxhuni et al., 2016a).
- ◆ De Santa,A., Gabrielli, S., Mayora,O. and **Maxhuni, A.**, 2015, Strumenti Innovativi per la Misura dello Stress Correlato al Lavoro, Medico competente Journal, Journal Article.

A.1.2 Conference and Workshop Publications

- ◆ **Maxhuni, A.**, Hernandez-Leal, P., Morales, E.F., Sucar, V.O., Muńoz-Meléndez, A. and Mayora, O., Using Intermediate Models and Knowledge Learning to Improve Stress Prediction. (Maxhuni et al., 2016c).

- Hernandez-Leal, P., **Maxhuni, A.**, Sucar, L.E., Osmani, V., Morales, E.F. and Mayora, O., 2015, December. Stress Modeling Using Transfer Learning in Presence of Scarce Data. In Ambient Intelligence for Health (pp. 224-236). Springer International Publishing. (Hernandez-Leal et al., 2015).
- Garcia-Ceja, E., Osmani, V., **Maxhuni, A.** and Mayora, O., 2014, November. Detecting walking in synchrony through smartphone accelerometer and wi-fi traces. In European Conference on Ambient Intelligence (pp. 33-46). Springer International Publishing. (Garcia-Ceja et al., 2014).
- Osmani, V., **Maxhuni, A.**, Grünerbl, A., Lukowicz, P., Haring, C. and Mayora, O., 2013, December. Monitoring activity of patients with bipolar disorder using smart phones. In Proceedings of International Conference on Advances in Mobile Computing & Multimedia (p. 85). ACM. (Osmani et al., 2013a).
- Matic, A., Osmani, V., **Maxhuni, A.** and Mayora, O., 2012, May. Multi-modal mobile sensing of social interactions. In 2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops (pp. 105-114). IEEE. (Matic et al., 2012).
- **Maxhuni, A.**, Matic, A., Osmani, V. and Ibarra, O.M., 2011, May. Correlation between self-reported mood states and objectively measured social interactions at work: A pilot study. In 2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops (pp. 308-311). IEEE. (Maxhuni et al., 2011).

A.1.3 Other publication work

- Matic, A., **Maxhuni, A.**, Osmani, V. and Mayora, O., 2013, September. Virtual uniforms: using sound frequencies for grouping individuals. In Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication (pp. 159-162). ACM. (Matic et al., 2013).
- Stevovic, J., **Maxhuni, A.**, Shrager, J., Convertino, G., Khaghanifar, I. and Gobbel, R., 2013, May. Adding individual patient case data to the Melanoma Targeted Therapy Advisor. In 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops (pp. 85-88). IEEE. (Stevovic et al., 2013).

A.2 ADDITIONAL RESULTS

– Motor Activity Results in Bipolar Disorder Patients

Table A.1: Intermediate models in supervised setting - Accuracy results from accelerometer frequency domain features and all audio features.

Classifier	**p0201**	**p0302**	**p0702**	**p0902**	**p1002**	**Mean (±SD)**	All P.
AdaBoost.M1	**86.91**	**87.18**	**78.15**	90.02	**86.8**	**85.81 (±4.49)**	**73.53**
Bagging	83.22	63.68	74.17	68.81	82.74	74.52 (±8.57)	73.35
C4.5	81.54	85.04	70.86	**92.72**	84.77	73.10 (±7.93)	73.06
k-NN	74.16	52.99	56.95	67.57	76.65	65.66 (±10.41)	72.83
NaiveBayes	76.85	61.97	62.25	61.54	74.62	67.45 (±7.61)	72.39
RandomForest	85.91	86.75	71.52	88.57	84.77	83.50 (±6.84)	71.94
SVM	83.22	63.68	74.17	68.81	82.74	74.52 (±8.57)	72.09

In order to evaluate the robustness of the proposed approaches, we evaluate the data extracted from motor activity of individuals in working environments. We present results using intermediate models in supervised learning setting. The average and standard deviations of all accuracy values from motor activity features in frequency domain and all audio features are presented in the Table A.1. From the results is easy to note that boosting methods has obtained better results compared to decision trees.

Table A.2: Intermediate Model and Semi-Supervised Learning - Accuracy results from Accelerometer Frequency Domain features and all Audio features.

Classifier	**p0201**	**p0302**	**p0702**	**p0902**	**p1002**	**Mean (±SD)**	All P.
AdaBoost.M1	80.48	87.76	78.65	90.02	85.78	84.54 (±4.82)	84.48
Bagging	75.93	71.82	74.53	68.81	81.78	74.57 (±4.86)	70.86
C4.5	**90.88**	**90.99**	**80.90**	**92.72**	**87.11**	**88.52 (±4.73)**	**88.54**
k-NN	69.52	60.97	66.29	67.57	74.22	67.71 (±4.82)	66.76
NaiveBayes	71.94	68.36	59.18	61.54	74.67	67.14 (±6.63)	68.32
RandomForest	85.04	89.15	75.66	88.57	84.44	84.57 (±5.40)	83.70
SVM	75.93	71.82	74.53	68.81	81.78	74.57 (±4.86)	70.86

Further improvements has been made in semi-supervised setting, where intermediate models improved all accuracy values from motor activity features in frequency domain and all audio features (shown in the Table A.3). Decision trees yielded significantly better performance level when more instances were included into classification.

We have evaluated accelerometer frequency domain and audio spectral features (shown in Table A.3) using intermediate models in supervised learning settings and intermediate models in semi-supervised setting (shown in Table A.4). Decision trees yielded better accuracy using audio spectral features. In this thesis we were focused in addressing scarce

Table A.3: Intermediate models in supervised setting - Accuracy results from accelerometer frequency domain features and audio spectral features.

Classifier	**p0201**	**p0302**	**p0702**	**p0902**	**p1002**	**Mean (±SD)**	All P.
AdaBoost.M1	85.57	**88.03**	**78.15**	89.40	82.23	**84.68 (±4.55)**	83.92
Bagging	84.90	68.38	74.17	66.74	82.23	75.28 (±8.10)	74.14
C4.5	86.91	84.62	70.86	**96.47**	**83.76**	84.52 (±9.16)	**84.32**
k-NN	79.87	54.27	55.63	67.98	80.71	67.69 (±12.68)	67.30
NaiveBayes	77.18	61.54	65.56	61.95	76.65	68.58 (±7.77)	69.72
RandomForest	**87.92**	86.75	72.85	91.48	**83.76**	84.55 (±7.10)	83.34
SVM	84.90	68.38	74.17	66.74	82.23	75.28 (±8.10)	74.14

data, however, in future work we plan to select extracted features in order to improve the classification accuracy.

Table A.4: Intermediate Models and Semi-Supervised Learning - Accuracy results from Accelerometer Frequency Domain features and Audio Spectral features.

Classifier	**p0201**	**p0302**	**p0702**	**p0902**	**p1002**	**Mean (±SD)**	All P.
AdaBoost.M1	88.30	86.37	78.28	89.40	83.56	85.18 (±4.45)	85.28
Bagging	83.63	73.21	75.28	66.74	80.44	75.86 (±6.56)	77.56
C4.5	**94.88**	**90.99**	**85.02**	**96.47**	**87.11**	**90.89 (±4.89)**	**90.82**
k-NN	70.76	53.81	66.29	67.98	76.89	67.15 (±8.47)	66.32
NaiveBayes	78.22	68.13	70.79	61.95	74.22	70.66 (±6.17)	70.28
RandomForest	92.40	89.61	75.28	91.48	83.56	86.47 (±7.14)	85.54
SVM	83.63	73.21	75.28	66.74	80.44	75.86 (±6.56)	77.56

– Result achieved from individuals at Stress@Work

Tables A.5 and A.6 show overall information about phone-conversations and SMS's for entire monitoring weeks of stress. In Table A.5 we demonstrate overall phone usage using the demographics of the individuals participated in the study. As discussed in the Chapter 6, incoming calls where in average higher when they perceived high stress level. Similarly, the length and number of responded SMS's where higher in the days when they perceived stress level.

Table A.7 presents prediction results for every subject monitored stress using the supervised and semi-supervised approaches. Results suggest that using semi-supervised settings can significantly improve the accuracy, reducing amount of scarce data and improving knowledge of individuals behaviour.

In Table A.8 and Table A.6 we demonstrate further details from Pearson correlation and multiple regression of all features extracted from our datasets. Both tables have show high correlation of stress with the objective variables measures.

Similarly, in Table A.10 and Table A.11 we provide results achieved from motor activity features from individuals at working environment with the aim at predicting perceived stress levels. Using intermediate models in semi-supervised setting has been shown to yield the best results. In the tables we provide different set a algorithms where decision trees have shown to perform the best accuracy.

Table A.5: The average phone duration (in minutes), number of calls per day, average length of SMS and number of SMS per day by demographics and perceived level of stress (30-subjects)

	Outgoing Calls			***Incoming Calls***			***Missing Calls***			***Outgoing SMS***			***Incoming SMS***		
Average:	***Duration(Number)***			***Duration(Number)***			***Number***			***Length(Number)***			***Length(Number)***		
	H	**M**	**L**	**H**	**M**	**L**	**H**	**M**	**L**	**H**	**M**	**L**	**H**	**M**	**L**
Age															
26-30	4.1 (2.0)	3.0 ***(2.6)***	2.1 (2.2)	4.4 (1.5)	4.3 (1.7)	4.1 ***(2.0)***	***1.3***	***1.3***	1.2	77.6 (9.5)	49.5 (8.4)	59.2 ***(11.2)***	184.8 (2.3)	356.5 ***(5.2)***	169.6 (2.0)
31-40	6.5 ***(4.1)***	6.5 (2.9)	8.1 (3.6)	8.1 ***(2.5)***	5.1 (2.1)	7.1 (1.9)	2.3	***3.2***	1.6	98.8 ***(2.0)***	81.7 (1.6)	115.2 (1.9)	273.6 (15.0)	205.1 (5.1)	181.0 ***(22.3)***
>41	9.0 (3.2)	7.4 (3.4)	10.0 ***(3.6)***	5.5 ***(2.5)***	7.1 (2.3)	9.0 ***(2.5)***	***2.1***	1.9	1.9	90.4 (3.2)	80.7 (3.4)	58.4 ***(3.6)***	451.6 (1.9)	441.0 ***(2.2)***	505.7 (2.1)
Gender															
– **Men**	8.2 (4.4)	8.5 (4.0)	11.0 ***(4.5)***	10.2 (2.0)	7.3 ***(2.3)***	9.1 (2.1)	***2.7***	1.8	1.9	115.5 (2.1)	71.6 ***(4.0)***	74.3 (2.8)	295.2 ***(3.3)***	288.1 (3.2)	271.5 ***(3.3)***
– **Women**	6.5 (2.7)	6.1 (2.7)	6.5 **(3.4)**	8.4 (2.0)	6.0 ***(2.3)***	6.0 (2.1)	1.5	***4.5***	2.0	33.0 ***(9.6)***	61.0 (6.1)	138.8 (3.4)	170.2 ***(18.1)***	139.3 (6.4)	143.9 (11.0)
Marital Status															
– **Married**	7.4 ***(3.8)***	7.4 (3.5)	8.0 (3.7)	9.5 ***(2.6)***	6.4 (2.2)	7.0 (2.2)	***2.4***	1.6	1.8	88.5 (1.6)	59.9 ***(4.1)***	58.3 (2.9)	278.9 ***(3.2)***	212.6 (2.4)	187.0 (2.3)
– **Never Married**	7.3 (3.5)	8.4 (3.8)	10.5 ***(4.6)***	9.2 (2.1)	7.3 ***(2.7)***	9.0 (2.5)	1.5	***5.1***	2.1	48.5 ***(10.0)***	73.4 (5.6)	130.3 (3.2)	209.4 ***(15.6)***	264.3 (6.5)	245.9 (9.6)
Number of children															
None	6.5 (3.3)	8.2 (3.4)	10.1 ***(4.4)***	9.3 (2.2)	6.2 (2.2)	8.3 ***(2.4)***	1.5	***5.0***	2.1	29.7 ***(10.7)***	55.6 (6.8)	127.0 (3.0)	183.0 ***(17.7)***	259.0 (7.3)	231.2 (8.6)
1-2	8.5 ***(4.4)***	6.2 (3.4)	6.5 (3.5)	10.3 ***(2.8)***	6.4 (2.4)	6.4 (2.2)	***3.2***	1.5	1.8	93.9 ***(3.2)***	48.1 (1.5)	42.2 (1.8)	349.0 ***(4.0)***	254.6 (2.9)	198.9 (2.3)
3-4	3.0 (2.7)	8.3 ***(4.4)***	9.5 (3.9)	11.5 (2.0)	4.0 (2.0)	7.5 (2.0)	1.3	***2.2***	1.7	144.4 ***(1.4)***	47.6 (1.2)	89.0 (1.2)	166.5 (2.5)	279.0 ***(2.8)***	219.1 (2.3)
Organisation															
– **A.**	6.5 (3.4)	7.4 (3.5)	9.3 ***(3.9)***	10.4 ***(2.4)***	6.5 (2.2)	8.0 (2.1)	2.3	***3.9***	1.9	35.6 ***(9.9)***	40.9 (5.7)	62.7 (2.8)	261.1 (12.9)	263.9 (6.5)	201.9 ***(13.3)***
– **B.**	9.3 ***(4.3)***	8.2 (3.7)	9.3 ***(4.2)***	6.5 ***(2.5)***	7.1 ***(2.5)***	8.0 (2.4)	1.6	1.6	***2.0***	115.5 (1.8)	85.5 ***(4.2)***	109.3 (3.1)	207.2 (2.7)	218.2 (2.8)	226.9 ***(3.5)***

(*) **H** - High, **M** - Moderate, **L** - Low Perceived Stress Level.

Table A.6: Overall mean of phone duration (in minutes), number of calls per day, average length of SMS and number of SMS per weekday by demographics and perceived level (PL) of Stress, Job-demand, and Job-control (30-subjects).

	Outgoing Calls			***Incoming Calls***			***Missing Calls***			***Outgoing SMS***			***Incoming SMS***		
Average:	***Duration(Number)***			***Duration(Number)***			***Number***			***Length(Number)***			***Length(Number)***		
	H (PL)	**M** (PL)	**L** (PL)	**H** (PL)	**M** (PL)	**L** (PL)	**H** (PL)	**M** (PL)	**L** (PL)	**H** (PL)	**M** (PL)	**L** (PL)	**H** (PL)	**M** (PL)	**L** (PL)
Perceived Stress															
Monday:	6.4 (3.1)	7.1 (3.7)	7.1 ***(4.2)***	10.3 (2.3)	6.5 ***(2.5)***	6.0 (2.1)	***2.4***	1.5	1.7	44.0 ***(6.4)***	74.8 (3.1)	93.4 (3.5)	252.8 (2.6)	241.2 (7.2)	212.4 ***(10.6)***
Tuesday:	5.0 (2.9)	9.4 (3.6)	8.5 ***(4.3)***	9.3 (2.1)	7.2 ***(2.5)***	6.5 (2.3)	1.6	1.8	***2.0***	47.9 ***(10.7)***	76.1 (4.2)	77.1 (3.5)	178.4 (1.8)	248.1 ***(3.1)***	201.4 (2.9)
Wednesday:	8.0 ***(4.3)***	8.5 (4.0)	11.2 (4.2)	12.1 ***(3.0)***	6.3 (2.4)	9.5 (2.5)	2.5	***5.8***	2.2	50.0 ***(12.6)***	78.6 (6.2)	148.5 (2.8)	330.5 ***(21.1)***	241.5 (3.0)	236.8 (10.4)
Thursday:	8.5 ***(4.2)***	7.25 (3.8)	9.4 (4.1)	8.4 (2.5)	6.4 ***(2.6)***	7.5 ***(2.6)***	***1.9***	1.7	***1.9***	86.3 ***(6.4)***	56.4 (5.1)	94.2 (3.1)	241.4 (2.8)	226.9 ***(6.7)***	242.2 (3.2)
Friday:	9.2 ***(4.0)***	6.2 (2.9)	9.4 (3.7)	7.4 (2.1)	7.3 ***(2.4)***	9.4 (2.3)	***2.2***	1.5	1.7	59.2 (4.1)	47.6 ***(5.0)***	69.3 (2.3)	204.7 ***(21.8)***	221.7 (2.7)	208.8 (2.8)
Perceived job-demand:	8.1 (3.0)	6.4 (3.4)	11.1 ***(4.1)***	7.4 ***(2.4)***	7.5 ***(2.4)***	8.5 ***(2.4)***	***2.5***	1.9	2.0	73.5 (4.4)	72.5 ***(5.8)***	106.5 (2.8)	220.0 (6.3)	216.0 ***(6.8)***	254.0 (5.9)
Perceived job-control:	9.5 (3.5)	7.4 (3.3)	7.3 ***(4.4)***	9.0 ***(2.5)***	7.4 (2.4)	6.3 (2.1)	2.1	1.7	***3.8***	98.5 (3.7)	76.0 (4.8)	57.0 ***(5.0)***	273.6 (5.8)	200.8 (4.5)	170.2 ***(13.0)***

(*) **H** - High, **M** - Moderate, **L** - Low by Perceived Stress, Job-Demands, and Job-Control.

Table A.7: Stress prediction using decision trees before and after applying a Semi-supervised learning (SSL) approach. Overall classes represent overall number of labeled classes in supervised learning and after performing unsupervised learning methods.

Subjects	**Supervised**	**Semi-Supervised**	**Overall Increase in Prediction (%)**
S02	87.50%	**88.89%**	+1.39%
S03	67.24%	**67.69%**	+0.45%
S04	67.35%	**75.00%**	+4.53%
S05	65.31%	**67.24%**	+1.93%
S06	**86.79%**	84.62%	**-2.17%**
S07	94.74%	**95.16%**	+0.42%
S08	62.96%	62.96%	0.00%
S09	53.85%	53.85%	0.00%
S01	61.43%	61.43%	0.00%
S10	73.47%	73.47%	0.00%
S11	33.33%	**76.36%**	+43.03%
S12	52.54%	52.54%	0.00%
S13	41.82%	**56.25%**	+14.43%
S14	**56.45%**	54.69%	**-1.76%**
S15	76.92%	**85.29%**	+8.37%
S16	53.70%	53.70%	0.00%
S17	53.73%	53.73%	0.00%
S18	**85.29%**	83.78%	**-1.51%**
S19	50.00%	50.00%	0.00%
S20	65.51%	**85.48%**	+19.97%
S21	84.29%	84.29%	0.00%
S22	79.10%	**79.17%**	+0.07%
S23	51.67%	51.67%	0.00%
S24	72.86%	72.86%	0.00%
S25	80.00%	**83.67%**	+3.67%
S26	90.38%	90.38%	0.00%
S28	52.94%	52.94%	0.00%
S29	85.37%	**94.83%**	+9.46%
S30	64.29%	64.29%	0.00%
S27	76.19%	**95.71%**	+19.52%
Accuracy Mean(±SD):	67.57% (±15.60%)	**71.73%** (±15.25%)	**4.20%** (±9.52%)
Overall Labeled Instances:	79.97% (1465/**1832**)	**94.00%** (1722/**1832**)	**14.03%**
Precision (%):	65.4%	**68.9%**	
Recall (%):	68.9%	**73.0%**	
F-Score (%):	66.0%	**70.0%**	

Table A.8: Pearson correlations between objective variables and Perceived Stress Level, Negative Mood Score, Positive Mood Score, and Overall Mood Score.

Objective Variables	Stress Level	Negative Mood	Positive Mood	Total Mood Score
Physical Activity Level	_-0.153**_	_-0.112**_	_0.071**_	_0.116**_
Cellular Locations	_-0.070 *_	_-0.070*_	0.033	_0.065*_
Sig. (2-tailed)	0.022	0.024	0.290	0.036
Nr.	1056	1056	1056	1056
Google-Maps Locations	0.051	0.017	_0.079*_	0.033
Sig. (2-tailed)	0.100	0.587	0.010	0.289
Nr.	1057	1057	1057	1057
Wifi Locations	_0.087**_	0.039	_-0.120**_	_-0.093**_
Sig. (2-tailed)	0.001	0.133	0.000	0.000
Nr.	1458	1458	1458	1458
Social-Interaction	0.032	_0.059*_	_-0.142**_	_-0.119**_
Sig. (2-tailed)	0.258	0.036	0.000	0.000
Nr.	1279	1279	1279	1279
Number-Outgoing-Calls	_-0.980**_	_-0.112**_	_0.083**_	_0.121**_
Sig. (2-tailed)	0.001	0.000	0.006	0.000
Nr.	1121	1121	1121	1121
Number-Incoming-Calls	-0.005	_-0.090**_	-0.019	0.05
Sig. (2-tailed)	0.866	0.002	0.522	0.093
Nr.	1122	1122	1122	1122
Missed-Incoming-Call	-0.006	-0.023	-0.012	0.009
Sig. (2-tailed)	0.847	0.441	0.688	0.769
Nr.	1132	1132	1132	1132
Duration-Outgoing-Call	_-0.098**_	_-0.097**_	_0.101**_	_0.123**_
Sig. (2-tailed)	0.006	0.007	0.005	0.123
Nr.	771	771	771	771
Duration-Incoming-Call	0.037	-0.034	_0.091*_	_0.074*_
Sig. (2-tailed)	0.313	0.354	0.013	0.044
Nr.	737	737	737	737
Number-SMS-Outgoing	_0.090**_	_-0.071*_	0.004	0.05
Sig. (2-tailed)	0.002	0.888	0.016	0.092
Nr.	1132	1132	1132	1132
Number-SMS-Incoming	0.006	-0.012	-0.044	-0.016
Sig. (2-tailed)	0.850	0.683	0.143	0.590
Nr.	1126	1126	1126	1126
Length-SMS-Outgoing	_-0.154**_	_-0.153**_	_0.106*_	_0.156**_
Sig. (2-tailed)	0.001	0.001	0.017	0.000
Nr.	505	505	505	505
Length-SMS-Incoming	0.013	-0.028	_0.088*_	0.069
Sig. (2-tailed)	0.737	0.478	0.028	0.069
Nr.	623	623	623	623
Duration-Apps-System	0.008	-0.021	-0.024	0.001
Sig. (2-tailed)	0.759	0.436	0.373	0.976
Nr.	1412	1412	1412	1412
Duration-Apps-Social	0.067	0.067	_-0.218**_	_-0.161**_
Sig. (2-tailed)	0.153	0.152	0.000	0.001
Nr.	460	460	460	460
Number-Apps-System	_-0.129**_	_-0.181**_	_0.194**_	_0.228**_
Sig. (2-tailed)	0.000	0.000	0.000	0.000
Nr.	1294	1294	1294	1294
Number-Apps-Social	-0.060	-0.040	-0.004	0.024
Sig. (2-tailed)	0.203	0.399	0.936	0.610
Nr.	450	450	450	450

– Significant at the level: $\mathbf{*}\rho < \mathbf{0.05}$; $\mathbf{**}\rho < \mathbf{0.01}$.

Table A.9: Significant results from the multiple regression using objective measurements with respect to Stress and Total Mood Score.

	Stress			**Total Mood Score**		
Objective Variables	β	**t**	ρ	β	**t**	ρ
Physical-Activity Levels	-.0111	-5.88	***0.001***	-.0111	-5.88	***0.001***
Cellular Location	-.2333	-2.29	***0.022***	.0376	2.10	***0.036***
Google-Maps Location	.0685	1.65	0.100	.0077	1.06	0.289
WiFi Location	.0057	3.34	***0.001***	-.0041	-3.58	***0.001***
Social Interaction (SI)	.0001	1.13	0.258	-.0008	-4.28	***0.001***
Number-Outgoing-Calls	-.0374	-3.31	***0.001***	.0081	4.07	***0.001***
Number-Incoming-Calls	-.0033	-0.17	0.866	.0058	1.68	0.093
Missed-Incoming-Call	-.0015	-0.19	0.847	.0004	0.29	0.769
Duration-Outgoing-Call	-.0125	-2.73	***0.006***	.0026	3.43	***0.001***
Duration-Incoming-Call	.0048	1.01	0.313	.0016	2.02	***0.044***
Number-SMS-Outgoing	.0188	3.05	***0.002***	.0018	1.68	0.092
Number-SMS-Incoming	.0003	0.19	0.850	-.0001	-0.54	0.590
Length-SMS-Outgoing	-.0015	-3.49	***0.001***	.0003	3.55	***0.001***
Length-SMS-Incoming	.0001	0.34	0.737	.0001	1.72	0.086
Duration-Application-System	.0001	0.31	0.759	.0001	0.03	0.976
Duration-Application-Social	.0001	1.43	0.153	-.0001	-3.47	***0.001***
Number-Application-System	-.0061	-4.69	***0.001***	.0020	8.42	***0.001***
Number-Application-Social	-.0189	-1.27	0.203	.0014	0.51	0.610

Significant at the level: ρ **<0.05**; ρ **<0.01**.

Table A.10: Comparison in terms of accuracy, precision, recall and f-measure of Supervised and Semi-supervised learning using different classifiers for predicting perceived stress.

Algorithms		**Supervised %**	**SSL %**
C4.5	**Accuracy: (Mean±SD)**	59.24 (±15.40)	68.66 (±15.53)
	Precision:	58.43	68.12
	Recall:	59.23	69.07
	F-Measure:	58.68	68.72
Random-Forest	**Accuracy: (Mean±SD)**	**65.50 (±12.72)**	69.21 (±12.91)
	Precision:	61.49	65.76
	Recall:	65.50	69.21
	F-Measure:	61.71	65.56
Naive-Bayes	**Accuracy: (Mean±SD)**	47.93 (±15.14%)	50.08 (±15.72)
	Precision	56.04	57.00
	Recall	47.88	50.09
	F-Measure	47.88	49.39
AdaBoost.M1	**Accuracy: (Mean±SD)**	61.88 (±17.21)	63.51 (±15.57)
	Precision:	54.19	54.91
	Recall:	61.88	63.51
	F-Measure:	56.24	56.91
SVM	**Accuracy: (Mean±SD)**	60.59 (±16.81)	61.70 (±16.53)
	Precision:	48.29	52.96
	Recall:	60.59	61.71
	F-Measure:	51.91	54.09
Bagging	**Accuracy: (Mean±SD)**	64.67 (±15.15)	**69.48 (±13.62)**
	Precision:	58.46	64.85
	Recall:	64.67	69.47
	F-Measure:	60.26	65.56
k-NN (1)	**Accuracy: (Mean±SD)**	55.90 (±14.09)	56.93 (14.22)
	Precision:	55.64	56.31
	Recall:	55.91	56.94
	F-Measure:	55.64	56.52

Table A.11: Comparison of the supervised learning method with the Intermediate Models (SL-IM) and SSL with the Intermediate Models (SSL-IM).

Algorithms		**SL & IM (%)**	**SSL & IM (%)**
C4.5	**Accuracy (Mean±SD)**	67.51 (±15.21)	77.24 (±16.80)
	Precision:	66.20	74.43
	Recall:	67.51	74.66
	F-Measure:	66.47	73.81
Random-Forest	**Accuracy(Mean±SD):**	**71.68** (±12.98)	**78.20** (±12.00)
	Precision:	68.15	73.09
	Recall:	71.49	75.45
	F-Measure:	68.58	72.74
Naïve-Bayes	**Accuracy(Mean±SD):**	57.42 (±16.02)	58.28 (±14.29)
	Precision:	59.37	73.19
	Recall:	55.33	74.54
	F-Measure:	54.93	72.41
AdaBoost.M1	**Accuracy(Mean±SD):**	66.51 (±16.4)	75.18% (±16.76)
	Precision:	59.82	65.14
	Recall:	64.29	56.33
	F-Measure:	59.95	57.36
SVM	**Accuracy(Mean±SD):**	68.70 (±15.84)	77.11% (±15.84)
	Precision:	63.42	69.03
	Recall:	66.60	72.67
	F-Measure:	63.62	68.89
Bagging	**Accuracy(Mean±SD):**	68.70 (±15.80)	77.11% (±15.84)
	Precision:	63.24	68.80
	Recall:	66.42	68.13
	F-Measure:	63.36	67.67
k-NN (1)	**Accuracy(Mean±SD):**	63.32 (±13.53)	70.48 (±17.27)
	Precision:	61.96	73.19
	Recall:	62.82	74.54
	F-Measure:	62.06	72.41