Journal of Biomedical Informatics 63 (2016) 344–356

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Stress modelling and prediction in presence of scarce data

Alban Maxhuni [a,*], Pablo Hernandez-Leal [c,d], L. Enrique Sucar [c], Venet Osmani [b], Eduardo F. Morales [c], Oscar Mayora [b]

[a] DISI, University of Trento, Via Sommarive 9, Povo, Trento, Italy
[b] CREATE-NET, Via alla Cascata 56/D, Povo, Trento, Italy
[c] Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro #1, Sta. María Tonantzintla, Puebla, Mexico
[d] Centrum Wiskunde & Informatica, Science Park 123, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 21 February 2016
Revised 18 August 2016
Accepted 26 August 2016
Available online 31 August 2016

Keywords:
Stress modelling
Transfer learning
Semi-supervised learning
Ensemble methods

ABSTRACT

Objective: Stress at work is a significant occupational health concern. Recent studies have used various sensing modalities to model stress behaviour based on non-obtrusive data obtained from smartphones. However, when the data for a subject is scarce it becomes a challenge to obtain a good model.
Methods: We propose an approach based on a combination of techniques: semi-supervised learning, ensemble methods and transfer learning to build a model of a subject with scarce data. Our approach is based on the comparison of decision trees to select the closest subject for knowledge transfer.
Results: We present a real-life, unconstrained study carried out with 30 employees within two organisations. The results show that using information (instances or model) from *similar* subjects can improve the accuracy of the subjects with scarce data. However, using transfer learning from dissimilar subjects can have a detrimental effect on the accuracy. Our proposed ensemble approach increased the accuracy by ≈10% to 71.58% compared to not using any transfer learning technique.
Conclusions: In contrast to high precision but highly obtrusive sensors, using smartphone sensors for measuring daily behaviours allowed us to quantify behaviour changes, relevant to occupational stress. Furthermore, we have shown that use of transfer learning to select data from close models is a useful approach to improve accuracy in presence of scarce data.

1. Introduction

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 [1–4]. All these negative effects are known to affect the well-being 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 [3].

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 [5]; 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 [6], 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 [7,8], therefore in our work we have used smartphones as non obtrusive approach to collect relevant behaviour data relative to stress levels.

The objective of this study is to model stress levels from different behavioural variables obtained from smartphones and in particular with the limitation that the labelled data for a person is scarce. The ultimate aim is to lessen reliance on self-reported,

* Corresponding author.
E-mail addresses: alban.maxhuni@disi.unitn.it (A. Maxhuni), Pablo.Hernandez@cwi.nl (P. Hernandez-Leal), esucar@inaoep.mx (L.E. Sucar), venet.osmani@create-net.org (V. Osmani), emorales@inaoep.mx (E.F. Morales), omayora@create-net.org (O. Mayora).

http://dx.doi.org/10.1016/j.jbi.2016.08.023
1532-0464/

subjective data for stress measurement and use objectively sensed data to allow continuous measurement of stress levels.

However, data scarcity is a common problem for *in situ* studies, since continuous annotation of current state is required, where the data derived from self-reports are considered ground-truth. For this study we collected data which includes information related to psychological self-assessments (obtained from a standardized validated questionnaire) and sensor data from smartphones used by 30 employees in two different organisations.

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 reduce the amount of unlabelled data, and transfer learning [9] to use previously learned models to improve the model of a person with scarce data.

Our approach learns a model for each subject in the study, this is useful not only to predict the stress levels but 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. This works expands upon our previous work [10] where we investigated the use of a single sensor modality, namely accelerometer to classify stress level.

Our study 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 use self-report instruments.

The rest of the paper is organized as follows. Section 2 reviews related work on stress detection using current technology. Section 3 introduces supervised, semi-supervised and transfer learning approaches. The data acquisition and extracted features are presented in Section 4. Section 5 discusses our proposed approach and results. Challenges and limitations are discussed in Section 6. Finally, the conclusions of the study are presented in Section 7.

2. Related work

Current methods to infer stress are mainly based on physiological signals, e.g., heart-rate variability, blood pressure, body temperatures and respiration [11]. Furthermore, recent works emphasize the importance of measuring physiological signals that would help providing short-term feedback to the users in everyday activities [12]. However, these methods have as drawback that they need to be carried at all times (and in specific places in the body) in order to allow accurately and continuous monitoring. Other approaches have tried to remove this limitations, for example, StressSense [13] proposes a method for detecting stress based on speech analysis and the variation of speech articulation. However, in real-life activities (e.g., crowded environments, noisy conditions) this approach may lead to misinterpretation of speech and therefore of emotion.

As mentioned in the previous section, several works [8,7,14–18] have provided evidence that smartphones are an appropriate tool to collect relevant data that can be used to measure various aspects of human behaviour and classify different mental conditions. In this line, the authors in [19] built a self-tracking system called MoodScope, to help its users manage their mood. The system detects users mood from smartphones usage data (e-mails, call and SMS logs, application usage, web history and location changes). The authors reported an initial 66% accuracy for 32 subjects from their daily mood and improving to 93% after two months of training.

Moreover, Sano and Picard [20] studied physiological signals (GSR) and behavioural patterns of the users from smartphone data aiming at stress detection. They collected data from accelerometer, GSR sensors and other smartphone logs (e.g., screen usage, SMS, calls, locations, etc.). For predicting stress, authors reported an accuracy of 75% and they were able to discriminate stress and no-stress states using information of smartphone usage and user activity (sitting, walking) information. In addition, Carneiro et al. [21] used video cameras, accelerometers and touch screens to extract different features while inducing different levels of stress during electronic game sessions. 19 subjects participated in the study, they used decision trees to build a model to infer stress. The authors achieved an accuracy of 78%.

Another relevant work is from Bauer and Lukowicz [22] whose work aimed at recognizing stress from 7 students before and after the exam period. The assumption is that students are likely to be under stress during the exam sessions. They acquired data from smartphones (location, social proximity through Bluetooth, phone calls and SMS logs) and they reported an average accuracy of 53% during the exam session. In recent work, Bogomolov et al. [23] used call logs, SMS logs, proximity data, and self-reported surveys about personality traits. The authors reported detecting daily stress levels with a 72% accuracy combining real life data from different sources. However, measuring stress in uncontrolled settings poses several difficulties since it requires the efforts of humans about their current perceived stress and other relevant variables. Our previous work on stress detection has focused on correlation of self-reported stress with verbal interaction [24] and app usage on the smartphone [25].

As it can be seen from the above overview, the issue of scarcity of self-reported data has not received enough attention. This is important for real-life studies, since it is often very challenging to collect large amount of self-reported data, especially in the health-care domain. For example, questionnaires are not always answered. Also there are different types of information which are generally scarce (i.e., phone calls data, calendar events) and therefore there will be few samples. Hence, to overcome these issues, in this study we use (1) semi-supervised learning methods, that combine the information in the unlabelled data with the explicit classification information of labelled data to improve the classification performance [26] and (2) transfer learning, which is used to improve the accuracy when available data is limited but when related models are available [9]. Although the usage of transfer learning is gaining huge interest in research, including in health-care (e.g., transfer learning for activity recognition [27]), to date there is no previous work on transferring knowledge in the problem of stress detection as we propose it in this work.[1]

3. Learning from data

The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience [29]. An important area of machine learning is called supervised learning.

[1] This work extends our previous conference paper [28] with two new approaches for transfer learning.

3.1. Supervised learning

One important task of supervised learning is classification, where usually the data is known before the learning task starts, which is called *offline* learning. Data consists of a set of examples containing a feature vector $\mathbf{x_i}$ and a label (class) y_i. A supervised learning algorithm produces a function $g : X \rightarrow Y$, with X and Y input and output spaces, respectively. There exists different techniques for performing classification such as Bayesian networks [30], support vector machines [31] and decision trees [32].

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 [29]. 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.

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 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 [32]. In our domain, trees are useful to represent how a person is affected by stress. For example, in Fig. 1 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. Four common measures are:

- *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 [33] and measures based on comparing the prediction results [34]. Miglio and Soffritti presented a dissimilarity measure that can combine the structure (the attributes of the nodes) and predictive (the predicted classes) similarities in a single value [35]. Let D_i and D_j be two trees with H and K leaves respectively used to classify n observations. We label $1, \ldots, H$ the leaves of D_i, and $1, \ldots, K$ the leaves of D_j to form the matrix:

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

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

The dissimilarity measure is defined as:

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

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

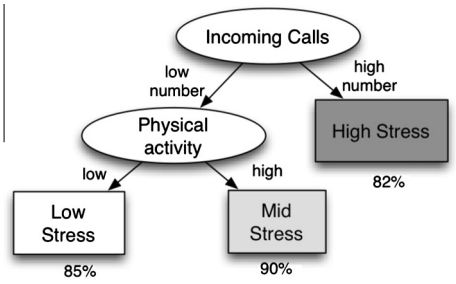

Fig. 1. An example of a decision tree that classifies the level of Stress of a subjects. Ovals represent decision nodes. Rectangles are leaves (terminal nodes) that give the classification value, in this case they represent low, mid or high level of stress. Below each leaf accuracy is presented as a percentage.

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

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 into:

$$s_h = \max\{s_{hk} \; k = 1, \ldots, K\}. \quad (2)$$

The coefficient $\alpha_h = q - p + 1$ is a dissimilarity measure computed between a leaf of D_i and with respect the leaf identified by Eq. (2) of D_j. When the 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 is then:

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

Thus, the normalized version of the dissimilarity is

$$d_n = \frac{d(D_i, D_j)}{\max d(D_i, D_j)} \quad (3)$$

where a $d_n = 0$ represents that the trees are very similar[2] and $d_n = 1$ that they are totally dissimilar.

Now, we present some trees with results using the dissimilarity measure presented in Eq. (3). We refer to the reader to [35] for a more detailed example. Fig. 2(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, Fig. 2(c) and (d) depict highly similar trees, ($d = 0.0$), note that the attributes in the nodes are the same (even when the split value is different they are considered the same).

3.2. 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 a single model [36].

In particular, one ensemble method commonly used is called *random forests* [37] 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. When dealing with real-world data it is likely to have missing data,

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

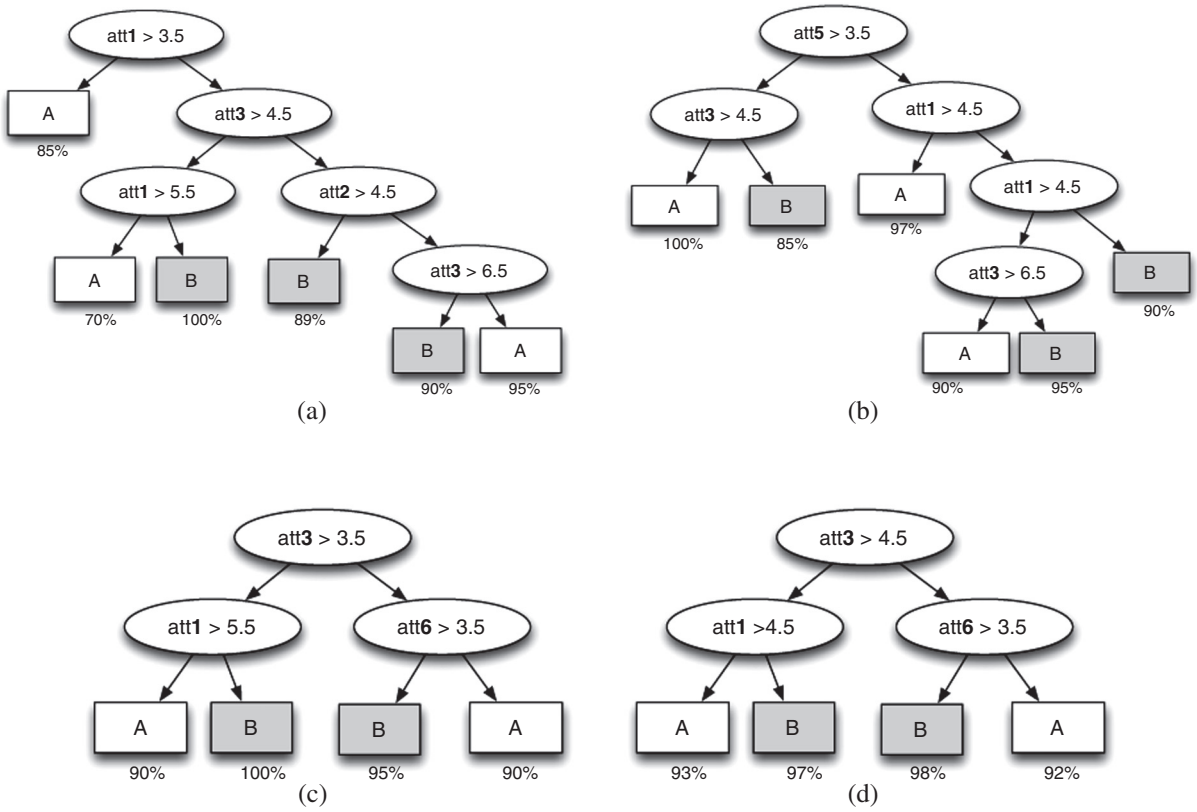

Fig. 2. Example of highly dissimilar decision trees (a) and (b) using the measure in Eq. (3) (since their paths and predictions differ); in contrast (c) and (d) depict highly similar trees since the attributes in the nodes are the same and the predictions are similar.

some techniques from machine learning that deal with this problem are called semi-supervised learning techniques.

3.3. Semi-supervised learning

In our domain, the efficiency of supervised learning is highly dependent on labelled instances that derive from human observer or from the participants themselves using the self-reported questionnaires. However, when it comes to monitoring human behaviour *in situ*, there are several issues where traditional supervised methods fail. This is because having a reasonable size of labelled instances is difficult, expensive and time consuming to obtain since they require human annotators or filling self-reported questionnaires. To address this issue, we can use semi-supervised learning methods that use both labelled and unlabelled data to construct a classifier and improve the classification performance. There are a number of different algorithms for semi-supervised learning, some are designed specifically for a classifier such as semi-supervised SVMs (S3SVM) and transductive support vector machine (TSVMs) [26]. Others offer a general approach for any classifier (e.g. self-training, co-training, and boosting [26,38]).

In this study, we focus on the self-training algorithm [26] that uses its own predictions to assign values to unlabelled data that achieved higher confidence in predictions (in our study we use confidence $\geqslant$80%). The unlabelled data with high confidence in its predicted class is added, with its class, to the labelled data. This new augmented labelled data is used to induce a new model from which new predictions over the reduced unlabelled data are produced (see Algorithm 1). The procedure is repeated until there are no more instances above the threshold value or until the unlabelled data becomes empty. Adding new labelled instances acquired from unlabelled data, is often shown to achieve a better accuracy than supervised learning that uses only the labelled data.

Algorithm 1. Self-training

Input: $\mathbf{L} = (\mathbf{x}_i, \mathbf{y}_i)$; *set of labelled instances*
$\mathbf{U} = (\mathbf{x}_i, ?)$; *set of unlabelled instances*
$\mathbf{T}$; *threshold for confidence*

1 **while** $U \neq \emptyset$ *or* $U' \neq \emptyset$ **do**
2 Train a classifier **C** with training data **L**
3 Classify data in **U** with **C**
4 Find a subset of **U'** of **U** with the most confident scores (confidence > **T**)
5 **L** + **U'** $\Longrightarrow$ **L**
6 **U** - **U'** $\Longrightarrow$ **U**

3.4. 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 labelled data is often very difficult and expensive to obtain in many domains. A system with the capability to use not only labelled but also unlabelled 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 labelled and unlabelled 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 [27]. 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 [39] in different domains such as social networking [40], text classification [40], image classification [41] and indoor and outdoor localization problems [42]. While these are only a handful of examples, TL has been used in many other applications as shown in the surveys in [9,43]. However, in the health-care domain, the use of TL is still in its infancy. For our work the *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 study, 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 2 where decision trees have been used to induce models of the subjects.

Table 1
Study demographics of the subjects in our study.

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%)
	Average (±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%)

Algorithm 2. Transfer Learning used in our study 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 $sim(M_T, M_i)|M_i \in M_{all}$
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
  $M_{sim} = \{\bigcup_i M_i | sim(M_T, M_i) > Th\}$)
  Select $D = \{\bigcup_i D_i |\ D_i$ wasusedtoinduce $M_i \in M_{sim}\})$
  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 $D = \{\bigcup_i D_i | D_i$ wasusedtoinduce $M_i \in M_K\})$
  Let $D' = \{\bigcup_i$ sample $D_i \in D \propto sim(M_T, M_i)\}$
  Induce new model M_T with $D_T \cup D\prime$
else if Ensemble then
  Select the L most similar models M_L = first L elements in M_{all}
  Create a weighted ensemble of models
  $\{M_T \bigcup_{i=1}^{L} w_i M_i | w_i = sim(M_T, M_i) \wedge M_i \in M_T\}$
end if

Next we review the data used in the study and the features extracted, after that we present the proposed transfer learning approach.

4. Data acquisition

In this study, we collected data from 30 healthy employees of two organisations located in the North-eastern part of Italy for a period of 8 weeks. Table 1 provides a summary of employees' demographics. Note that there is a balanced mix of gender, age and education level, marital status and number of children among the subjects in the study.

All employees were given a smartphone[3] where the application used for this study collected data continuously as a background application. The extracted features for each subject are categorized into two types, the first group contains *subjective* information obtained from the self-reported questionnaires, that include mood and work-relevant stress items. The second group of variables includes information of user's behaviour that was collected from the smartphone sensors during work hours, these are called *objective* variables.

4.1. Self-assessment questionnaires

Self-reports include subjective information related to subjects stress and mental state. In order to collect information relevant to 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

[3] Samsung Galaxy S3 mini 32 GB.

Table 2
Overall number and percentage of stress-responses.

Variable	Level	Nr. responses (%)	Nr. subjects
Perceived Stress	*High*	325 (22.18%)	27
	Moderate	515 (35.15%)	30
	Low	625 (42.66%)	30
	Total	1465 (100.00%)	30

related to work stress. The questionnaire developed is clinically validated to capture subjects perceived stress and mood states of the employees at work. Questionnaires were organized to prompt automatically three times a day (9am -at the beginning of working day, 2 pm -after lunch, and 5 pm -before leaving workplace). The questionnaire used for this study was derived from the POMS (Profile of Mood State) scale [44] which has two dimensions related to mood states: 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. In our study, each item had five response alternatives, which assessed five stress-related factors on a Likert scale ranging from 1 to 5. The answers were stored on the mobile device and constituted part of the analysis. The first section of the questionnaire, collected information about occupational health outcomes of the participants (i.e., job induced stress, job-control, job-demand and energy perceived) during working days. The second section includes the scales to measure mood such as feelings of anxiety, cheerfulness, friendliness, sadness, angriness, and quality of sleep.

In Table 2 we present the overall stress responses for the whole period (8 weeks), where we include only the questionnaires obtained from (2:00 pm and 5:00 pm). The total number of responses was 1465. In order to simplify the measurements of the work-related stress, we have classified the stress-level into three classes: $\geqslant 2$ as *Low-Stress*, 3 as *Moderate-Stress*, and $\leqslant 4$ as *High-Stress*. Results show that during the entire monitoring period, 27-subjects perceived *High-Stress* at some point.

4.2. Objective data acquisition and feature extraction

In Table 3 we provide an overview of the sensors and features extracted from smartphones acquired for the study. These are divided into four categories:

- **Physical Activity Level:** Physical activity and the impact of sedentary behaviour in the development of psychological complaints has been gaining a lot of interest over the last decades [50]. Psychological stress has been demonstrated to decrease physical-level and physical-wellness, e.g., experiencing fatigue, weakness [51], reducing frequency and duration of physical activity of the employees [52]. Hence, in our study we want to explore the existing association between objectively measurements of physical activity and subjective reported stress level in working environments. We measured the level of activity using accelerometer data capturing 3-axial linear acceleration continuously at a rate of 5 Hz, which was sufficient to infer physical activity levels. For extracting features, we used the method developed in the framework in [45] and we measured the magnitude and the variance sum of 26 s (non-overlapping fixed length windows of $n = 128$ samples) accelerometer readings. Each segment was classified into *"high"-(variance sum $\geqslant 7$)*, *"low"-(variance sum $\geqslant 3 < 7$)*, and *"none"-(variance sum < 3)* activity levels. Activity counts were measured in percentage using Eq. (4) and were divided into periods of 9am–2 pm and 2 pm–5 pm.

$$\mathbf{pACL}_{(h,l,n)} = \frac{Number\ of\ High\ Activities\ (h)}{Total\ Classified\ Activities\ (h,l,n)} X100\% \qquad (4)$$

- **Location:** Stress can produce behaviours such as smoking, caffeine consumption and skipping lunch [53]. Thus, it is important to analyse locations of subjects with the focus in understanding frequent locations changes during working hours. Location patterns and the location changes were measured using the list of WiFi access points available with their respective BSSID address, cell tower location and Google location information (latitude, longitude). We performed clustering for WiFi by means of the received signal strength (RSS) from each access point (AP). Density-based clustering (DBSCAN) [46] was used to obtain a number of different locations (clusters) in hourly basis. Similarly, DBSCAN was used to cluster Google location and cellular tower location. Location information was clustered in hourly basis. Our objective is to test whether subjects show changes of location in each interval (9am–2 pm and 2 pm–5 pm). For this we compared locations every hour counting when different clusters appeared with respect to the previous hour.
- **Social Interaction (SI):** Social isolation and withdrawal have been associated with perceiving high stress in daily activities [54,55]. Therefore, we used the microphone embedded on smartphones to capture verbal interaction within the employees when they where involved in conversation in a close proximity. We have extracted two main audio features (Pitch [48] and Mel-MultiBand Spectral Entropy Signature (Mel-MBSES) [49]) to perform speech recognition. We built a SVM [31] classifier using MEL-MBSES coefficients trained on frames coming from 3 min of voiced data and 3 min of background data. We sampled audio frequency of 8000 Hz and set a frame every

Table 3
Number of sensors used and features extracted from smartphones for this study.

Category	Smartphone Sensors	Attributes (Feature - extracted)
1. Physical Activity Level	Accelerometer	- 3-Axis (Magnitude) - 3-Axis (Variance Sum [45])
2. Location	Cellular	– CellID and LACID (Number of clusters (DBSCAN) [46])
	WiFi	– Access Points (Number of clusters (DBSCAN) [46])
	Google-Maps	– Latitude and Longitude (Number of clusters (DBSCAN)[46], and distances [47])
3. Social Interaction	Microphone	– Proximity Interaction – Pitch [48], Mel-MBSES [49]
4. Social Activity	Phone Calls	– Number of Incoming and Outgoing Calls – Duration of Incoming and Outgoing Calls – Duration of Incoming and Outgoing SMS – Most common Contact-SMS
	Application usage	– Number of used applications (Social, System) – Duration of used applications (Social, System)

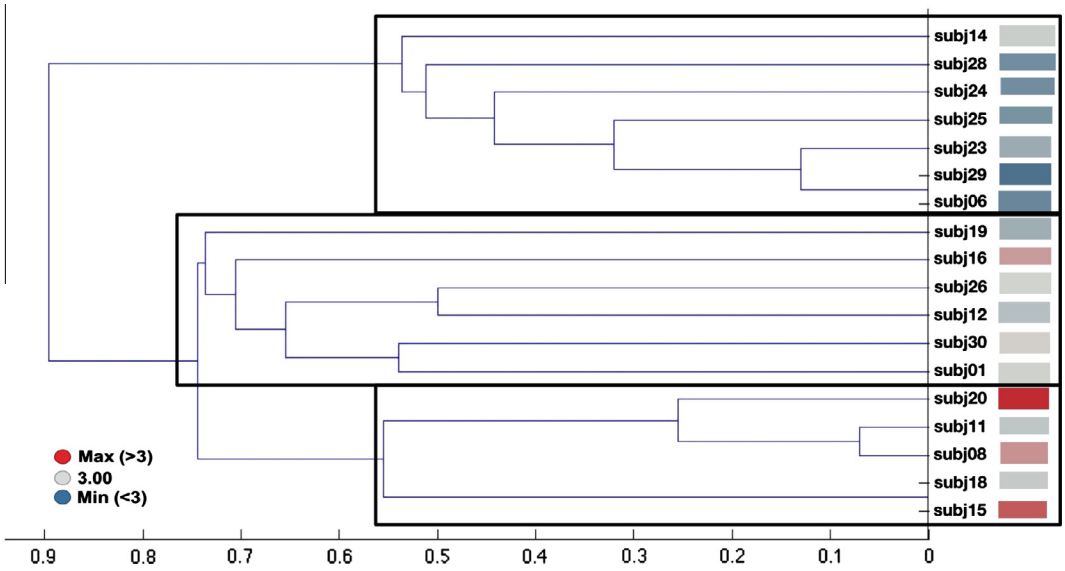

Fig. 3. Dendrogram obtained by computing similarities between models of each subject (using only 18 subjects). Three major clusters can be noted, colour boxes correspond to average stress for different subjects (best seen in colour). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

256 samples where we calculated Pitch and Mel-MBSES features for each frame, then each frame is labelled either as human voice or not a human voice. Approximately every 0.7 s (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) of conversations as shown in Eq. (5):

$$SI = \sum_{i=1}^{n} \frac{\mathbf{True}_{Classified}}{\mathbf{Total}_{Classified}} \times 100\% \quad (5)$$

- **Social Activity:** Since social interaction includes not only face to face conversations but also phone calls and messages, we extracted the number and length of conversations (incoming, outgoing and missing), SMS messages (incoming and outgoing) performed by the employees.[4] Finally, another aspect that may have impact on the stress levels is application usage of the smartphones. For this, our software captures the time and duration when an app is in use. Then, we extracted 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:** Viber, WhatsApp, Facebook, Skype and other user downloaded apps (e.g., games other entertainment apps).

Due to the variety of the selected sensors there is a wide difference in the number of data available for each one. For example, the accelerometer has a rate of 5 Hz which results in ≈12 million points for the 30 days for each person. In contrast, the number of phone calls and SMSs varied in the range of (50–200) per person. The number of location per person varied in the range of (20–100) per subject and the frequency of application usage in the range of (20–30) per subject. Since our objective is to predict stress levels in periods of the working day, we need to align each sensor data to those two periods from 9am to 12 pm and from 1 pm to 5 pm to match the responses obtained from the questionnaires.

[4] To protect users privacy, phone call events were anonymised registering only the five last numbers of each calling or called contact.

We have presented the extracted features used to predict stress levels of each employee. The next section presents how to learn a model to predict stress and how semi-supervised and transfer learning techniques are used to improve the prediction accuracy.

5. Stress modelling using transfer learning

In this section we present how to model stress using decision tress. Then, we use semi-supervised learning techniques to reduce the size of missing data. Finally, we propose to use transfer learning and ensemble methods to improve the accuracy of the learned models.

5.1. Modelling stress

Predicting perceived stress of a person can be modelled as a classification problem. We used decision trees [32] 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 Section 4.2 and the class to predict is the self-reported stress level (Section 4.1) (*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 Eq. (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 Fig. 3, 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

Table 4
Stress Prediction using decision tress before and after applying a Semi-supervised learning approach. Overall classes represent overall number of labelled 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

Bold values show the best score.

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 (grey boxes). The third cluster with only 4 subjects shows subjects with high and mid level of stress.

5.2. 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 study, we use self-training (ST) [26] 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 [56] with the default parameters of C4.5 classifier. The new classified data with high confidence ($\geqslant$80%) is added to the training set, the classifier is re-trained and the procedure repeated. Using ST we were able to reduce the unlabelled data (improving the labelled 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 Table 4).

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. 4(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 Fig. 4(b). Finally, in Fig. 4(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.

$$\Delta i,j(original, modified) = |e_{i,j}^{original} - e_{i,j}^{modified}| \quad (6)$$

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.

5.3. 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.

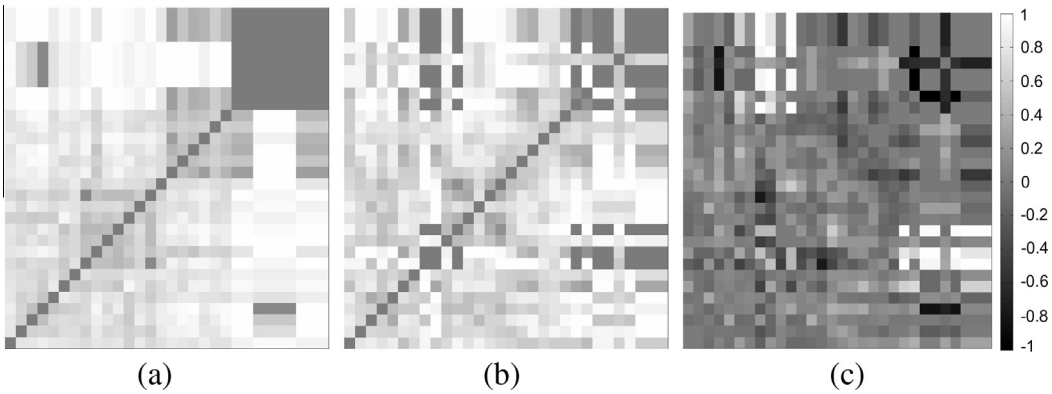

Fig. 4. Similarity matrices of 30 users using (a) all data (after semi-supervised learning) and (b) with 50% of instances removed –darker cells indicate high similarity. (c) depicts the difference between (a) and (b); a white cell indicates a + difference, black a − negative difference, and grey no difference.

Table 5
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. ± St. dev.	**62.09 ± 11.32**	60.61 ± 10.71	0.42 ± 0.12	−1.47 ± 8.42	75.91 ± 9.70

Bold values show the best accuracy score.

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 Eq. (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:

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

   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.

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

   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, farthest model and average for every subject in the study. From the results we obtained an average distance of 0.42 (using Eq. (3)) to the nearest subject, in contrast the, 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 5).

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 5 summarizes 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 Fig. 5(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, Fig. 5(c) and (d) shows trees which have a $d = 0.60$. Note, that in this case the trees show different decision nodes.

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 summarizes 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 $d \geqslant \beta$ it is better not to use transfer learning since the models are far from each other and this causes a negative transfer effect.

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

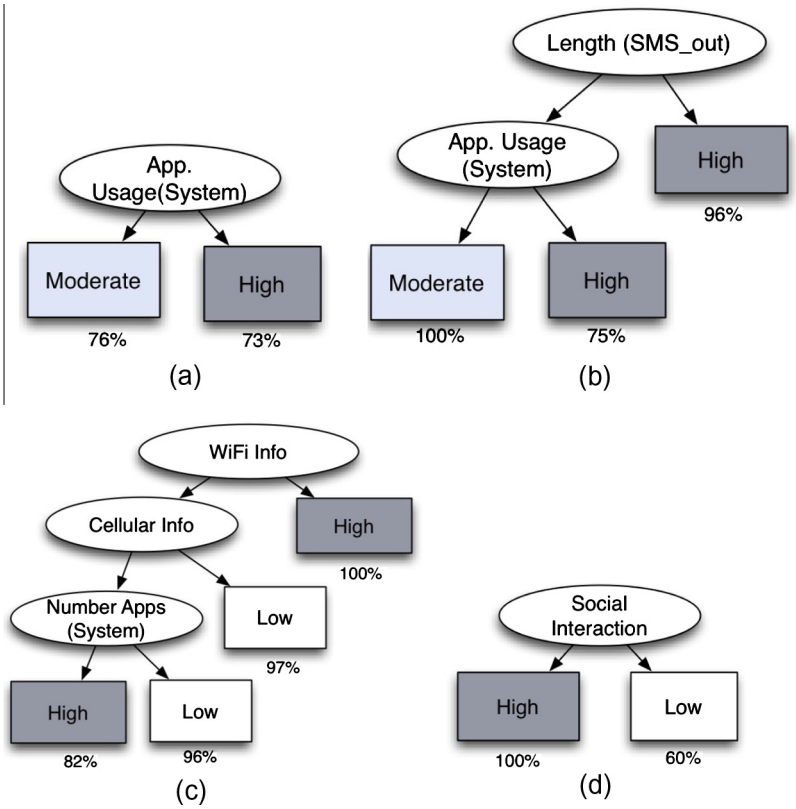

Fig. 5. Learned models of different subjects: *Subj*30 (a) and its most similar *Subj*17 (b). *Subj*29 (c) and its most similar model *Subj*05 (d).

Table 6
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. ± St. 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. ± St. 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

Bold values show the best accuracy score.

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).

Table 7
Classification accuracies using the proposed approaches and using all original data (upper bound).

Subject ID	No transfer	Transfer learning approaches				All data
		Naive	Threshold	Sampling weighted	Ensemble weighted	
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. ± St. 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

Bold values show the best accuracy score.

5.4. 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).

5.5. Summary of analysis

We proposed four different transfer learning approaches to cope with scarce data. Table 7 summarizes 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.

6. Challenges and limitations

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 deriving from smartphone sensors) and subjective data (self-reported questionnaires). 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 a large number of participants [57]. 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. In our future work, 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 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) [35]. Therefore, other approaches might be explored [33,34,58,59]. Finally, one last limitation is that the participants were recruited through two different organizations (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.

7. Conclusions and future work

In this paper we have used semi-supervised learning as a pre-processing technique to reduce the amount of unlabelled data. Then, we have analyzed 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 models to improve 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.

Conflict of interest statement

The authors manifest that they have work and/or study relations with:

- CREATE-NET, Italy
- Universita di Trento, Italy
- Instituto Nacional de Astrofisica, Optica y Electronica, Mexico
- Centrum Wiskunde & Informatica, The Netherlands

Acknowledgements

The work on this paper was partially funded by EC Marie Curie IRSES Project UBIHEALTH - 316337.

References

[1] K. Glanz, B.K. Rimer, K. Viswanath, Health Behavior and Health Education: Theory, Research, and Practice, John Wiley & Sons, 2008.
[2] K. Korabik, L.M. McDonald, H.M. Rosin, Stress, coping, and social support among women managers, in: B.C. Long, S.E. Kahn (Eds.), Women, Work, and Coping: A Multidisciplinary Approach to Workplace Stress, University of British Columbia Academic Women's Association, Montreal, 1993, pp. 133–153 (Chapter 7).
[3] C. Maslach, W.B. Schaufeli, M.P. Leiter, Job burnout, Annu. Rev. Psychol. 52 (1) (2001) 397–422.
[4] A. Parent-Thirion, P. Paoli, Working Conditions in the Acceding and Candidate Countries, European Foundation for the Improvement of Living and Working Conditions, Dublin, 2003.
[5] M. Milczarek, E. Schneider, E. Rial-González, Occupational Safety and Health in Figures: Stress at Work-Facts and Figures, European Agency for Safety and Health at Work, Luxemburg, 2009.
[6] P. Näätänen, A. Aro, S. Matthiesen, K. Salmela-Aro, Bergen Burnout Indicator 15, Edita, Helsinki, 2003.
[7] M. Al-Mardini, F. Aloul, A. Sagahyroon, L. Al-Husseini, Classifying obstructive sleep apnea using smartphones, J. Biomed. Infor. 52 (2014) 251–259.
[8] R. Guidoux, M. Duclos, G. Fleury, P. Lacomme, N. Lamaudière, P.-H. Manenq, L. Paris, L. Ren, S. Rousset, A smartphone-driven methodology for estimating physical activities and energy expenditure in free living conditions, J. Biomed. Inform. 52 (2014) 271–278.
[9] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[10] E. Ceja, V. Osmani, O. Mayora, Automatic stress detection in working environments from smartphones' accelerometer data: a first step, IEEE J. Biomed. Health Infor. 20 (4) (2015) 1053–1060, http://dx.doi.org/10.1109/JBHI.2015.2446195.
[11] J. Bakker, M. Pechenizkiy, N. Sidorova, What's your current stress level? detection of stress patterns from GSR sensor data, in: IEEE 11th International Conference on Data Mining Workshops,, Vancouver, BC, 2011, pp. 573–580.
[12] K.K.-L. Liu, A Personal, Mobile System for Understanding Stress and Interruptions Master's Thesis, MIT Media Arts and Science, 2004.
[13] H. Lu, D. Frauendorfer, M. Rabbi, M.S. Mast, G.T. Chittaranjan, A.T. Campbell, D. Gatica-Perez, T. Choudhury, Stresssense: Detecting stress in unconstrained acoustic environments using smartphones, in: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, USA, 2012, pp. 351–360.
[14] A. Muaremi, B. Arnrich, G. Tröster, Towards measuring stress with smartphones and wearable devices during workday and sleep, BioNanoScience 3 (2) (2013) 172–183, http://dx.doi.org/10.1007/s12668-013-0089-2.
[15] A. Grunerbl, A. Muaremi, V. Osmani, G. Bahle, S. Ohler, G. Troster, O. Mayora, C. Haring, P. Lukowicz, Smartphone-based recognition of states and state changes in bipolar disorder patients, IEEE J. Biomed. Health Infor. 19 (1) (2015) 140–148.
[16] V. Osmani, A. Maxhuni, A. Grünerbl, P. Lukowicz, C. Haring, O. Mayora, Monitoring activity of patients with bipolar disorder using smart phones, in: Proceedings of International Conference on Advances in Mobile Computing & Multimedia – MoMM '13, ACM Press, 2013, pp. 85–92, http://dx.doi.org/10.1145/2536853.2536882. <http://dl.acm.org/citation.cfm?id=2536853.2536882> .
[17] V. Osmani, Smartphones in mental health: detecting depressive and manic episodes, IEEE Pervasive Comput. 14 (3) (2015) 10–13.
[18] A. Maxhuni, A. Muñoz-Meléndez, V. Osmani, H. Perez, O. Mayora, E.F. Morales, Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive Mobile Comput. (in press). http://dx.doi.org/10.1016/j.pmcj.2016.01.008, URL <http://www.sciencedirect.com/science/article/pii/S1574119216000109>.
[19] R. LiKamWa, Y. Liu, N.D. Lane, L. Zhong, Moodscope: building a mood sensor from smartphone usage patterns, in: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services, Taipei, Taiwan, 2013, pp. 389–402.
[20] A. Sano, R.W. Picard, Stress recognition using wearable sensors and mobile phones, in: Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), Geneva, Switzerland, 2013, pp. 671–676.
[21] D. Carneiro, J.C. Castillo, P. Novais, A. Fernández-Caballero, J. Neves, Multimodal behavioral analysis for non-invasive stress detection, Expert Syst. Appl. 39 (18) (2012) 13376–13389.
[22] G. Bauer, P. Lukowicz, Can smartphones detect stress-related changes in the behaviour of individuals?, in: 2012 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), Lugano, Switzerland, 2012, pp 423–426.
[23] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, A.S. Pentland, Daily stress recognition from mobile phone data, weather conditions and individual traits, in: Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, Florida, USA, 2014, pp. 477–486.
[24] R. Ferdous, V. Osmani, J.B. Marquez, O. Mayora, Investigating correlation between verbal interactions and perceived stress, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), Milan, Italy, 2015, pp. 1612–1615, http://dx.doi.org/10.1109/EMBC.2015.7318683.
[25] V. Osmani, R. Ferdous, O. Mayora, Smartphone app usage as a predictor of perceived stress levels at workplace, in: 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), Istanbul, Turkey, 2015, pp. 225–228, http://dx.doi.org/10.4108/icst.pervasivehealth.2015. 260192.
[26] X. Zhu, Semi-supervised learning, in: Encyclopedia of Machine Learning, Springer, 2010, pp. 892–897.
[27] P. Rashidi, D.J. Cook, Multi home transfer learning for resident activity discovery and recognition, in: Proceedings of International Workshop on Knowledge Discovery from Sensor Data (SensorKDD), 2010, pp. 3–63.
[28] P. Hernandez-Leal, A. Maxhuni, L.E. Sucar, V. Osmani, E.F. Morales, O. Mayora, Stress modelling using transfer learning in presence of scarce data, in: Ambient Intelligence for Health, Springer, Puerto Varas, Chile, 2015, pp. 224–236.
[29] T.M. Mitchell, Machine Learning, 1st edition., McGraw-Hill Higher Education, New York, 1997.
[30] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Francisco, 2014.
[31] V. Vapnik, S.E. Golowich, A. Smola, Support vector method for function approximation, regression estimation, and signal processing, in: Advances in Neural Information Processing Systems, Denver, Colorado, USA, 1997, pp. 281–287.
[32] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Francisco, 1993.
[33] W. Shannon, D. Banks, Combining classification trees using MLE, Stat. Med. 18 (6) (1999) 727–740.
[34] R. Miglio, Metodi di partizione ricorsiva nell'analisi discriminante Ph.D. Thesis, Dipartimento di Scienze Statistiche, Bologna, 1996.
[35] R. Miglio, G. Soffritti, The comparison between classification trees through proximity measures, Comput. Stat. Data Anal. 45 (3) (2004) 577–593.
[36] L. Rokach, Ensemble-based classifiers, Artif. Intell. Rev. 33 (1-2) (2010) 1–39.
[37] T.G. Dietterich, Ensemble methods in machine learning, in: International Workshop on Multiple Classifier Systems, Cagliari, Italy, 2000, pp. 1–15.
[38] K. Nigam, A.K. McCallum, S. Thrun, T. Mitchell, Text classification from labeled and unlabeled documents using EM, Mach. Learn. 39 (2) (2000) 103–134.
[39] R. Luis, L.E. Sucar, E.F. Morales, Inductive transfer for learning bayesian networks, Mach. Learn. 79 (1) (2010) 227–255.
[40] S.D. Roy, T. Mei, W. Zeng, S. Li, Socialtransfer: cross-domain transfer learning from social streams for media applications, in: Proceedings of the 20th ACM International Conference on Multimedia, Nara, Japan, 2012, pp. 649–658.
[41] R. Raina, A. Battle, H. Lee, B. Packer, A.Y. Ng, Self-taught learning: transfer learning from unlabeled data, in: Proceedings of the 24th International Conference on Machine Learning, Corvallis, Oregon, USA, 2007, pp. 759–766.
[42] S.J. Pan, V.W. Zheng, Q. Yang, D.H. Hu, Transfer learning for wifi-based indoor localization, in: AAAI Workshop on Trading Agent Design and Analysis, Chicago, IL, USA, 2008, pp. 43–48.
[43] K. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, J. Big Data 3 (1) (2016) 1–40.
[44] D. McNair, M. Lorr, L. Droppleman, Manual for the profile of mood states, Educational and Industrial Testing Services, San Diego, CA, 1971.
[45] N. Aharony, W. Pan, C. Ip, I. Khayal, A. Pentland, Social fMRI: investigating and shaping social mechanisms in the real world, Pervasive Mobile Comput. 7 (6) (2011) 643–659.

[46] D. Birant, A. Kut, ST-DBSCAN: an algorithm for clustering spatial–temporal data, Data Knowl. Eng. 60 (1) (2007) 208–221.
[47] C. Robusto, The cosine-haversine formula, Am. Math. Monthly 64 (1) (1957) 38–40.
[48] P. Hedelin, D. Huber, Pitch period determination of aperiodic speech signals, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP-90), Albuquerque, NM, USA, 1990, pp. 361–364.
[49] F.J. Harris, On the use of windows for harmonic analysis with the discrete fourier transform, Proc. IEEE 66 (1) (1978) 51–83.
[50] C. Bernaards, M. Jans, S. Van den Heuvel, I. Hendriksen, I. Houtman, P. Bongers, Can strenuous leisure time physical activity prevent psychological complaints in a working population?, Occup Environ. Med. 63 (1) (2006) 10–16.
[51] C.D. Spielberger, P.R. Vagg, C.F. Wasala, Occupational stress: job pressures and lack of support, in: J.C. Quick, L.E. Tetrick (Eds.), Handbook of Occupational Health Psychology, American Psychological Association, Washington, DC, 2003, pp. 185–200.
[52] R.S. Lutz, M.A. Stults-Kolehmainen, J.B. Bartholomew, Exercise caution when stressed: stages of change and the stress–exercise participation relationship, Psychol. Sport Exerc. 11 (6) (2010) 560–567.
[53] T.L. Conway, R.R. Vickers Jr, H.W. Ward, R.H. Rahe, Occupational stress and variation in cigarette, coffee, and alcohol consumption, J. Health Soc. Behav. 22 (2) (1981) 155–165.
[54] Stress Effects – American Institute of Stress, 2016. <http://www.stress.org/stress-effects/> (accessed: 07-07-16).
[55] S. Cohen, T.A. Wills, Stress, social support, and the buffering hypothesis, Psychol. Bull. 98 (2) (1985) 310–357.
[56] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explor. Newslett. 11 (1) (2009) 10–18.
[57] S. Xiang, L. Yuan, W. Fan, Y. Wang, P.M. Thompson, J. Ye, 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, 2013, pp. 185–193.
[58] E.B. Fowlkes, C.L. Mallows, A method for comparing two hierarchical clusterings, J. Am. Stat. Assoc. 78 (383) (1983) 553–569.
[59] H.A. Chipman, E.I. George, R.E. McCulloch, Managing multiple models, in: Eighth International Workshop on Artificial Intelligence and Statistics, Key West, Florida, USA, 2001, pp. 11–18.