

# Unsupervised Stress Detection Algorithm and Experiments with Real Life Data

Elena Vildjiounaite<sup>1</sup>(✉), Johanna Kallio<sup>1</sup>, Jani Mäntyjärvi<sup>1</sup>,  
Vesa Kyllönen<sup>1</sup>, Mikko Lindholm<sup>1</sup>, and Georgy Gimelfarb<sup>2</sup>

<sup>1</sup> VTT Technical Research Centre of Finland, 02044 VTT Espoo, Finland  
{elena.vildjiounaite, johanna.kallio, jani.mantyljarvi,  
vesa.kyllonen, mikko.lindholm}@vtt.fi

<sup>2</sup> University of Auckland, Private Bag 92019, Auckland 1149, New Zealand  
g.gimelfarb@auckland.ac.nz

**Abstract.** Stress is the major problem in the modern society and a reason for at least half of lost working days in European enterprises, but existing stress detectors are not sufficiently convenient for everyday use. One reason is that stress perception and stress manifestation vary a lot between individuals; hence, “one-fits-all-persons” stress detectors usually achieve notably lower accuracies than person-specific methods. The majority of existing approaches to person-specific stress recognition, however, employ fully supervised training, requiring to collect fairly large sets of labelled data from each end user. These sets should contain examples of stresses and normal conditions, and such data collection effort may be tiring for end users. Therefore this work proposes an algorithm to train person-specific stress detectors using only unlabelled data, not necessarily containing examples of stresses. The proposed method, based on Hidden Markov Models with maximum posterior marginal decision rule, was tested using real life data of 28 persons and achieved average stress detection accuracy of 75%, which is similar to the accuracies of state-of-the-art supervised algorithms for real life data.

**Keywords:** Stress detection · Unsupervised learning · Hidden Markov Models

## 1 Introduction

Stress is a state of mental tension and worry caused by problems in life, work, etc.; something that causes strong feelings of worry or anxiety [1]. As stress is one of the major problems in modern society [2], intelligent environments should be able to recognise human stresses. According to a recent review [3], however, the majority of studies into stress detection by environmental sensors were performed in a lab and hence are not yet ready for real life conditions, that is, when subjects do not face video, depth or hyperspectral cameras or do not sit in pressure-sensitive chairs. Studies with environmental audio sensors [4] and keyboard/mouse dynamics [5, 6] use real life data, but these modalities do not work when the subjects stay silent or do not use computers.

Wearable sensors allow to monitor humans in greater range of situations than environmental sensors, tested to date, but again, not all wearable sensors suit to real life

use. Most of physiological monitoring devices provoke too much discomfort [3] for everyday use, and their improper attachment may cause notable data losses [7]. To date, only wrist bracelets and mobile phones were found sufficiently convenient in real life use [8, 9], but system convenience depends also on algorithm choice. Unfortunately, existing stress detectors typically employ fully supervised algorithms, such as SVM (Support Vector Machines) [2, 4, 9–11] and decision trees [12, 13]. Training of these methods requires fairly large sets of labelled data, and stress labels (i.e., information about stress times and (possibly) severity levels) are usually obtained by asking test subjects to provide self-reports upon periodical system prompts. The reason why it is beneficial to obtain stress labels from each end user is strong person-dependency of stress perception and stress manifestation (stress influence on physiological parameters and behaviour). Hence stress detection models, which training utilised labelled data of the target individual, usually achieve significantly higher accuracies than any other models. Such results were reported by both lab tests [11, 14] and field studies [4, 10, 12], and in the latter training person-specific models on the data of each subject increased stress recognition accuracy by 20% on average compare with general models. This accuracy gain, however, was achieved at the cost of notable data labelling efforts: training data in [10, 12] contained about 100 labelled instances per person; dataset in [4] - even more. Although dataset size of 100 samples is not large for machine learning methods, it seems to be too large for humans: in [8] not every subject provided even 100 labels, and on average the test subjects answered only 28% of system prompts to provide self-reports.

Therefore several alternatives to fully supervised training were proposed to date. One of them is to train a general model using data of many subjects and to adapt it to each target person in a certain way, e.g., by using person-specific input features (such as deviations of sensor values from the average values of the target individual in neutral state [11]) or by modifying class priors according to individual tendency to report more or less stressful events [4]. Another alternative is to exploit similarity between human beings, e.g., by clustering test subjects and training a separate model for each cluster. Then the target individual is assigned to one of these clusters using either his/her labelled [12] or unlabelled [4, 11, 14] data, and the corresponding model is used for detecting his/her stresses. Other proposed methods include combining outputs of models of similar subjects and training a model using a mixture of the target person data with data of similar individuals [13]. Success of similarity-based methods, however, was found to depend on the chosen numbers of similar persons. Another similarity-based reasoning method is *k* Nearest Neighbours classifier, but this approach was tested only in a limited range of real-life settings: users working on their computers [6]. In addition, although the above-listed approaches reduce the need in labelled data of each target individual, they nevertheless require collecting labelled data of many persons. To the best of our knowledge, only the work [15] proposed an unsupervised stress detection method: to recognise stresses by calculating so-called “additional heart rate” (a deviation between current and recent physiological data) and comparing it with dynamically calculated threshold, taking into account current physical activity. This method required two additional sensors, however: accelerometers on the chest and the right thigh, which is not a realistic setup for long-term real life application.

This work proposes stress detection system for a fairly broad range of real life settings, convenient both sensor-wise and algorithm-wise: it uses mobile phones and wrist bracelets and requires no data labelling. As unsupervised learning typically results in lower accuracies than supervised one, this work aims at recognising only the most dangerous stresses: high level stresses. The main contribution of this work is the following: first, we propose a novel unsupervised method to recognise stress-related data patterns. Then, using real life data, we demonstrate that for training of this method fairly small datasets (two days) suffice and that these datasets should not necessarily contain examples of stresses. In addition, we discuss how the stress recognition results correlate with personality traits of the test subjects.

## 2 Unsupervised Stress Detection Algorithm

We assume that high stresses occur in human lives less frequently than normal conditions; hence, we detect high stresses by learning a model of normal human condition and evaluating current deviations from this model. As normal behavioural and physiological data may notably vary between different individuals, we learn normal models for each person separately. Although stress may be caused by a short event and may reflect itself in physiological data as a short-term deviation, realising what happened and coping with stress takes some time. Furthermore, due to great varieties of short-term human behaviours, it is difficult to reliably distinguish between normal and unusual data patterns in a short term. Thus we detect stresses by evaluating data sequences within time windows of certain duration, and we employ overlapping time windows instead of consecutive ones to obtain stress detection results as frequently as desirable.

We suggest to classify each time window into two classes, “normal” vs. “unusual”, at least when training datasets are small. After the system collects sufficient amount of training data, it may switch to finer classification. For unsupervised classification we propose to employ HMM (Hidden Markov Model) with two hidden states (“normal” and “unusual”) and discrete observations because training of such models does not require large datasets. We propose to train HMM in fully unsupervised way, namely:

- use no pre-defined thresholds for classifying data samples;
- use no labels in training, i.e., use all collected data samples in exactly same way.

### 2.1 Unsupervised Hidden Markov Model Training

First, we use all training data to create a reference model, i.e., a vector of reference samples of physiological, acceleration and mobile phone usage data features. As we don’t use any labels in training, we don’t know which data samples denote normal condition and which ones denote stress; hence, we calculate feature-wise average over all samples according to formula (1), where  $V_{M:i}$  is a reference sample of feature  $i$ ,  $V_{T:i}$  is a data sample of this feature at time  $T$ , and  $m$  is total number of training data samples.

$$V_{M:i} = \frac{1}{m} \sum_{T=1}^m V_{T:i} \tag{1}$$

Then for each data sample we calculate its deviation from the reference model. A deviation  $D$  of the time moment  $T$  with  $n$  features from the reference model  $M$  is:

$$D = \frac{1}{n} \sum_{i=1}^n (V_{T:i} - V_{M:i}) \tag{2}$$

This deviation could be straightforwardly used as degree of normality of the current time moment, but in our tests it did not work well. Instead, we discretise this deviation by dividing an interval  $[-1, 1]$  into  $K$  appropriate sub-intervals and use the sub-interval's number as an observation in the HMM. The experiments below employed the following  $K = 6$  sub-intervals:  $[-1.0, -0.6]$ ;  $[-0.6, -0.3]$ ;  $[-0.3, 0]$ ;  $[0, 0.3]$ ;  $[0.3, 0.6]$ ;  $[0.6, 1.0]$ . A sequence of the discretised deviations of data samples inside selected time window from the reference model is a sequence of the HMM observations for this time window.

The HMM of a time window has finite sets of states  $S = \{1, \dots, N\}$  (output classes, i.e., “normal” and “unusual”) and observations  $X = \{1, \dots, K\}$ . An example configuration of the proposed HMM is illustrated in Fig. 1.

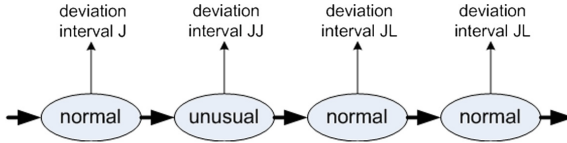


Fig. 1. HMM model of a time window

Let  $S_T = \{s_t : t = 1, \dots, T\}$  be a sequence of the hidden states and  $X_T = \{x_t : t = 1, \dots, T\}$  be a corresponding sequence of obtained observations at  $T$  time moments. The HMM assumes that every observation  $x_t$  at time  $t$  depends, in a probabilistic sense, only on a hidden state  $s_t$  and the latter (excluding  $s_1$ ) depends in turn only on the previous state  $s_{t-1}$ . Let  $\alpha = [\alpha(s|v) : s, v \in S]$ ;  $\beta = [\beta(x|s) : x \in X, s \in S]$ , and  $\pi = [\pi(s) : s \in S]$  denote conditional probabilities of transitions between the discrete states; conditional probabilities of observations, given the state, and unconditional probabilities of the initial state, respectively. Then the HMM is characterised by the joint probability of the sequences of evolving states and observations [17]:

$$p(S_T, X_T) = \pi(s_1) \beta(s_1|x_1) \prod_{t=2}^T \alpha(s_t|s_{t-1}) \beta(x_t|s_t) \tag{3}$$

Given a sequence of unlabelled observations  $X_T$ , HMM is trained, i.e., its parameters  $(\alpha, \beta, \pi)$  are learned in a fully unsupervised mode by applying Baum-Welch forward-backward algorithm [16]. In other words, mapping of deviations from a normal model into output classes is learned from the training data and does not require

defining any thresholds. This method can be used for training HMM on data of each target individual separately (so-called “*personal model*”) as well as on data of all individuals (“*general model*”) or of similar individuals (“*similarity-based model*”). Reference models can be built using data of each individual in all cases because stress labels are not required.

## 2.2 Inference with Hidden Markov Models

To classify the current time window as “normal” or “unusual”, a sequence of observations within this window (i.e., a sequence of discretised deviations of data samples from the reference model) is created first. Then a sequence of the hidden states can be obtained by the Bayesian maximum *a posteriori* (MAP) decision rule using the well-known Viterbi dynamic programming algorithm [16]. The MAP rule minimising the error probability assumes that the cost of errors for a given sequence of observations is just the same, irrespectively of their number (i.e., a single erroneous state or all  $T$  such errors are equally bad). An alternative way is to account for all the individual errors and minimise their expected number. In this case the hidden states are to be recovered with the Bayesian MPM (maximum posterior marginal) rule that selects for each time moment the hidden state with the maximum posterior marginal probability. The posterior marginals  $\{p_t(s|X_T) : s \in \mathbf{S}; t = 1, \dots, T\}$  are calculated for each hidden state  $s$  of time  $t$  by the forward and backward message propagation [16]:

$$p_t(s|X_T) = \frac{\mu_t(s|x_1, \dots, x_t)m_t(s|x_T, \dots, x_t)}{\sum_{v \in \mathbf{S}} \mu_t(v|x_1, \dots, x_t)m_t(v|x_T, \dots, x_t)} \quad (4)$$

where  $\mu_t(s|x_1, \dots, x_t)$  and  $m_t(s|x_T, \dots, x_t)$  denote the forward and backward message, respectively, for the state  $s$  at each instant  $t$ . These messages are computed successively from the beginning and the end of the observed sequence  $X_T : \mu_1(s, x_1) = \pi(s)\beta(x_1|s)$ ;

$$\begin{aligned} \mu(s; x_1, \dots, x_t) &= \sum_{v=1}^N \mu_{t-1}(v; x_1, \dots, x_{t-1}) \alpha(s|v) \beta(x_t|s) \\ \text{and } m_T(s|x_T) &= 1; m_t(s|x_T, \dots, x_t) = \sum_{v=1}^N m_{t+1}(v|x_T, \dots, x_{t+1}) \alpha(v|s) \beta(x_{t+1}|s) \end{aligned} \quad (5)$$

Both the conventional MAP and less conventional MPM decisions produce a recovered sequence of hidden states classifying each time moment, for example, “normal, unusual, normal, normal, ...”. Due to the HMM based reasoning, the classification of each time moment depends on all the observations and hidden states.

A stress score of the whole window can be then calculated as a conventional likelihood of generating a given sequence of observations by the learned HMM (the lower the likelihood, the less normal the sequence), but this estimation takes into account the order of recovered states. In our experience, learning to recognise truly unusual order of states requires fairly large training datasets. As we had fairly small datasets per subject in this study, we calculated a stress score of each window in a different way, by assigning numerical scores  $A_s$  to the recovered hidden states:

$$A_s = \begin{cases} 1, & \text{if state } s \text{ is classified as normal} \\ -1, & \text{if state } s \text{ is classified as unusual} \end{cases} \quad (6)$$

and using an average quantified state as a window stress score:

$$A = \frac{1}{T} \sum_{s=1}^T A_s \quad (7)$$

For example, the stress score of a sequence “normal, unusual, normal, normal” is  $A = \frac{1-1+1+1}{4} = 0.75$ , indicating a normal time window.

In the proposed method all HMM model parameters are learned from the data, just time window length has to be specified. Choice of window length should depend on human behaviour patterns and number of data samples in a window. Short time windows are likely to result in high false stress detection rate because of diversity of short-term human behaviours. HMM does not work well with short sequences of observations either. The proposed HMM inference may additionally benefit from longer sequences: due to crisp digitalisation of deviations between data samples and reference models, HMM observations are sensitive to small changes in sensor values lying on the boundaries between  $K$  sub-intervals. This problem is softened by the probabilistic nature of HMM inference and dependency of each hidden state classification on the neighbouring states (HMM may assign exactly same observation to “normal” or “unusual” state depending on its estimations of other states), but this “smoothing” effect is more notable in longer time windows. On the other hand, long time windows are more likely to include a mixture of “stressed” and “normal” (e.g., “not yet stressed”) human conditions.

For time window classification it is needed to specify also decision threshold  $TS$ : if window score  $A$  falls below  $TS$ , it denotes stress; otherwise human condition is normal. For classifying each day into “stress” and “normal” classes we also need to specify, how many windows in a day should be classified as “stress” to classify the whole day as stressful. The experiments below compared several lengths of time windows and decision thresholds for the following two inference schemes:

- (a) *HMM-Viterbi*: sequence of hidden states is obtained via the MAP decision rule by the Viterbi algorithm and the window score is calculated using Eqs. (6) and (7).
- (b) *HMM-MPM*: Each sequence of hidden states is obtained via the MPM decision rule and the window score is formed in accord with Eqs. (6) and (7).

### 3 Experiments with Field Data

#### 3.1 Data Collection

In this study we used data, collected by the Institute of Behavioural Sciences at the University of Helsinki (Finland). Participants were recruited from the university; they had to satisfy the following criteria: good health, non-smoking, interest in technology, willingness to use mobile applications and possession of an Android smart phone. The

subjects were monitored during normal course of their lives during four days, although monitoring of a few subjects lasted for three or five days. Before data collection the subjects answered a questionnaire to identify their Big Five personality traits.

Physiological data were collected once per minute by wrist-worn Basis device [17] that included an optical heart rate sensor and 3-axis accelerometer. Mobile phone data included two parts: (1) an Android service collecting digital behaviour data and (2) self-reporting. Digital behaviour data consisted of logs of usage of six different application types: social (Skype, social networks etc.), entertainment (games, music etc.), infotainment (news, books etc.), business (calendar, editing etc.), wellbeing (weight watching, exercise monitoring etc.) and any other interaction with a phone. These logs only contained information whether a user interacted with an application of certain type or not during each minute; contents of the web pages or keystrokes were not logged.

Self-reporting was prompted by Android notification every 45 min during daytime from 9 am to 9 pm. The subjects had to answer whether stress had occurred during the current reporting period and if so, evaluate it on 7-level Likert scale. The subjects also provided free-form comments on their activities. 28 subjects answered questionnaires fairly regularly. These persons aged from 20 to 47 years old (mean 25.5 years, standard deviation 6); among them were 4 males and 24 females. High stress levels (6–7 on Likert scale) were reported on 12% of days, by the female subjects only.

### 3.2 Experimental Protocol

In the tests we first evaluated accuracy of classifying each self-reporting period as “stress” (i.e., high stress) vs. “normal” (other conditions). Although Android phone prompted the test subjects to provide self-reports every 45 min, in practice intervals between reports were not so regular because the subjects did not always answer immediately. Thus it was not possible to use HMM time window of 45 min in the evaluation and to compare HMM results with the self-reports directly. Instead, we compared several HMM window sizes in the following protocol: first, HMM labelled each time window as either “stress” or “normal”. These HMM results were stored along with the window timestamps. Then for each time interval between self-reports we retrieved HMM results for time windows falling inside this interval.



**Fig. 2.** Evaluation of time intervals between self-reports

If a self-report stated that high stress has occurred and at least one HMM result within this interval was “stress”, we considered this interval as real stress, otherwise - as missed stress. Similarly, if self-report stated that high stress has not occurred, we

considered this interval as “true normal” if HMM had not labelled any window within this interval as “stress”; otherwise we considered this interval as false stress. For example, in Fig. 2 time interval A includes HMM results 1 and 2. If one of these results is “stress”, it means that HMM evaluated interval A as “stress”, and this result is compared with the self-report 1. Then we calculated the following criteria:

- stress detection rate, i.e., a ratio between correctly classified “high stress” intervals and the total number of “high stress” intervals;
- non-stress detection rate, i.e., a ratio between correctly classified intervals, not labelled as “high stress”, and the total number of these intervals.

These criteria were used to compare the following stress modelling approaches:

- **Personal**: model of each target subject was trained using data of this person only; we used first half of his/her data for training and another half for testing, then swapped training and test sets and averaged the results over these runs;
- **Similarity-based**: first the subjects were clustered via k-means clustering; then for each target subject a model was trained using all data of other subjects in the same cluster and tested on all data of the target subject;
- **General**: for each target person, a model was trained on all data of all other subjects and tested on all his/her data (i.e., leave-one-person-out protocol).

In addition, as timing of self-reports may be imprecise, we evaluated ability of *personal* models to evaluate days, i.e., to detect whether stress occurred during a day or not. For this we needed one more hyper-parameter, TW: if number of HMM results for some day exceeded TW, this day was classified as “stress”; otherwise as “normal”.

In all cases neither training of HMM-based stress classification models nor clustering utilised self-reports; hence all above-listed approaches were fully unsupervised.

### 3.3 Experimental Results

Figure 3 presents accuracies of recognising “stress” vs. “normal” human conditions by personal models for different window sizes and decision thresholds TS.

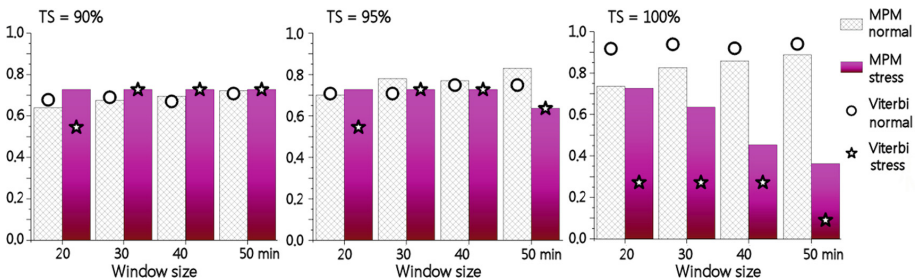
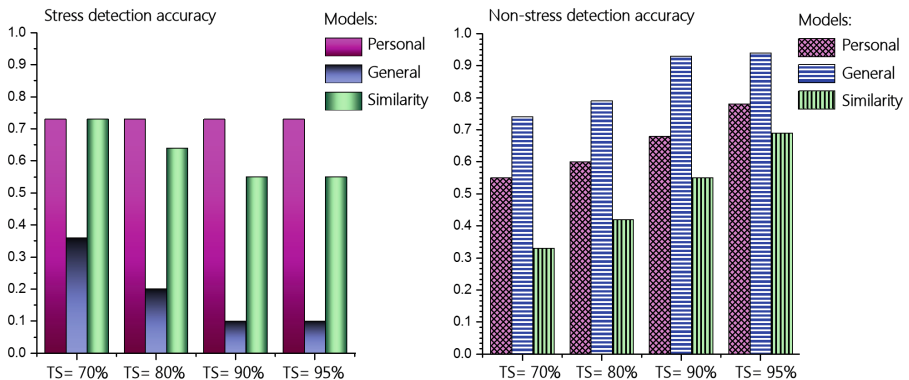


Fig. 3. Classification of time intervals between self-reports by personal HMM models



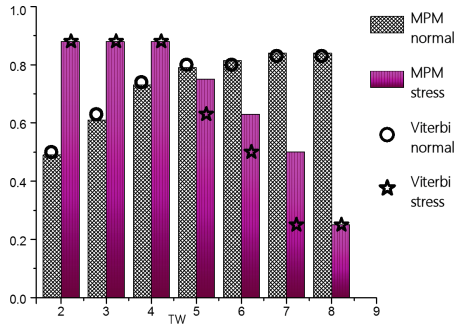
A window is classified as “stress” if number of its hidden states, classified as “unusual”, is greater than TS. For example, window size 50 min and TS = 95% mean that at least 48 min in a window should be classified as “unusual” to classify this window as “stress”. As Fig. 3 shows, HMM-MPM behaviour is consistent with respect to hyper-parameter changes: increase in window size and TS leads to increase in true negative rate and decrease in high stress recognition rate. HMM-Viterbi behaves less consistently: its recognition rate of high stress for window size 20 min and TS = 90–95% is lower than that for window sizes 30 and 40 min for the same values of TS. HMM-MPM also achieved higher accuracy for TS = 100% than HMM-Viterbi.

Figure 4 compares accuracies of personal, general and similarity-based models. For the latter we used the following numbers of clusters: two, three, four, five and six. Best result was achieved in case of four clusters; with other numbers of clusters accuracy of recognising stress was slightly lower, whereas accuracies of recognising normal conditions were fairly similar to that with four clusters. Figure 4 shows that general models tend to classify nearly all intervals as normal. Similarity-based modelling resulted in higher false detection rate and lower stress detection rate than that of personal models.



**Fig. 4.** Accuracies of personal, general and similarity-based HMM-MPM models for window size 30 min and different TS values; similarity-based results are presented for 4 clusters

Figure 4 shows that personal models achieved higher accuracies than other models: 73% accuracy of recognising stress and 75–80% accuracy of recognising normal conditions. Unfortunately, in real life these accuracies would mean large number of false detections if number of normal time intervals is large. On the other hand, in real life it may be sufficient to evaluate stresses on daily basis: to detect whether stress occurred today. Figure 5 presents accuracies of evaluating days by personal models for different values of threshold TW: the best results were achieved for TW equal to 4 or 5, i.e., for our test subjects it was fairly normal to have 4–5 periods of low activity per day, but not more. Figure 5 shows that HMM-MPM again outperformed HMM-Viterbi.



**Fig. 5.** Accuracies of detecting stresses by personal HMM models on daily basis for window size 30 min and TS = 95% for different values of threshold TW.

## 4 Discussion

During data collection the test subjects were asked to provide free-form comments regarding their activities in addition to stress labels. Although not all self-reports included free-form comments, from available comments we deduced that the proposed method best of all recognised high stresses which caused subjects' distraction from the current activity and contemplation: e.g., one subject stated that stress caused her to sit and wonder about received message. We also observed that HMM did not recognise stresses when the subjects dealt with them actively, e.g., one stress was not detected because the subject had a long phone call with her husband; another one - because the person went shopping. After checking trained HMM models we found that indeed HMM classified time windows as "unusual" if test subjects were notably inactive during this time: used their phones and moved less than normally. This finding corresponds to findings of the work [2], which reported that high stress significantly correlated with shorter "phone screen on" times, and to observations that stresses decrease physical activity level [13]. Hence HMM detected stress-induced decrease in values of physiological and behavioural features, unlike more common way to treat activity-induced changes in physiological data as non-indicative of stresses [15, 18].

As the subjects' "Big Five" personality data were acquired, we calculated Pearson correlation for our results and the subjects' factors. We found two statistically significant correlations: accuracy of recognising high stress negatively correlated with conscientiousness ( $R = -0.91$ ,  $p = 0.01$ ) and positively correlated with openness ( $R = 0.82$ ,  $p = 0.04$ ). Higher conscientiousness value means that the person is more efficient/organised and less easy-going/careless; higher openness value means that the person is more inventive/curious and less consistent/cautious. Hence the observed correlations suggest that organised and consistent persons were more inclined to keep doing what they were doing when stress occurred, whereas careless and inventive subjects were more inclined to stop and contemplate stresses, which sounds reasonable.

Classification of periods of inactivity as stresses, unfortunately, resulted in errors in cases when test subjects were notably inactive because they focused on studying or slept during daytime. For such cases it appeared important to have accelerometer in a

wrist bracelet: normally human beings move their hands fairly often even if other parts of the body are still. In addition, classifying daytime sleeping as unusual behaviour may be correct if it happens only occasionally, because then daytime sleeping is a sign of tiredness and hence is related to stress. Learning whether daytime sleeping is usual or unusual behaviour for each person is fairly straightforward, but it requires longer observation histories than that available for this study. Number of misclassifications of cases when the subjects are highly concentrated on studying can be decreased by analysing user activities in greater detail, e.g., by classifying computer applications. An alternative approach is to learn models of normal behaviour depending on time, location and other available context data: for example, to learn what is normal for each user in the afternoon at home vs. what is normal for the same user in the morning in other places. As the proposed approach does not require any self-reports from the end users, it allows to learn large numbers of context-dependent models effortlessly for the users. Hence in future we plan to evaluate accuracy of context-dependent stress detection.

To the best of our knowledge, this work is the first study into unsupervised learning of person-specific stress detection models on the basis of wrist sensors and mobile phone data. Hence we only aimed at recognising most dangerous stresses (high stresses), moreover that training data, available for this study, were not abundant. Nevertheless the proposed unsupervised HMM-MPM stress detection algorithm achieved accuracies, comparable with the state-of-the-art results of fully supervised methods. For example, two works that trained general models to recognise stresses on the basis of data from wrist sensors (either alone [9] or in combination with mobile phone usage data [2]), achieved 75% accuracies for two-class context-independent stress detection problem in the tests with real life data. Context-dependent stress recognition in [9] was notably more accurate than context-independent one; hence, we also plan to train context-dependent models in future, but unlike [9], in our case learning will not require data labelling. Person-specific models, trained to recognise stresses on the basis of mobile phone usage data, achieved 70–71% accuracy in three-class classification problem in [10, 13], but such training requires significant labelling efforts from end users. We also plan to study capability of unsupervised HMM to recognise several stress levels in future, after collecting larger datasets.

Fully supervised training not only requires fairly large sets of labelled data; it requires these sets to contain examples of all classes. High stresses typically occur less often than normal conditions; furthermore, highly stressed users may feel too badly to report all such cases. Therefore collecting suitable databases for training fully supervised methods may require long time. The proposed unsupervised HMM-based stress detection approach does not suffer from this drawback: in our tests high stresses of several users occurred during one day only and hence recognition of these stresses was performed by HMM models, trained only on the normal data. Nevertheless these stresses were recognised correctly when the users reacted by decrease in activities.

## 5 Conclusion

This work proposed unsupervised stress detection algorithm, based on discrete HMM with MPM decision rule. The proposed algorithm first learns person-specific models of normal behaviour and then classifies notable deviations from normal behaviour as stresses. In the experiments with real life data of 28 subjects the proposed HMM-MPM approach outperformed more conventional HMM-Viterbi inference method, based on the MAP decision rule.

The main advantage of the employed devices (a wrist bracelet and a mobile phone) is ease of wearing in a broad range of real life settings. The main advantage of the proposed reasoning algorithm is its unobtrusiveness: algorithm training requires no data labelling from the end users. Hence the proposed method allows to provide for person-dependency of stress perception and stress influence on physiological parameters and behaviour of human beings in more realistic way than existing approaches, either employing supervised training (and thus requiring end users to collect fairly large sets of labelled data), or employing data collection devices that suit only to a limited range of conditions (e.g., only to cases of computer work [6], or only to cases of constant monitoring of sensor attachment quality [7]). The main limitation of the proposed approach is that it best of all recognises stresses which cause subjects' contemplation, but may easily fail to detect cases of active coping with stresses (e.g., when the subjects discuss the problems with trusted persons or comfort themselves by shopping). On the other hand, some studies reported that active coping with stresses, such as positive orientation, seeking advice and assistance etc., is correlated with overall wellness [19] and hence is less dangerous than rumination. The main limitation of this study is a fairly small dataset and gender bias: high stresses were reported by females only. The latter may be not so important, however, as emotion-focused coping with stresses, such as rumination, was found to be gender-independent [20].

In future we plan to increase accuracy of HMM-MPM by learning context-dependent person-specific models, i.e., to learn what kind of behaviour is normal in different contexts, and then to evaluate deviations from the normal model in context-dependent way. We also plan to exploit relations between Big Five personality factors and stress detection results in more detail. The initial experiments, presented in this work, seem to be encouraging for further algorithm development and evaluations because in this study the proposed unobtrusive system achieved accuracy, similar to that of the state-of-the-art fully supervised methods, reported by other studies.

## References

1. Merriam-Webster dictionary. <http://www.learnersdictionary.com/definition/stress>. Accessed 23 Mar 2017
2. Sano, A., Picard, R.W.: Stress recognition using wearable sensors and mobile phones. In: ACII 2013, pp. 671–676 (2013)

3. Alberdi, A., Aztiria, A., Basarab, A.: Towards an automatic early stress recognition system for office environments based on multimodal measurements: a review. *J. Biomed. Inform.* **59**, 49–75 (2016)
4. Hernandez, J., Morris, R.R., Picard, R.W.: Call center stress recognition with person-specific models. In: *ACII 2011*, pp. 125–134 (2011)
5. Carneiro, D., Novais, P., Pêgo, J.M., Sousa, N., Neves, J.: Using mouse dynamics to assess stress during online exams. In: Onieva, E., Santos, I., Osaba, E., Quintián, H., Corchado, E. (eds.) *HAI 2015. LNCS*, vol. 9121, pp. 345–356. Springer, Cham (2015). doi:[10.1007/978-3-319-19644-2\\_29](https://doi.org/10.1007/978-3-319-19644-2_29)
6. Pimenta, A., Carneiro, D., Novais, P., Neves, J.: Detection of distraction and fatigue in groups through the analysis of interaction patterns with computers. In: Camacho, D., Braubach, L., Venticinque, S., Badica, C. (eds.) *Intelligent Distributed Computing VIII. SCI*, vol. 570, pp. 29–39. Springer, Cham (2015). doi:[10.1007/978-3-319-10422-5\\_5](https://doi.org/10.1007/978-3-319-10422-5_5)
7. Rahman, M.M., et al.: Are we there yet?: feasibility of continuous stress assessment via wireless physiological sensors. In: *ACM-BCB 2014*, pp. 479–488 (2014)
8. Adams, P., Rabbi, M., Rahman, T., Matthews, M., Voids, A., Gay, G., Choudhury, T., Voids, S.: Towards personal stress informatics: comparing minimally invasive techniques for measuring daily stress in the wild. In: *PervasiveHealth 2014*, pp. 72–79 (2014)
9. Gjoreski, M., Gjoreski, H., Lutrek, M., Gams, M.: Continuous stress detection using a wrist device: in laboratory and real life. In: *Ubicomp 2016 Adjunct*, pp. 1185–1193 (2016)
10. Ferdous, R., Osmani, V., Mayora, O.: Smartphone app usage as a predictor of perceived stress levels at workplace. In: *PervasiveHealth 2015*, pp. 225–228 (2015)
11. Shi, Y., et al.: Personalized stress detection from physiological measurements. In: *International Symposium on Quality of Life Technology (2010)*
12. Garcia-Ceja, E., Osmani, V., Mayora, O.: Automatic stress detection in working environments from smartphones’ accelerometer data: a first step. *IEEE J. Biomed. Health Inform.* **20**(4), 1053–1060 (2016)
13. Maxhuni, A., Hernandez-Leal, P., Sucar, L.E., Osmani, V., Morales, E.F., Mayora, O.: Stress modelling and prediction in presence of scarce data. *J. Biomed. Inform.* **63**, 344–356 (2016)
14. Xu, Q., Nwe, T.L., Guan, C.: Cluster-based analysis for personalized stress evaluation using physiological signals. *IEEE J. Biomed. Health Inform.* **19**(1), 275–281 (2015)
15. Kusserow, M., Amft, O., Troster, G.: Modeling arousal phases in daily living using wearable sensors. *IEEE Trans. Affect. Comput.* **4**(1), 93–105 (2013)
16. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**(2), 257–286 (1986)
17. <https://www.mybasis.com/>. Accessed 12 Jan 2016
18. Hovsepian, K., al’Absi, M., Ertin, E., Kamarck, T., Nakajima, M., Kumar, S.: cStress: towards a gold standard for continuous stress assessment in the mobile environment. In: *ACM International Joint Conference on Pervasive and Ubiquitous Computing (2015)*
19. Aalto, J.K., Brotheridge, C.M.: Resources, coping strategies and emotional exhaustion: a conservation of resources perspective. *J. Vocat. Behav.* **63**(3), 490–509 (2003)
20. Folkman, S., Lazarus, R.S.: An analysis of coping in a middleaged community sample. *J. Health Soc. Behav.* **21**, 219–239 (1980)