Computers in Industry 90 (2017) 42–49

Contents lists available at ScienceDirect

Computers in Industry

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

# Detecting work-related stress with a wearable device

Lu Han[a,b], Qiang Zhang[a,b], Xianxiang Chen[a,b], Qingyuan Zhan[c], Ting Yang[c,*], Zhan Zhao[a,b,**]

[a] Institute of Electronics, Chinese Academy of Sciences, China
[b] University of Chinese Academy of Sciences, China
[c] China-Japan Friendship Hospital, China

ARTICLE INFO

*Article history:*
Received 2 December 2016
Received in revised form 22 May 2017
Accepted 25 May 2017
Available online xxx

*Keywords:*
Stress detection
Wearable computing
Electrocardiogram
Respiration
Support vector machine

ABSTRACT

Excessive stress will lower work efficiency, lead to negative emotions and even various illnesses. This paper aims at detecting work-related stress based on physiological signals measured by a wearable device. Different from common binary stress detection, this study detects three levels of stress, i.e., no stress, moderate stress and high perceived stress. The Montreal Imaging Stress Task (MIST) is used to simulate the different stress conditions, including both mental stress and psychosocial stress factors that are relevant at the workplace. A sensor-based wearable device is used to acquire the electrocardiogram (ECG) and respiration (RSP) signals from 39 participants. We extract stress-related features from ECG and RSP, and the Random Forest is used to select the optimal feature combination, which is later fed to the classifier. Four classifiers are investigated about their ability to predict the three stress levels. Finally, the combination of Random Forest and Support Vector Machine (SVM) achieve the best performance. With this method, the accuracy is improved from 78% to 84% in three states classification. And in binary stress detection, the accuracy is 94%.



## 1. Introduction

Work-related stress has drawn great interest in modern society. In 2007, stress was identified to be one of the most common health problem inducements in the European Union [1]. The poor match between people's working ability and demands leads to work-related stress [2]. Moderate stress can stimulate people's potential, while chronic and heavy stress may cause a series of negative effects including depression and even health problems, such as cardiovascular diseases, cerebrovascular diseases and musculoskeletal disorders [1–5]. Excessive workload and stress may make employees absent from job, which results in high economic costs [3]. If high work-related stress could be detected and monitored in time, it is less possible to cause health problems. Further, if the moderate level of stress could be recognized, it could help people maintain the appropriate working state. Therefore detection of different levels of stress is meaningful.

Physiological response corresponds to psychological change and can't be manipulated by people. The mechanism to maintain the body under a stable condition is realized by the autonomic nervous system (ANS), which contains sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). It's known that stress can activate the SNS [6]. And the PNS can bring the body back to a rest state. Intuitively, SNS activation increases the heart rate, whereas PNS decreases it. Activity of SNS and PNS can be monitored through some physiological signals, such as heart rate, heart rate variability (HRV), blood pressure and so on. Also, the respiration under stress is short and rapid, whereas it's deep and slow at a rest state. In our study, we select ECG and respiration signal to measure stress.

There have been many studies on stress detection. Liao used mental arithmetic and an alphabetic task to emulate mental stress. They used facial, physiological signals (heart rate, skin temperature, galvanic skin response), behavioral and task performance (e.g., error rate) as factors [7]. Zhai and Barreto used an interactive 'Paced Stroop Test' to emulate stress. In the test, participants had to select the font color of a word shown on the screen and the word itself named a color [8]. Katsis simulated car races to detect high stress, low stress, disappointment, optimistic and neutral state. They extracted features from facial electromyogram, RSP, electrodermal activity and ECG [9].

* Correspondence author at: 2 Yinghua Dongjie, Hepingli Beijing 100029, China.
** Correspondence author at: No. 19 North 4th Ring Road West, Haidian District, 100190 Beijing, China.
*E-mail addresses:* dryangting@qq.com (T. Yang), zhaozhan@mail.ie.ac.cn (Z. Zhao).

http://dx.doi.org/10.1016/j.compind.2017.05.004
0166-3615/

Muaremi combined the recording of smartphones with subjective assessments and voice messages during workday and recording of HRV data during night. From the smartphone, they extracted audio (microphone), physical activity (accelerometer and GPS) and social interaction (phones calls, address book, calendar and battery) features. They got an accuracy of 55% using only smartphone features, while 59% using HRV features [10]. McDuff measured physiology parameters of Heart Rate and HRV captured at a distance of 3 m using a digital camera. Ball control and card sorting tasks were used to emulate mental stress [11]. In this work, the head or body motions and changes of ambient light might easily impact the accuracy in stress detection.

In stress emulating, the mental arithmetic, alphabetic task, ‘Paced Stroop test’ and other forms of mental stress were often used. Most studies used only mental workload to elicit mental stress, which is certainly an important stress factor in office. However, there are other factors that can elicit psychosocial stress, such as social threat from leaders and colleagues. In our study, to recognize work-related stress, we pick an experiment setting that is very close to a real office situation. Therefore we consider both mental and psychosocial factors by using the MIST which is a standardized task based on computer and psychology [12]. Furthermore, the MIST contains a no stress condition, a moderate stress condition and a high stress condition. As moderate stress is beneficial to work efficiency, one highlight of our study is to detect three levels of stress rather than simply binary stress condition (with or without stress).

After data collection, different algorithms were used to detect stress. Setz compared performances of Linear Discriminant Analysis (LDA), nearest class center and SVM with linear, quadratic, polynomial and rbf kernels in stress detection. With the features extracted from electrodermal activity, they achieved a maximum accuracy of 82.8% by LDA [13]. Liao used a Dynamic Bayesian Network to estimate a continuous stress level [7]. Researchers from the University of Memphis developed a Bayesian Network model of self-reported stress and used a SVM model to predict the instantaneous self-report. With ECG and RSP features, they obtained an accuracy of 72% on filed data [14,15].

In most studies, they only investigated the appropriate classifier to adapt to some kind of feature set. However, the combination of features was often neglected. In our study, we use the Random Forest to find the optimal feature combination, which help improving the performance of classifiers. And to get the best performance, four different classifiers are investigated.

Until now, there is no universally accepted definition of stress or standard database for stress recognition both in lab and in field. In our study, to detect different levels of work-related stress, we make contributions from four aspects.

Firstly, we combine both mental and psychosocial stress factors closing to a real-life office condition. Secondly, we try to detect three levels of stress (no, moderate and high stress) rather than simple binary classification between rest and stress. Thirdly, we use a wearable device to collect ECG and respiration signals, which can provide continuous measurement of stress levels. Finally, we use the Random Forest to find the optimal feature combination, which help improving the performance of classifiers. And to get the best performance, four different classifiers are investigated.

## 2. Data collection

In this study, we consider both mental and psychosocial stress by using the MIST to make the experiment closer to the real office situation. During the experiment, the participants wear the wearable device for collecting ECG and RSP signals.

There are total 39 healthy participants (male: 24; female: 15; mean age: 23.9) participating in the experiments. To ensure the validity and authenticity, the participants are told that they are taking part in an experiment investigating the relationship between cognitive performance and physiological characteristics. Actually, they are confronted with both mental and psychosocial stress.

### *2.1. Experiment*

The Montreal Imaging Stress Task was originally created to evaluate the effects of psychological stress on physiology and brain activity [12]. It has been shown to induce moderate stress response [12]. MIST is an experimental paradigm based on computer and psychological, which mainly consists of four processes: rest, moderate stress, high stress and recovery. The no stress condition is just a rest state. The high stress condition consists of mental arithmetic under time pressure and social-evaluative threat, whereas the moderate stress condition contains only mental arithmetic and moderate social evaluation without any other pressure, which is similar to working under moderate stress.

Fig. 1 shows the experiment procedure for inducing stress in our study. After each condition, we will ask the subjects to give a self-report and questionnaires, which serve as the ground-truth stress level. In 39 subjects, 38 self-reports are consistent with the MIST processes.

We develop the MIST program using the Visual Studio application for Windows. The basic algorithm of the program creates the arithmetic tasks. The algorithm uses up to 4 numbers ranging from 0 to 99 and up to 4 operands containing addition, subtraction, multiplication and division. It is designed to create arithmetic tasks automatically and the solution will be an integer between 0 and 9. The arithmetic tasks are divided into 5 categories. For the first two easiest categories, tasks are only about 2 or 3 one-digit integers and the operands are only addition or subtraction, for example: 6 + 8–9. For the medium two categories, tasks are about 3 or 4 integers with up to 2 integers in 2-digit range and multiplication is allowed, for example: 64–5*11. For the most difficult category, tasks are about 4 integers that can be in 2-digit range. Multiplication and division will be used, for example: 12*14/21-2.

During the rest condition (for 5 min), the participants do not have any task, which is the most relaxing condition.

The moderate stress condition (for 4 min) contains only mental arithmetic and moderate social evaluation, which is similar to working on a computer under moderate stress. When the participant submits the answer, the screen will display “Right!” or “Wrong!” as a feedback. And the leader will give some friendly and moderate giveback, such as “Just do as much as you can!” or “Is there a problem for solving the tasks?”. During moderate stress condition, only mild social stress is induced, whereas strong social stress is induced during high stress condition. During the moderate

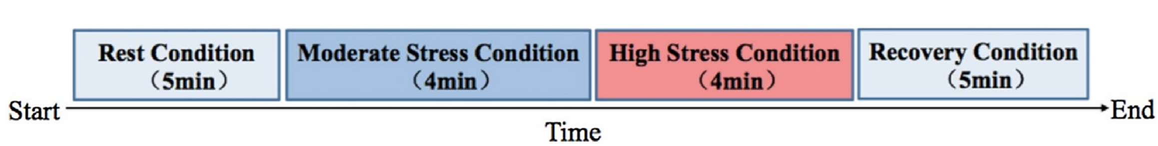


**Fig. 1.** The experiment procedure for inducing stress. After each condition, we will ask the subjects to give a self-report and questionnaires.

stress condition, the program will record the subject's average time needed to solve problems at various difficulty categories. And the recorded time is used to set a time limit for the high stress condition.

During the high stress condition (for 4 min), the participants do mental arithmetic tasks under a time limit. The time limit is 10% less than the subject's average response time recorded in the moderate stress condition. In addition, the program will continuously record the number of correct responses and average response time. The program reduces the time limit to 10% less than the average time for 3 correctly solved tasks, if the subject correctly solves 3 consecutive arithmetic tasks. The program can adaptively adjust time constraints and difficulties in order that the participants can only solve 45%–50% of the problems correctly. This is similar to the high stressful work situation where the work requirements do not match one's ability. Fig. 2 shows the computer program display during the high stress condition. Besides arithmetic tasks and the numbers for submitting answer, a time bar reminds the participant of the remaining time for this task and "Time Out!" will be displayed on the screen when time is out. When the participant submits the answer before the given time, the screen will also display "Right!" or "Wrong!" as a feedback. The color bar indicates the performance of the participant and a simulated average performance level. The participants will be told that the experiment will fail if his performance can't reach the blue area of the bar, which will help add the factor of social evaluative threat. Additionally, after some task the study leader will give a feedback about the subject's performance, such as "Your math is really poor!", "Make more effort to solve more!" or other similar words to increase the stress level. It is similar to the real-life situation when the boss or leader complains about the subject's working performance.

The last condition of recovery (for 5 min) helps the subjects to return to normal.

### 2.2. Devices

During the experiment, the subjects wear a sensor-based wearable device. The device mainly consists of a MCU micro-controller, biomedical sensors and a wireless transmission module. The MCU contains an ARM Cortex-M micro-controller which reduces the power consumption of the device. The biomedical sensors include a three-lead ECG sensor, a 9-axis accelerometer, a body temperature sensor and a photoplethysmography sensor. The three-lead ECG sensor can also measure respiratory according to the impedance pneumography. The measurements of the sensors are transmitted wirelessly using Bluetooth communication protocol controlled by the micro-controller.

To balance the power consumption and quality of the signals, we set the sampling rate for the sensor to 250 Hz and the samples are transmitted at the rate of 25 packets/s. The format of data frame is an HCI command, an identification code, the data length and information field. After receiving the samples, we distinguish different kinds of data according to the identification code. Then we parse the data according to the data length and information field. The device can work around 5 days normally between successive battery recharges. The appearance of the device is shown in Fig. 3.

## 3. Data processing and model development

In this part, some details of data processing will be described, including feature extraction, feature selection and classifier training.

With the MIST experiment, we collect ECG and RSP from 39 participants. The collected data is segmented into one-minute intervals. We extract features from ECG and RSP for each interval. To find the highly correlated features with the stress, we use the Random Forest to sort the significance of the features and select the optimal feature combination. Finally, the optimal features are used to train and test the classifier. The overview of our model development is shown in Fig. 4.

### 3.1. Feature extraction

The received data is segmented into one-minute intervals. For each interval, we will calculate base features from time and frequency domains. Based on these base features, statistical features such as mean, variance, standard deviation are calculated.

#### 3.1.1. ECG features

According to the characteristics of ECG as shown in Fig. 5, R peaks are all automatically detected and extracted using Pan and Tompkins's algorithm [16]. The time difference between two successive R peaks is called RR interval or inter beat interval (IBI) [15]. The missed R peak will lead to RR interval invalid, which can't be detected by PT algorithm. In order to improve the accuracy of RR interval extraction, we filter the extracted RR intervals according to the criterion beat difference (CBD) [17]. Combined with this algorithm, the accuracy of RR intervals extraction can be increased from 98.6% to 99.04% [15].

To eliminate the personal specific components from the RR intervals distribution, we use the z-score normalization for the RR

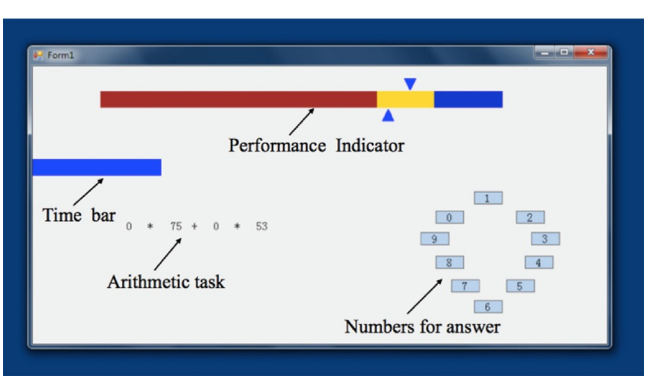


**Fig. 2.** MIST screen during high stress condition.

**Fig. 3.** The appearance of the wearable device.

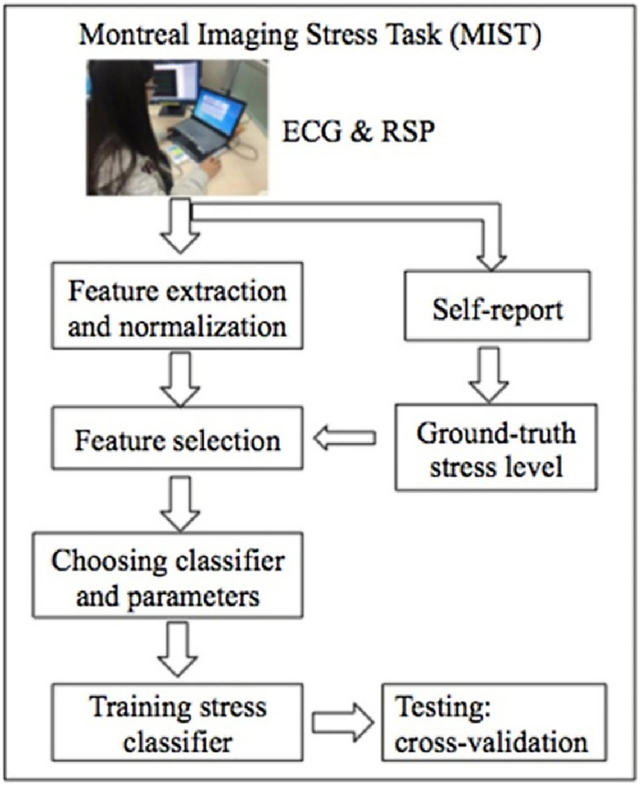


**Fig. 4.** Overview of our model development. It mainly contains data collection, feature extraction, feature selection, classifier and parameters selection, classifier training and model testing.

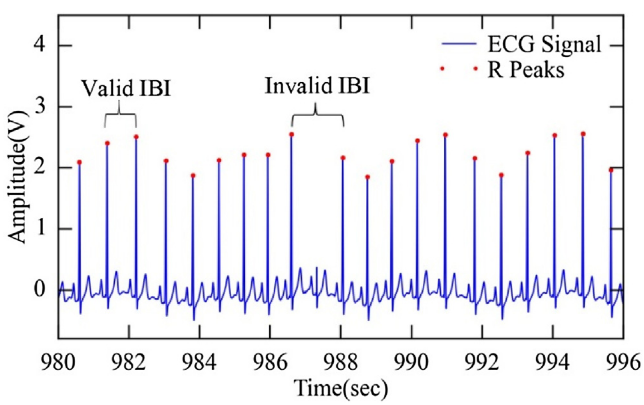


**Fig. 5.** ECG signal and valid or invalid IBI. The accuracy of RR intervals extraction is increased by detecting the missed R peaks and filtering the invalid IBI.

intervals. Then there is no need to construct a personalized model after extensive training. Anyone who is not in the training set can use this model.

The z-score normalization is calculated as in (1). In the formula, $x$ is the value of the original feature, $z$ is the z-score normalized value, $\mu$ is the mean of the feature and $\sigma$ is the standard deviation of the feature.

$$z = \frac{(x - \mu)}{\sigma} \tag{1}$$

Then we calculate the statistical features of RR intervals for all one-minute windows based on the normalized RR intervals. All ECG features are listed in Table 1. They are used for training and testing the model of stress recognition together with RSP features. Specially, the quartile deviation is one half of the difference obtained by subtracting the first quartile from the third quartile in a feature set. The $n^{th}$ percentile of a set of features is the value at which n% of the features are below it, for example the 20th percentile is the value at which 20% of the features are below it.

### 3.1.2. Respiration features

The peaks and valleys of the respiration signal are detected firstly and then each cycle is identified [18]. As in Fig. 6, we calculate the moving average curve according to the period T and the amplitude of the RSP. The RSP frequency $f$ can be estimated by the peak location in the power spectrum. Consequently, the period T (T = 1/f) can be obtained. Then the inspiratory (up) and expiratory (down) phases can be determined. The peaks and valleys are the maximum and minimum values between pairs of the inspiration and expiration phases. If the respiration amplitude is less than 20% of the mean amplitude, the pair of peak and valley will be deleted. In addition, the upper limit of the respiration cycle is set to be 12.5 s and the lower limit to be 0.9 s [19,20].

The base features extracted from the RSP are shown in Table 2. As shown in Fig. 6, the inspiration duration is the time of inspiration, the expiration duration is the time of expiration and the respiration duration is the total time of the respiration cycle. I/E ratio is the ratio of the inspiration and expiration duration. Stretch is the difference of the peaks and valleys within a cycle. In Fig. 7, we can see that respiration under stress is short and rapid, whereas it's deep and slow at a rest state.

The excitability of the SNS and PNS changes during the breath, which leads to the changes of RR interval. That's the respiratory sinus arrhythmia (RSA). So we also calculate RSA feature combined ECG with RSP. RSA is calculated by computing the difference

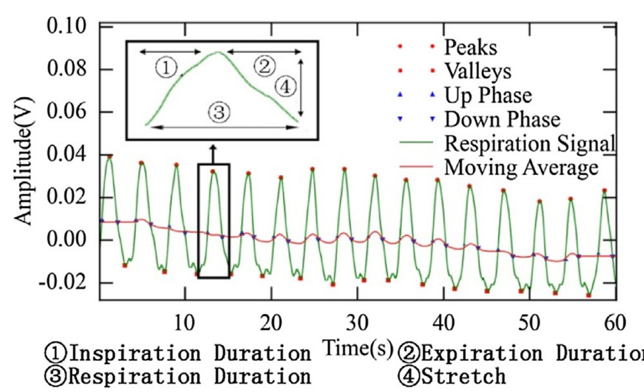


**Fig. 6.** Respiratory signal and calculation of base respiration features.

**Table 1**
All ECG features, extracted using the filtered and normalized RR intervals. They are used for training and testing the stress classifier with RSP features.

| Base Features | Statistical Features | |
|---|---|---|
| | Time Domain | Frequency Domain |
| RR Interval | mean, median, variance, quartile deviation, 20th percentile, 80th percentile, heart-rate | low frequency energy (0.04–0.15 Hz), high frequency (0.15–0.4 Hz), low/high frequency energy ratio |

**Table 2**
All RSP features. The statistical features are calculated using the normalized base features. They are used for training and testing the stress classifier with ECG features.

| Base Features | Statistical Features |
|---|---|
| Inspiration duration, expiration duration, respiration duration, I/E ratio, stretch, respiratory sinus arrhythmia (RSA), breath rate | Mean, median, 80th percentile, 20th percentile, quartile deviation |

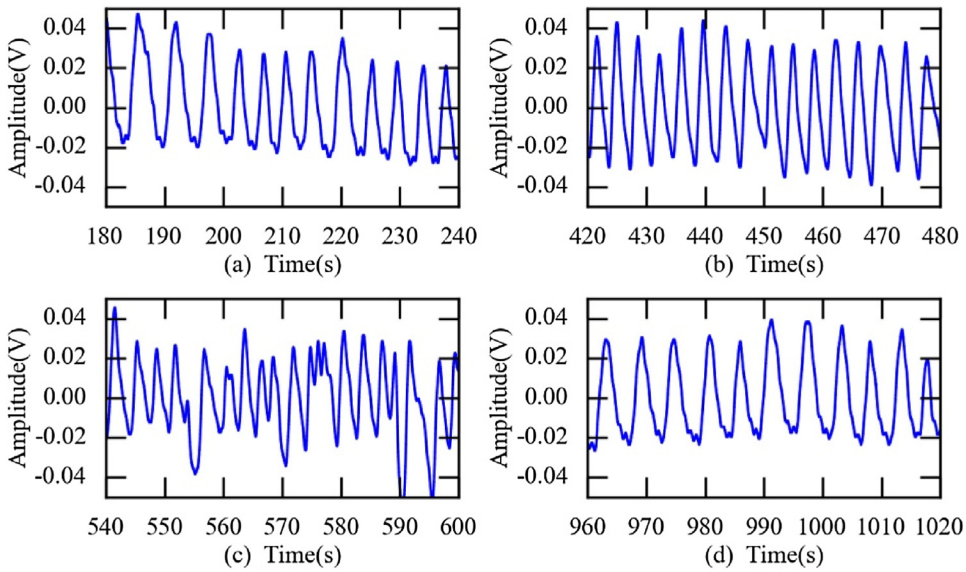


**Fig. 7.** The respiratory signals of rest, moderate stress, high stress and recovery conditions. (a) is the respiratory signal of rest condition, (b) is the respiratory signal of moderate stress condition, (c) is the respiratory signal of high stress condition, (d) is the respiratory signal of recovery.

between the longest RR interval and the shortest RR interval within every respiratory cycle [21].

The base features of the RSP are also normalized in a similar way with the RR intervals. We calculate the statistical features for each one-minute phrase as listed in Table 2. As with ECG, these parameters are all used as the final features to train and test the stress model.

### 3.2. Feature selection

After normalized by the Min-Max method, the features are calibrated according to the results of self-reports and questionnaires. To find the highly correlated features with the stress, we use the Random Forest to sort the significance of the features. And the Random Forest also helps us to select the small number of features that can predict stress states sufficiently.

For each decision tree in a Random Forest, the corresponding out-of-pocket (OOB) data is used to calculate its out-of-pocket data error, denoted as errOOB1. And then, adding random noise to the characteristic X of all samples of the OOB, which can change the value of the sample at feature X randomly. Again we calculate the out-of-pocket data error, denoted as errOOB2. Assuming N trees in a random forest, then the significance of the feature X is [22]:

$$x = \sum (errOOB2 - errOOB1)/N \tag{2}$$

Firstly, the Random Forest sorts the features in a descending order according to the significance. Then it removes some unimportant features to obtain a new feature set. It creates a new random forest using the new feature set and repeats the above steps to obtain some new feature sets. According to each new feature set and new random forest, we calculate the errOOB1 and select the feature set with the lowest errOOB1 as the final feature set [23].

### 3.3. Classifier training

Four classifiers are investigated in our study. They are SVM, Linear Discriminant Analysis (LDA), Adaboost and Nearest Neighbors (KNN).

The principle of KNN method is to find a number of training samples closest to the test sample in distance and predict it from these training samples. The number of samples can be a user defined constant or vary based on the local density of samples. The KNN is a non-generalizing machine learning method for only remembering all of training samples. However, it's often useful when the decision boundary is irregular [24].

The Adaboost classifier is a meta-estimator. It starts by fitting the classifier on the original dataset and fits the additional copies on the same dataset. But the weights of the incorrectly classified samples will be adjusted so that the classifiers later focus more on these cases [25]. And the LDA is a classifier with a linear decision boundary. The fitted LDA model can be used to reduce the dimensionality of the input [26].

The SVM is a kind of classification algorithm based on statistics. It will map the features to high dimension spaces and then use a simple linear model to classify. Given training dataset in two classes $\{x_i, y_i\}$, i = 1, 2, $\cdots$ n, x ∈ RD, $y_i \in \{1, -1\}$. For a linear problem, there is a hyperplane of w·x + b = 0 that can classify the dataset. To non-linear problem, the optimal hyperplane for classification is [27]:

$$\min_{w,b,\xi_i} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i \tag{3}$$

Then the optimization problem for SVM is:

$$y_i [(w \cdot \Phi(x_i)) + b] \geq 1 - \xi_i, \xi_i \geq 0, i = 1, 2, 3, \cdots, n \tag{4}$$

Its dual is:

$$\min_{\alpha} \frac{1}{2}\alpha^T \mathbf{Q}\alpha - \mathbf{e}^T\alpha \tag{5}$$

Subject to:

$$y^T\alpha = 0, 0 \le \alpha_i \le C, i = 1, 2, \cdots, n \tag{6}$$

Where $\boldsymbol{e}$ is the vector of all ones, $\mathbf{Q}$ is an n by n positive matrix, $\alpha$ is the Lagrangian coefficient, $\xi_i$ is the slack variable. Training vectors are implicitly mapped into a higher dimensional space by the kernel function $\Phi(x_i)$. The penalty coefficient C determines the degree of penalty when a sample is misclassified [27].

The SVM has a good performance in solving non-linear problems. However, it's sensitive to parameters and kernels. Different kernel functions can be used for various usages.

Among the four classifiers, the SVM classifier achieves the best performance in training and testing. So we use the SVM to train the classifier model. 80% of the data set is used as the training set and the remaining 20% as the testing set. In training, the RBF Kernel is used. And the RBF Kernel function is:

$$-\gamma|x - x^{'}|^2 \tag{7}$$

In three states classification, we use the OneVsRestClassifier function [28]. Namely, in training, three estimators are created to distinguish each state from the other two states. The learned model is a hyper-plane defined in high-dimensional function space and selected by the SVM algorithm to maximize the margin of separation. The performance of the SVM model is highly sensitive to the penalty coefficient C and RBF $\gamma$. In order to select the best value of C and $\gamma$, the grid-search method is used. And we use the cross-validation to evaluate the model performance. Namely, we use the features of each participant in testing and use all other participants' features to train the model. The output of the learned model is the probability that the input belongs to.

After training, we chose the following parameters to evaluate the performance of the model. TPR is the true positive rate. FPR is the false positive rate. ROC curve is the receiver operating characteristic. AUC is the area under ROC curve which will be a value between 0.5 and 1. And the AUC is larger, the performance of the classifier is better. F1-score is the weighted average of the precision and recall, where an F1-score reaches its best value at 1 and worst at 0. Its value is:

$$F1 - score = \frac{2}{1/precision + 1/recall} \tag{8}$$

## 4. Results

In this section, we present the results of feature selection and classification. The Random Forest gets the significance of each feature set based on the score. Among all the feature sets, the scores of mean of RR interval, 80th percentile of RR interval, mean of expiration duration, mean of respiration duration, 80th percentile of respiration duration are relatively low. We remove the five highly correlated features. Fig. 8 shows the scores of the top five and last five features. And Table 3 shows the results of the four classifiers before and after the feature selection in three conditions classification. With the SVM classifier, the accuracy is increased from 78% to 84% with Random Forest feature selection.

Then we compare the predictive power of different feature sets, including the whole set of ECG and RSP features, only ECG features and only the RSP features with SVM classifier. Table 4 shows performance of different feature sets in classification of rest and stress conditions with the SVM classifier. Compared to using only ECG or RSP feature, the combination of ECG and RSP features achieves a higher accuracy.

Most importantly, we can not only distinguish between rest and stress conditions, also can discriminate three levels of stress that contains no stress, moderate stress and high perceived stress. After feature selection, with the remaining 36 features, the results of the four classifiers in three states classification are listed in Table 3. From it we can see that the SVM classifier gets the best performance. In Table 5, it's the result of SVM in three states classification. And in Table 6, it's the confusion matrices in three states classification with SVM. In Fig. 9, it shows the ROC curve for

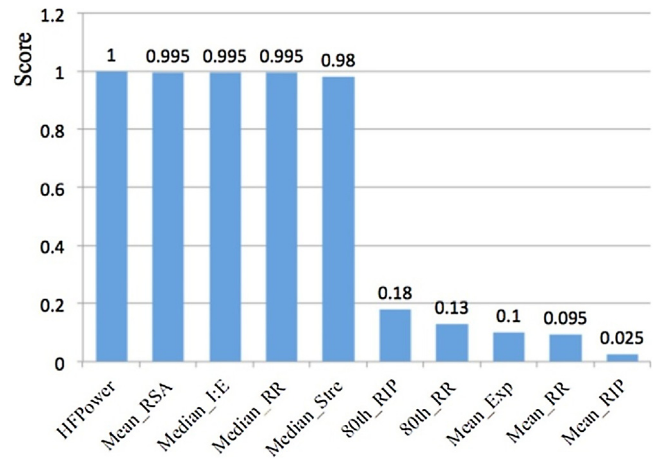


**Fig. 8.** Scores of top 5 and last 5 features calculated by the Random Forest.

**Table 3**
The results of four classifiers before and after the feature selection in three conditions classification.

| Feature set | Classifier | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|---|
| All Features (41) | SVM | 0.78 | 0.79 | 0.78 | 0.78 |
| | LDA | 0.77 | 0.78 | 0.77 | 0.78 |
| | KNN | 0.72 | 0.72 | 0.72 | 0.72 |
| | Adaboost | 0.79 | 0.77 | 0.77 | 0.77 |
| Remaining Features (36) | SVM | 0.83 | 0.84 | 0.83 | 0.84 |
| | LDA | 0.80 | 0.80 | 0.80 | 0.80 |
| | KNN | 0.74 | 0.72 | 0.73 | 0.72 |
| | Adaboost | 0.79 | 0.79 | 0.79 | 0.79 |

**Table 4**
Predictive power of different feature sets in two states classification with SVM.

| Feature set | Data set | Precision | Recall | FPR | Accuracy |
|---|---|---|---|---|---|
| All | Train | 0.96 | 0.96 | 0.01 | 0.96 |
| | Test | 0.94 | 0.94 | 0.01 | 0.94 |
| ECG | Train | 0.87 | 0.87 | 0.05 | 0.87 |
| | Test | 0.85 | 0.85 | 0.06 | 0.85 |
| RSP | Train | 0.94 | 0.94 | 0.02 | 0.94 |
| | Test | 0.93 | 0.93 | 0.03 | 0.93 |

**Table 5**
The result of SVM classifier in three conditions classification with the remaining 36 features in training and testing.

| Data set | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| Train | 0.89 | 0.89 | 0.89 | 0.89 |
| Test | 0.83 | 0.84 | 0.83 | 0.84 |

**Table 6**
Test confusion matrix of three conditions classification with SVM.

| | | Classified by Model | | | |
|---|---|---|---|---|---|
| | | No Stress | Moderate Stress | High Stress | Total |
| Actual | No Stress | 73 (99%) | 0 (0%) | 1 (1%) | 74 |
| | Moderate Stress | 2 (6%) | 25 (71%) | 8 (23%) | 35 |
| | High Stress | 9 (28%) | 2 (6%) | 21 (66%) | 32 |
| | Total | 84 | 27 | 30 | 141 |

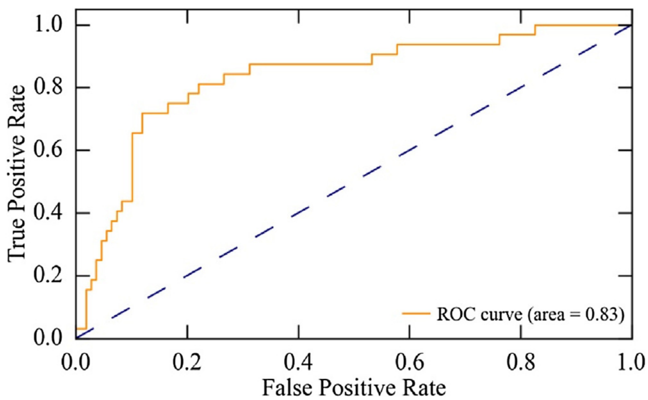


**Fig. 9.** ROC curve for high stress condition with SVM classifier.

prediction of high stress condition with SVM. And Fig. 10 shows the ROC curves for three classes with SVM.

In the studies of stress detection, scholars use various kinds of experiments to eliminate stress. For example, public speaking, mental arithmetic and physical stimulation are all commonly used experimental methods. Setz also used the MIST experiment to induce stress. They achieved a maximum accuracy of 82.8% for discriminating stress from cognitive load. However, they measured the electrodermal activity to detect stress [13]. So the results cannot be compared directly.

## 5. Discussion, conclusion and outlook

As presented in the previous section, 36 features are selected by Random Forest. After feature selection, the accuracy is increased from 78% to 84% in three states classification with SVM classifier. In testing, the accuracy of 94% is achieved in distinguishing only between rest and stress conditions with SVM classifier. Furthermore, in discriminating three stress levels of no, moderate and high stress conditions, the SVM classifier performs best compared to LDA, Adaboost and KNN classifier. Its accuracy is 84%. The performance of classification model is better with the combination of ECG and RSP features.

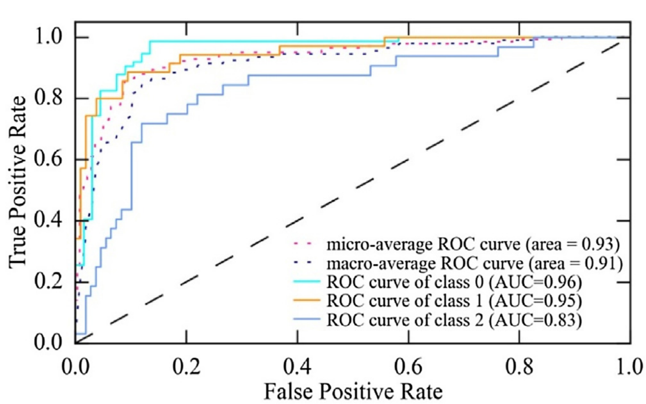


**Fig. 10.** ROC curve for three classes with SVM classifier.

However, there is no universally accepted definition of stress. The widely used methods to assess stress are cortisol and self-reporting. But the correlations between them are only limited to 0.26–0.36 [29,30]. Furthermore, there is no standard database for stress recognition either in lab or in field. Scholars use various kinds of experiments to eliminate stress and different types of physiological data and features to recognize stress. And many physiological signals are collected using various wearable devices, such as ECG, blood pressure, electrodermal activity, photoplethysmogram. So it's not possible to compare the accuracy among so many different studies directly.

In this work, we have got an accuracy of 84% in discriminating three stress levels. However, this work has several limitations and significant potential for future works. Firstly, we achieve this in an experimental setting. As the MIST represents a closer to real-life work stress condition, we expect similar results in real-life and long-term work experiments for field usage. Secondly, several approaches could be adopted to improve the accuracy such as better processing of data normalization, handling of physical activity confounds and combining with more physiological signals such as skin temperature and galvanic skin response. Thirdly, real-time detection and effective visualizations could be developed that permit users to visualize their stress patterns on mobile devices. This will give feedback to the users and remind them to pay attention to break. Finally, since stress may be private information for someone, we should pay more attention to privacy management for the sensor data.

In conclusion, in our study, we combine both mental and psychosocial stress factors closing to a real-life office condition. In our experiment, the monitoring of ECG and RSP signals can discriminate no, moderate and high stress levels rather than only distinguishing between rest and stress conditions. It's more meaningful. We achieve a higher accuracy by combining Random Forest feature selection with SVM classifier. Totally we got an accuracy of 84% in discriminating three stress levels in a more comfortable way by using a wearable device.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China (No 61302033), National Key Research and Development Project2016YFC1304302 and Key Project of Beijing Municipal Natural Science FoundationZ16003.

## References

[1] G. Nema, Y.M. Dhanashree Nagar, A study on the causes of work related stress among the college teachers, Pac. Bus. Rev. Int. (2010) 1–7.
[2] S.L. Sauter, L.R. Murphy, J.J. Hurrell, Prevention of work-related psychological disorders: a national strategy proposed by the National Institute for Occupational Safety and Health (NIOSH), Am. Psychol. 45 (10) (1990) 1146.
[3] J. Hillebrandt, Work-related Stress and Organizational Level Interventions Addressing the Problem at Source, (2007) .
[4] R. Rosmond, P. Björntorp, Endocrine and metabolic aberrations in men with abdominal obesity in relation to anxio-depressive infirmity, Metabolism 47 (10) (1998) 1187–1193.
[5] R. Rosmond, M.F. Dallman, P. Björntorp, Stress-related cortisol secretion in men: relationships with abdominal obesity and endocrine, metabolic and hemodynamic abnormalities 1, J. Clin. Endocrinol. Metab. 83 (6) (1998) 1853–1859.
[6] C. Tsigos, G.P. Chrousos, Hypothalamic-pituitary-adrenal axis: neuroendocrine factors and stress, J. Psychosom. Res. 53 (4) (2002) 865–871.
[7] W. Liao, W. Zhang, Z. Zhu, et al., A real-time human stress monitoring system using dynamic Bayesian network, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Diego, California, 2005, pp. 70–78.
[8] J. Zhai, A. Barreto, Stress detection in computer users based on digital signal processing of noninvasive physiological variables, 28th Annual International

Conference of the IEEE on Engineering in Medicine and Biology Society, New York, NY, USA, 2006, pp. 1355–1358.
[9] C.D. Katsis, G. Ganiatsas, D.I. Fotiadis, An integrated telemedicine platform for the assessment of affective physiological states, Diagn. Pathol. 1 (1) (2006) 16.
[10] A. Muaremi, B. Arnrich, G. Tröster, Towards measuring stress with smartphones and wearable devices during workday and sleep, BioNanoScience 3 (2) (2013) 172–183.
[11] D.J. McDuff, J. Hernandez, S. Gontarek, et al., Cogcam: contact-free measurement of cognitive stress during computer tasks with a digital camera, Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, Santa Clara, California, USA, 2016, pp. 4000–4004.
[12] K. Dedovic, R. Renwick, N.K. Mahani, et al., The Montreal Imaging Stress Task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain, J. Psychiatry Neurosci.: JPN 30 (5) (2005) 319.
[13] C. Setz, B. Arnrich, J. Schumm, et al., Discriminating stress from cognitive load using a wearable EDA device, IEEE Trans. Inf. Technol. Biomed. 14 (2) (2010) 410–417.
[14] E. Ertin, N. Stohs, S. Kumar, et al., AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field, Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems, Seattle, WA, USA, 2011, pp. 274–287.
[15] K. Hovsepian, M. al'Absi, E. Ertin, et al., cStress: towards a gold standard for continuous stress assessment in the mobile environment, Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Osaka, Japan, 2015, pp. 493–504.
[16] J. Pan, W.J. Tompkins, A real-time QRS detection algorithm, IEEE Trans. Biomed. Eng. 32 (3) (1985) 230–236.
[17] G.G. Berntson, K.S. Quigley, J.F. Jang, et al., An approach to artifact identification: application to heart period data, Psychophysiology 27 (5) (1990) 586–598.
[18] W. Lu, M.M. Nystrom, P.J. Parikh, et al., A semi-automatic method for peak and valley detection in free-breathing respiratory waveforms, Med. Phys. 33 (10) (2006) 3634–3636.
[19] D.H. McFarland, Respiratory markers of conversational interaction, J. Speech Lang. Hear. Res. 44 (1) (2001) 128–143.
[20] J.A. Neder, S. Dal Corso, C. Malaguti, et al., The pattern and timing of breathing during incremental exercise: a normative study, Eur. Respir. J. 21 (3) (2003) 530–538.
[21] C.L. Stephens, I.C. Christie, B.H. Friedman, Autonomic specificity of basic emotions: evidence from pattern classification and cluster analysis, Biol. Psychol. 84 (3) (2010) 463–473.
[22] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (8) (1998) 832–844.
[23] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[24] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (8) (1998) 832–844.
[25] Y. Freund, R. Schapire, N. Abe, A short introduction to boosting, J.-Jpn. Soc. Artif. Intell. 14 (771–780) (1999) 1612.
[26] R. Fisher, The use of multiple measures in taxonomic problems, Ann Eugenics 7 (2) (1936) 179–188.
[27] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 1992, pp. 144–152.
[28] C.W. Hsu, C.J. Lin, A comparison of methods for multiclass support vector machines, IEEE Trans. Neural Netw. 13 (2) (2002) 415–425.
[29] M. Al'Absi, S. Bongard, T. Buchanan, et al., Cardiovascular and neuroendocrine adjustment to public speaking and mental arithmetic stressors, Psychophysiology 34 (3) (1997) 266–275.
[30] M. Al'Absi, D. Hatsukami, G.L. Davis, et al., Prospective examination of effects of smoking abstinence on cortisol and withdrawal symptoms as predictors of early smoking relapse, Drug Alcohol Depend. 73 (3) (2004) 267–278.