AN INVESTIGATION OF DECISION ANALYTIC METHODOLOGIES FOR STRESS IDENTIFICATION

Yong Deng[1], Chao-Hsien Chu[2], Huayou Si[3], Qixun Zhang[4], Zhonghai Wu[4]

1. School of Electronic Engineering and Computer Science, Peking University, Beijing, China

2. School of Information Systems, Singapore Management University, Singapore

3. School of Computer Science and Technology, Hangzhou Dianzi University, Zhejiang, China

4. School of Software and Microelectronics, Peking University, Beijing, China

Emails: dengyong@pku.edu.cn; chchu@smu.edu.sg; sihy@hdu.edu.cn; zhangqx@ss.pku.edu.cn; wuzh@pku.edu.cn

Submitted: Jan. 18, 2013 Accepted: Jan. 31, 2013 Published: Sep.05, 2013

Abstract- In modern society, more and more people are suffering from some type of stress. Monitoring and timely detecting of stress level will be very valuable for the person to take counter measures. In this paper, we investigate the use of decision analytics methodologies to detect stress. We present a new feature selection method based on the principal component analysis (PCA), compare three feature selection methods, and evaluate five information fusion methods for stress detection. A driving stress data set created by the MIT Media lab is used to evaluate the relative performance of these methods. Our study show that the PCA can not only reduce the needed number of features from 22 to five, but also the number of sensors used from five to two and it only uses one type of sensor, thus increasing the application usability. The selected features can be used to quickly detect stress level with good accuracy (78.94%), if support vector machine fusion method is used.

Index terms: Stress detection, physiological sensors, feature selection, information fusion, classification

I. INTRODUCTION

In modern society, more and more people are suffering from some type of stress. There is a strong link between stress and overall health condition of human beings. According to a latest survey by the American Psychological Association [3], more than half (56%) of the Americans indicated that stress is a main source of their personal health problems. Also, more than 94% of the adults believed that stress can contribute to the development of major illnesses, such as heart disease, depression and obesity, and that some types of stress can trigger heart attacks, arrhythmias and even sudden death, particularly for people who have cardiovascular disease. In another research, Nako [19] reported that work related stress is a key cause of mental illness health in worldwide populations. For instance, in Canada, 28% of workers reported that they are either 'quite a bit' or 'extremely' stressful most days at work [6]. Similarly, in the United Kingdom, the Labor Force Survey [20] showed that there are 760 incidence cases of work-related stress, depression or anxiety for every 100,000 workers.

Although people perceive that stress can have negative impact on health and well-being, they normally do not take action to prevent stress or manage it. The APA survey [3] suggested that time management may be a significant barrier preventing people from taking the necessary steps to improve their health. Effectively detecting the stress of human beings in time not only provides a way for people to better understand their stress condition but also provides physicians with more reliable data for intervention and stress control.

In general, there are two streams of approaches to identify the stress level people are suffering. One stream uses psychological self-assessment in the form of questionnaires and the other one is through the analysis on the information acquired by the physiological sensors that people wear. Identifying the stress of human beings using psychological sensors has been a hot research topic in recent years. Existing studies have shown that psychosocial stress can be recognized by the physiological information of human being. The physiological information, which can be acquired by biological or physiological sensors, usually includes Elecardiagram (ECG), Galvanic Skin Response (GSR), Electromyogram (EMG), and Respiration (RESP).

The process of detecting stress using physiological sensors normally consists of three major phases. See Figure 1. First, features are extracted from the raw physiological sensor data using feature extraction algorithms. In order to effectively identify the stress level or patterns, many

features must be extracted from a variety of physiological sensors. Secondly, most relevant features are selected by using some feature selection heuristics. The more features extracted does not necessarily mean the better performance of stress identification. On the other hand, more features may bring in useless information or even misleading information. Malhi and Gao [17] have shown that some features provide contradictory information and thus decrease the quality of data analysis. Also, in real time stress detection more features mean more data processing, which may reduce or limit the real time performance. Moreover, it is not realistic for people to wear too many physiological sensors, as that will bring inconvenience as well as discomforts to them. Therefore, selecting as least features and predicting as accurate as possible for stress detection is a challenging research work to do. Finally, based on the selected features, information fusion procedure is applied to identify the stress level or patterns. The fusion methods used in previous studies include linear discriminant function (LDF), C4.5 induction tree, support vector machine (SVM), Naïve Bayes (NB), and others (see Section II for details).

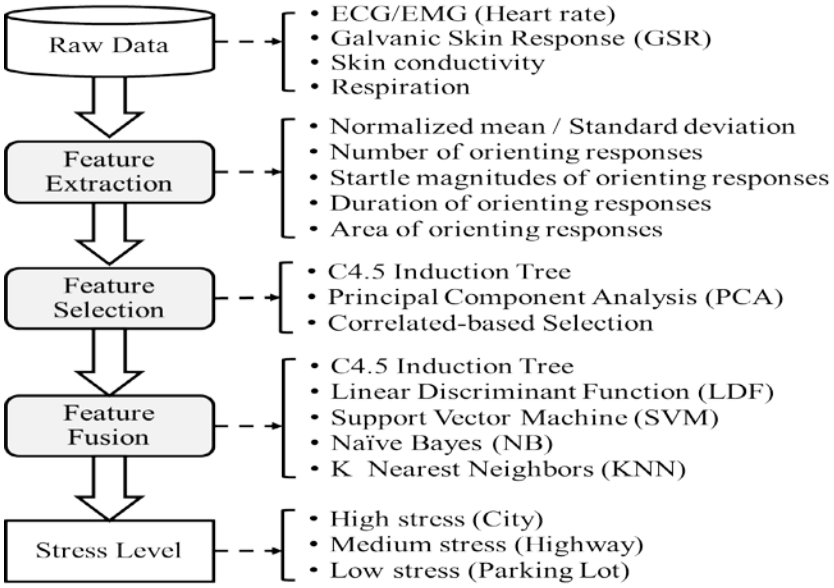

Figure 1. The generic process of stress identification

The main purposes of this study are threefold: (1) to identify and examine features that are relevant to stress identification, (2) to select an effective feature selection method and (3) to evaluate the relative performance of different information fusion methods. We use the same data set and feature extraction methods as used in Healey and Picard [13, 14] to benchmark our

analysis. We first screen the data set for potential errors and noise and extract a total of 22 features for each data segment from the data set. We then apply three feature selection heuristics: Decision Tree, Principal Component Analysis (PCA) and Correlation-based Analysis to obtain reduced sets of features. The effectiveness of these heuristics are then investigated by comparing the correct rate and computational time using five information fusion algorithms – LDF, C4.5, SVM, NB and K-Nearest Neighbors (KNN) – to benchmark with the full feature set.

The organization of this paper is as follows. In section II we briefly review related work. Decision analytic methodologies, including feature extraction, feature selection algorithms, and information fusion algorithms, are introduced in Section III. The data set and the experimental setting used for performance evaluation are described in Section IV. The analysis on the results, including the effects of feature selection approaches and fusion methods are presented in Section V. We then provide conclusion and discuss limitation of the paper in Section VI.

II. RELATED WORK

The questionnaires-based self-assessment approach has a long history. Davidson, et al. [8] developed a self-rated scale tailored to 17 symptoms of the Diagnostic and Statistical Manual of Mental Disorders, which can be used to measure symptom frequency as well as severity of stress treatment; for example, measurement of stress symptom change over time, response prediction, and evaluation of differences between stress therapy modalities in the research setting. A stress management system based on the questionnaires was developed to assist the determination of the negative stress level and resolve the problem for lessening it [16]. However, Watson and Pennebaker [26] found that the self-report measures of stress as well as health contain a significant negative affectivity component, thus, correlation between such measures likely overestimate the true association between stress and health.

Some work has been done to detect stress from physiological measurements. Jovanov, et al. [15] used heart-rate variability (HRV) to quantify stress level prior to and during training as well as to predict stress resistance. Angus and Zhai [4] reviewed methods as well as challenges toward the automated assessment of emotional stress by monitoring and recording three psychophysiological signals: Blood Volume Pulse (BVP), GSR and Skin Temperature (ST). Zhai and Barreto [30] monitored four kinds of physiological signals -- GSR, BVP, Pupil Diameter (PD), and ST -- in

the computer users and used three machine learning approaches, NB, SVM and Decision Tree, to classify stress types. Bakker et al. [5] used GSR sensor to detect changes in the stress level by both performance monitoring-based change detection with the non-parametric test and change detection based on raw data using adaptive windowing.

Healey and Picard [13, 14] have conducted in-depth studies in stress detection for real world driving tasks. They continually recorded ECG, EMG, GSR, and RESP signals of drivers on a fixed route through downtown Boston, covering three different conditions: rest, highway and city road. A total of 22 features were extracted from the recorded signals and proposed to use LDF as the fusion method to predict drivers' stress during their driving. They have partially released their physiological signal data record on the PHYSIONET website [21]. Although the driver data set they contributed does not contain the exact same data of all the drivers as that in their experiments, the data set allows other researchers to further explore stress detection. Their work used the full feature set, which might result in higher computational burden and user resistance especially in real time stress recognition.

Akbas [2] presented an evaluation based on the driver dataset of Healey and Picard [13, 14]. His evaluation contained the mean as well as standard deviation of Instant Heart Rate (IHR), Hand-based Skin Conductance, Foot-based Skin Conductance, Amplitude of EMG, and Instantaneous Respiratory Rate (IRR). Besides, he has also evaluated the segment-based data arrays such as IRR and average number of Contractions per Minute (CPM) derived from the RESP and EMG signals respectively by using a peak detection algorithm. Zhang, et al. [31] presented a systematic approach using a structurally learned Bayesian Network to fuse the sensor feature information and concluded that good correct rate can be acquired. In the paper, feature selection approach was referred, however, neither original data segments from which features were extracted are mentioned nor are the feature selection results shown in details.

Table 1 provides a summary of the data and methods used in previous studies. As shown, previous studies on stress detection using physiological data have predominately focused on using (1) a variety of physiological sensors; (2) methods of extracting features from sensor data; and (3) methods of detecting stress. Additional work still needs to be explored to understand: (1) the importance of feature types; (2) the effect of feature selection; (3) the relative performance of different feature selection algorithms; (4) the relative performance of different fusion algorithms

and (5) integrating self-assessment and physiological data for improving stress detection. In this study, we focus on investigating the first four research issues.

Table 1：Summary of previous research

Reference	Data/Sensors Used	Feature Extraction	Feature Selection	Stress Detection
[15]	HRV	N/A	N/A	N/A
[4]	BVP, GSR and ST.	N/A	N/A	N/A
[14]	ECG, EMG, GSR, and RESP	Normalized mean, standard deviation, #,startle magnitudes, duration and area of orienting response	N/A	LDF
[30]	GSR, BVP, PD, and ST	N/A	N/A	NB, SVM and Decision Tree
[2]	IHR, GSR, EMG, RESP	Mean, standard deviation and amplitude of EMG	N/A	Peak Detection Algorithm
[31]	GSR	N/A	N/A	Bayesian Network
[5]	GSR	N/A	N/A	Change Detection Algorithm and Non-parametric test

N/A: Not Available

III. DECISION ANALYTIC METHODOLOGIES

This section briefly reviews the core processes used in stress detection (refer to Figure 1), with focus on the proposed feature selection method.

a. Feature Extraction

We clean up the data set before proceeding to feature extraction. First, we segment the data. We use five minutes as the basis to segment the signals from physiological sensors. The segments for “low” stress level are taken from the last five minutes of the rest periods. The segments for the “medium” stress level are taken from the middle five minutes of the highway driving periods. And the segments for the “high” stress level are taken from the middle five minutes of the city driving periods. A total of 65 segments are acquired. We then extract the features applying some simple algorithms such as calculating mean, standard deviation, magnitude, number, frequency, duration and area, etc. as summarized in Table 2. The algorithms used depend on the type of sensor signal we collected. For each segment, 22 features are extracted.

Table 2: Feature symbol and description

Category	Number	Symbol	Feature Description
EMG	1	EMG_mean	The normalized mean of the EMG data
Skin Conductivity	12	FGSR_mean	The normalized mean of the foot GSR data
		FGSR_std	The standard deviation of foot GSR data
		FGSR_freq	The total number of orienting responses of a segment for foot GSR
		FGSR_mag	The summary of the startle magnitudes of orienting responses of a segment for foot GSR
		FGSR_dur	The summary of the duration of orienting responses of a segment for foot GSR
		Fgsr_area	The summary of the area of orienting responses of a segment for foot GSR
		HGSR_mean	The normalized mean of the hand GSR data
		HGSR_std	The standard deviation of hand GSR data
		HGSR_freq	The total number of orienting responses of a segment for hand GSR
		HGSR_mag	The summary of the startle magnitudes of orienting responses of a segment for hand GSR
		HGSR_dur	The summary of the duration of orienting responses of a segment for hand GSR
		HGSR_area	The summary of the area of orienting responses of a segment for hand GSR
Respiration	6	RESP_mean	The normalized mean of the Respiration data
		RESP_std	The standard deviation of Respiration data
		RESP0~0.1	The summary of respiration energy in the bands 0~0.1Hz
		RESP0.1~0.2	The summary of respiration energy in the bands 0.1~0.2Hz
		RESP0.2~0.3	The summary of respiration energy in the bands 0.2~0.3Hz
		RESP0.3~0.4	The summary of respiration energy in the bands 0.3~0.4Hz
Heart Rate	3	HR_mean	The normalized mean of the heart rate data
		HR_std	The standard deviation of heart rate data
		HR_lr	The total energy of Heart Rate in the low frequency band (0-0.08 Hz)

b. Feature Selection

Feature selection plays an important role in predicting both accuracy and real time performance. Feature selection is aiming to reduce the dimensionality of the input features and the number of sensors to wear, which will bring more friendliness to users in real world. We consider three popular algorithms for feature selection.

(1) Induction tree algorithm

C4.5 induction tree algorithm, an improvement over ID3 algorithm, is one of the most popular and practical methods for inductive inference [22, 24]. The algorithm uses information entropy (equation 1) as the metric to evaluate performance and uses information gain (equation 2) to select features. Those features which can result in higher information gain are kept; while those resulting in lower information gain will be discarded. Information entropy can be expressed as:

$$Entropy(S) = \sum_{i=1}^{c} -P_i \log_2 P_i \tag{1}$$

Where, *S* is the data set; *c* is the number of target classes; P_i is the proportion of *S* belonging to class *i*. The information gain obtained from pruning the tree can be expressed as:

$$Gain(S,f) = Entropy(S) - \sum_{v \in Value(f)} \frac{|s_v|}{|S|} Entropy(s_v) \tag{2}$$

Where, *f* is the feature set; *Value (f)* is the set of all possible values for feature *f*; S_v is the subset of *S* whose feature *f has* value *v* (i.e., $s_v = \{s \in S | f(s) = v\}$).

(2) Principal component analysis (PCA)

PCA is often regarded as the simplest true eigenvector-based multivariate analyses [27, 28]. Its operations can be thought as revealing the internal structure of the data in a way which best explains the variance in the data. If a multivariate dataset is visualized as a set of coordinates in a high-dimensional data space, PCA can supply the user with a lower-dimensional picture, a "shadow" of this object when viewed from its most informative viewpoint. Often we only use the first few principal components, which are regarded to represent the whole original dataset, so that the dimensionality of the transformed data is reduced.

According to Duda et al. [9], PCA approach transforms *n* vectors $(X_1, X_2, \ldots, X_i, \ldots, X_n)$ from a *d* dimensional space to vectors $(X'_1, X'_2, \ldots, X'_i, \ldots, X'_n)$ in a new *d'* dimensional space:

$$X'_i = \sum_{k=1}^{d'} a_{k,i} e_k, d' \leq d \tag{3}$$

Where, e_k is the set of eigenvectors corresponding to the largest eigenvalues for the scatter matrix; $a_{k,i}$ are the projections of the original vectors on the eigenvectors. These projections are called the principal components of the original data set. Both *d* and *d'* are positive integers and

the dimension d' cannot be greater than *d*. The *d* by *d* scatter matrix for the original data set is defined as:

$$SM = E[x_i x_i^T], \text{for } i = 1 \text{ to } n \tag{4}$$

Where, $E[x_i x_i^T]$ is the statistical expectation operator applied to the outer product of x_i and its transpose. The representation shown in equation (3) minimizes the error between the original and transformed vectors. This is illustrated by considering the variance of the principal components given by Haykin [12]:

$$\sigma^2(e_k) = E[a_{k,j}^2] = e_k^T S \tag{5}$$

Where, e_k represents the *d* by1 vector. It is evident that the variance of the principal components is a function of the magnitude of the components of the vectore_k. At the local maxima and minima for the variance function in equation (3), the following relationship exists:

$$\sigma^2(e_k + \delta e_k) = \sigma^2(e_k) \tag{6}$$

Equation (4) is satisfied when $(\delta e_k)^T S e_k - \lambda(\delta e_k)^T e_k = 0$, where λ is a scaling factor, which leads to $Se_k = \lambda e_k$. Equation (6) can be recognized as an eigenvalue problem with nontrivial solutions only when λ is the set of eigenvalues for the scatter matrixSM. Thus, the associated vectors (*k*=1 to *d'*) are the eigenvectorse_k. If the condition $d' < d$is satisfied, then the above representation also reduces the dimensionality of the vectors. The error in representation of the original dataset $(X_1, X_2, \ldots, X_i, \ldots, X_n)$ due to the reduction in number of dimensions to d' is given by Haykin [12] as:

$$E_{d'} = \frac{1}{2}\sum_{i=d'+1}^{d} \lambda_k \tag{7}$$

Where, λ_k is the set of eigenvalues of the scatter matrix SM corresponding to the eigenvectorse_k. It is seen from equation (7) that using the eigenvectors corresponding to the largest eigenvalues would give the smallest error in representation. Thus, the variance is maximized in the direction of the eigenvectors. Also, the variance in the directions of the eigenvectors $(e_1, e_2, \ldots, e_k, \ldots, e_d)$ decreases in the same order when$\lambda_1 > \lambda_2 > \cdots \lambda_k \ldots > \lambda_d$. This property has been exploited for several feature selection studies. We selected the most important components from the PCA results, which can almost represent the whole original data set. After that, we chose the most

sensitive features which have the highest correlation with the principal components that we have selected. The correlation between a component and a feature is also called a loading, which can estimate the information they share [1]. The larger the value of loading square is, the more information the feature contains about the corresponding component. Through this way, we can acquire the most important features selected by PCA approach as:

$$C_i = \sum_{j=1}^{n} {l_{i,j}}^2 * w_j \tag{8}$$

Where, C_i is the contribution of feature *i* to the whole feature set; *n* is the total component number; $l_{i,j}$ is the loading between feature *i* and component *j*; ${l_{i,j}}^2$ is the value of the square of $l_{i,j}$; w_j is the contribution of component *j* to the whole feature set. Based on equation (8), these features with highest C_i should be selected.

(3) Correlation-based feature selection (CFS)

Correlation is another useful approach to select features. Normally a good feature subset is the one that contains features highly correlated to the class, yet uncorrelated to each other [11, 29]. If the correlation between two features is high, it means these two features have similar characters for the classification prediction, so we can just select one of them and discard the other one. The following formulas are used in computing the correlation between vectors A and B:

$$r_{A,B} = \frac{cov(A,B)}{\sigma_A \sigma_B} = \frac{E((A-\mu_A)(B-\mu_B))}{\sigma_A \sigma_B} = \frac{E(AB)-E(A)E(B)}{\sqrt{E(A^2)-E(A)^2}\sqrt{E(B^2)-E(B)^2}} \tag{9}$$

c. Feature Fusion Algorithms

We use five algorithms -- LDS, C4.5, SVM, NB, and K-NN -- to fuse data from features to obtain meaningful results. These are popular data mining methods which have been widely used in a wide variety of applications and four of them have been applied to detect stress. Among which two of them (LDF and K-NN) belong to conventional statistical method and the other three are normally part of the intelligent technology.

In Healey's work [13, 14], a linear discriminant function was used to classify the stress levels:

$$g_c(\hat{y}) = 2{m_c}^T K^{-1} \hat{y} - {m_c}^T K^{-1} m_c + 2\ln(\Pr[W_c]) \tag{10}$$

Here the stress class is assumed to be Gaussian distribution with m_c as the mean. The covariance K is the pooled covariance. A linear classifier is implemented by assigning each test sample to the class c for which the value of the function is the maximum. $P_r[W_c]$ is the priori probability of belonging to class c. $P_r[W_c]=1/n_k$. n_k is the numbers in class c. C4.5 algorithm, is one of the most popular and practical methods for inductive inference [22, 24], which uses information entropy as the metric to evaluate performance and uses information gain to select the nodes of the tree. See section III.B.(1) for technical details. Pioneered by Vapnik [25], SVM is a statistical learning algorithm, whose basic idea is to find an optimal hyper-plane that can maximize the margin between two groups of samples [7]. The vectors nearest to the optimal hyper-plane are called support vectors. A NB classifier is a simple probabilistic classifier based on Bayes' theorem with strong (naive) independence assumptions [23]. KNN is a kind of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification [9, 18].

IV. PERFORMANCE EVALUATION

a. The Data Set

We use the driver data set for stress detection from PHSIONET for evaluation. The data set is similar to but not as complete as the one reported in [14]. In the original work, a total of sixteen drivers participated in the experiment and for each driver eight types of raw data (Time Stamp, ECG, EMG, Foot GSR, Hand GSR, IHR, Marker, and Respiration) are acquired from the sensors that the drivers wear. However, in the released data set, only 10 drivers' data can be used, among which seven drivers' data (drivers 6, 7, 8, 10, 11, 12 and 15) are complete, which include all the sensor information as well as have clear mark identification. Three drivers' data (drivers 5, 9, and 16) are partially complete but can be used in the experiment. The remaining seven drivers' data (drivers 1, 2, 3, 4, 13, 14 and 17) do not contain all the sensor information and the mark of different driving periods is not clear. Table 3 gives a detailed illustration about the incompleteness of the driver stress data set. For example, the data set for driver 5 lacks heart rate signal during the time from 1881.2s to 2194.0s. Both the data of drivers 9 and 16 do not have clear end mark for the final rest period, so we will not use the signal data from the final rest period. The data of drivers 1 and 3 have no mark to separate different driving periods. The data of

driver 2 lacks EMG data as well as hand GSR. The driver 4 data set lacks EMG data. The driver 13 data set lacks hand GSR data. And driver 14 data set lacks heart rate data. After the cleanup, a total of 65 data are available for evaluation.

Table 3: Illustration of driver stress data set

Driver No.	Data Type Complete	Marker Clear	All periods >= 5 Minutes	Other Issues	Usefulness for our Experiment
1	Yes	No	---	No	No
2	No EMG or HGSR	No	---	No	No
3	Yes	No	---	No	No
4	No EMG	Yes	Yes	No	No
5	Yes	Yes	The first city period<5 minutes	HR is 0 during period from 1881.2s to 2194.0s.	Yes, except the first city period
9	Yes	No	The second highway period<5 minutes.	Final rest unavailable.	Yes, except the second highway period and the final rest period
13	no HGSR	Yes	Yes	No	No
14	no HR	Yes	Yes	No	No
16	Yes	Yes	The second highway period<5 minutes.	Final rest unavailable.	Yes, except the second highway period and the final rest period
17	Yes	No	---	No	No

b. Experimental Setting

We use the same feature extraction methods to extract features from each data. We then use three feature selection approaches, C4.5, PCA, and CFS, to obtain six reduced feature sets and apply five fusion methods, LDF, C4.5, SVM, NB, and K-NN, to detect stress levels. In order to better evaluate the results, we use the popular 10-fold cross validation method to prepare training and testing data sets. The results were evaluated in terms of correct rate and computational time.

The computer that our experiment ran on is a Lenovo Think Pad T410i. Its CPU is Intel T3i with 2.40GHz. Its memory is 4.0 GB. The software that we use in the experiment includes: Matlab, Weka and Eclipse. Matlab was used in data extraction and perform statistical analysis such as K-NN, PCA, ANOVA, and paired-t test. Weka was used to perform key data mining analysis such as C4.5, SVM, and Naïve Bayes. Eclipse was used to run java code calling Weka package.

V. RESULTS AND ANALYSIS

a. Results of Feature Selection

(1) Results of induction tree analysis

Figure 2 shows the decision tree for driver stress condition generated by the C4.5 algorithm. As shown, a total of five features remained as the branch node of the tree, which are: FGSR_dur, RESP0~0.1, RESP_std, HGSR_mag and HR_std. Where, FGSR_dur is the rise time duration feature from foot GSR sensor. RESP0~0.1 is the spectral power feature within frequency 0 to 0.1 Hz for Respiration sensor. RESP_std is the standard deviation feature of RESP sensor. HGSR_mag is the rise magnitude feature from hand GSR sensor. HR_std is a standard deviation feature of Heart Rate. In this case, four sensors in three different types need to be used to capture the data for detection.

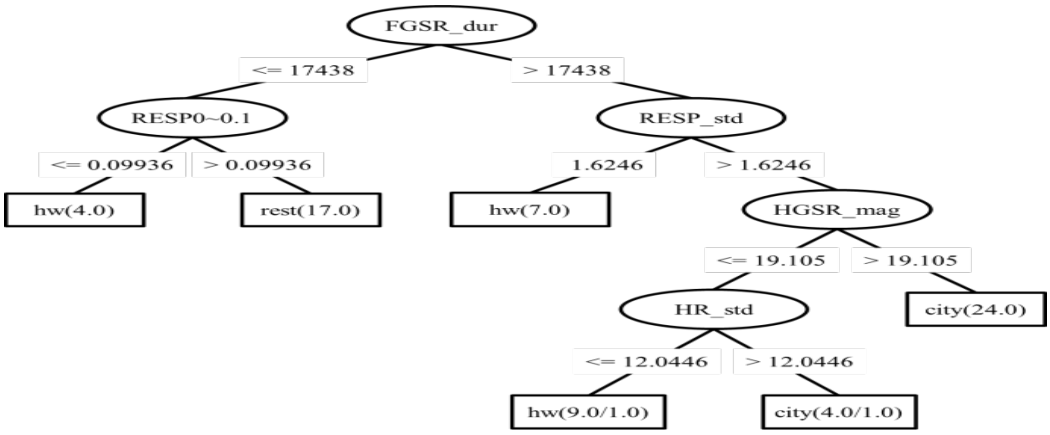

Figure 2. Decision tree generated by C4.5 algorithm

(2) Results of PCA

Figure 3 depicts the cumulative weight results from the PCA. We can see that, the first component contributes about 79% of the original data set. The first three components contribute to about 90% of the whole original data set. The first five components can contribute to almost 100% of the whole original data set. The values of w_j for the first five components are 0.786, 0.102, 0.072, 0.039, and 0.001 respectively.

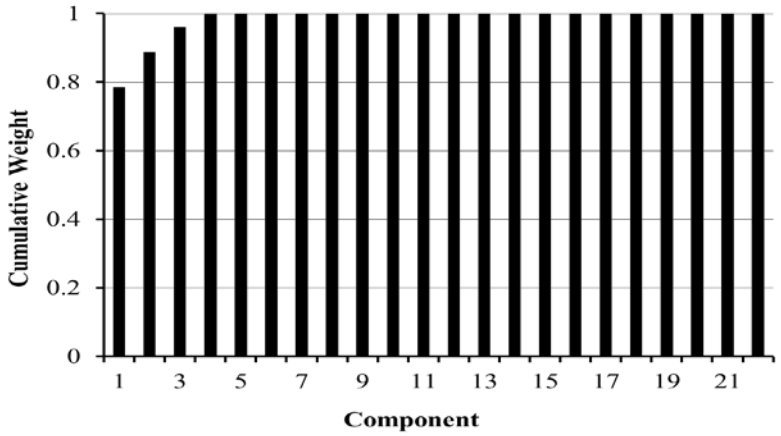

Figure 3. Results from PCA analysis

Table 4: Contribution of features for PCA

Feature (*i*)	**Comp 1**	**Comp 2**	**Comp 3**	**Comp 4**	**Comp 5**	**Contribution (*Ci*)**
Weight	0.786	0.102	0.072	0.039	0.001	
mean_EMG	0.260	-0.040	0.050	0.310	-0.010	5.72%
mean_FGSR	-0.030	0.120	-0.220	-0.240	-0.450	0.81%
mean_HGSR	-0.190	-0.020	0.000	-0.020	-0.250	2.85%
mean_HR	0.660	0.150	-0.010	0.010	-0.180	34.47%
mean_Resp	-0.230	0.090	-0.140	0.130	0.010	4.45%
std_FGSR	-0.320	0.020	-0.040	-0.100	-0.140	8.11%
std_HGSR	-0.430	-0.100	0.130	0.060	-0.090	14.77%
std_HR	0.160	-0.230	0.140	-0.220	0.070	2.88%
std_RESP	0.170	0.190	0.130	-0.060	0.090	2.78%
Resp0~0.1	-0.360	-0.090	-0.120	0.220	0.050	10.56%
Resp0.1~0.2	0.370	0.060	0.110	-0.170	-0.010	11.00%
Resp0.2~0.3	0.280	0.120	0.120	-0.300	-0.180	6.77%
Resp0.3~0.4	0.220	0.180	0.070	-0.370	-0.140	4.71%
Fgsr_freq	0.770	0.370	-0.170	-0.140	-0.240	**48.29%**
Fgsr_Mag	0.740	0.006	0.390	0.350	-0.360	44.63%
Fgsr_Dur	0.940	-0.320	0.030	-0.060	0.000	**70.52%**
Fgsr_Area	0.750	0.470	0.170	-0.280	-0.080	**46.98%**
Hgsr_freq	0.770	-0.070	-0.100	0.050	-0.620	46.77%
Hgsr_Mag	0.490	-0.020	0.670	0.560	0.000	23.33%
Hgsr_Dur	0.930	0.260	-0.240	0.120	0.000	**69.14%**
Hgsr_Area	0.760	0.460	0.370	-0.270	0.000	**48.83%**
Ihr_LR	0.350	0.060	0.040	-0.200	-0.330	9.84%

From Table 4, we can get the correlation of each feature and the first 5 components. According to equation (8), we can then calculate the contribution of each feature and acquire the top 5 features with the maximum contribution values (highlighted in bold face).

(3) Results of CFS

Figure 4 shows the most significant correlated features and the correlation coefficient between the features in our study. Where, p-value < 1.0*E-10 and correlation coefficient >= 0.70. As shown, there are three groups of features (e.g., those extracted from RESP, FGSR, and HGSR), in which, within each group, each feature has high correlation with others. In this case, maybe only one feature from each subclass should be considered in the evaluation.

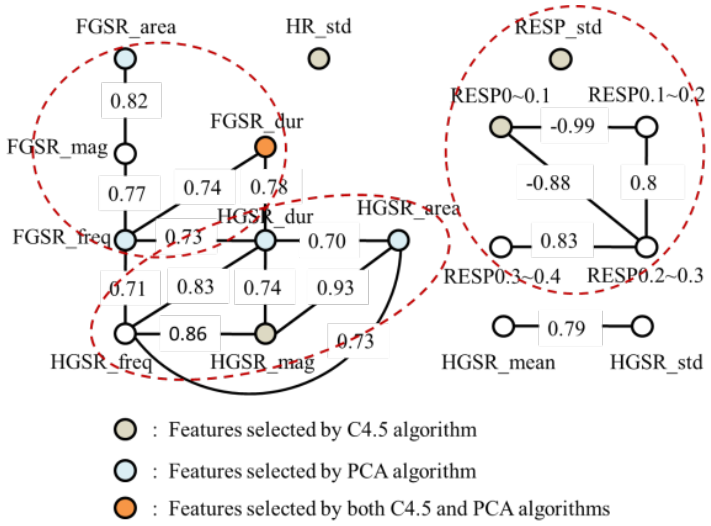

Figure 4. Feature correlation schematic diagram

(4) Comparison of feature selection methods

Table 5 presents the initial feature selection results from this study. Both C4.5 and PCA approaches select five features while the features are different except FGSR_dur feature. MIX4 is the merge of results of C4.5 and of PCA. MIX1, MIX2, and MIX3 are the subsets of MIX4, which are generated by reducing some features from MIX4 using correlation-based feature selection approach. For instance, from Figure 7, we can see that, HGSR_dur has high correlation with FGSR_freq, FGSR_dur, HGSR_mag and HGSR_area. So, if HGSR_dur is selected, the other four features should be removed to prevent from high internal correlation; thus, five features are included in MIX1 from MIX4. Similarly, if FGSR_dur and HGSR_area are selected, FGSR_freq, HGSR_mag and HGSR_dur should be removed; thus, six features are included in MIX2 from MIX4. Seven features of MIX3 can also be derived from MIX4.

Table 5: Summary of feature selection results

Method	No. of Feature	Features Selected	No. of Sensors
C4.5	5	FGSR_dur, RESP0~0.1, RESP_std, HGSR_mag, HR_std	4
PCA	5	FGSR_dur, HGSR_dur, HGSR_area, FGSR_area, FGSR_freq	2
MIX1	5	RESP_std, HR_std, RESP0~0.1,FGSR_area, HGSR_dur	4
MIX2	6	RESP_std, HR_std, RESP0~0.1, FGSR_area, FGSR_dur, HGSR_area	4
MIX3	7	RESP_std, HR_std, RESP0~0.1, FGSR_dur, HGSR_mag, HGSR_dur	4
MIX4	9	RESP_std, HR_std, RESP0~0.1, FGSR_dur, FGSR_area, FGSR_freq, HGSR_area, HGSR_mag, HGSR_dur	4

By carefully examining the types of sensors used, we notice that the data features selected by C4.5 Induction Tree need three different types and four sensors, ECG, respiration, and hands and foot galvanic skin response sensors. The data features selected by PCA need to use only one type of sensor, galvanic skin response sensor, but place at both hands and foot. Similar to C4.5 feature set, other feature sets need three different types and four sensors. Considering the importance of user-friendliness, we would select the feature set suggested by PCA method.

b. Results of Stress Detection

We randomly generate 10 folds from the whole data sets and in turn randomly pick one fold as the test set and the remaining folds as the training set. Meanwhile, we repeat the above procedure for six times, resulting in 60 test results. The average of these results is summarized in Table 6.

We can see that, the seven features selected by the method of MIX3 can result in the best average correct rate for all the fusion algorithms, which reaches 75.74% in average and using the C4.5 fusion method results in second best correct rate, 84.33% in average. The second best average correct rate comes from the nine features selected by method MIX4. The five features selected by C4.5 method can result in the best individual correct rate when using C4.5 algorithm as the fusion method, which is 85.46%. The five features selected by PCA method can result in 70.83% accuracy as the average performance, and its highest value is 78.94% when using SVM as the fusion method. On the other hand, the average correct rate for the full 22 features can only reach 70.99% when using all fusion algorithms and the best correct rate can only reach to78.05% when using Naïve Bayes as the fusion method.

Table 6: Results from 10-fold cross validation method

Feature Set	Statistics	Fusion Algorithm					Average (%)
		LDF	C4.5	SVM	NB	KNN	
All/22*	Mean	68.13	65.46	74.50	78.05	68.80	70.99
	Std	15.90	8.40	6.80	6.76	6.81	8.93
C4.5/5	Mean	72.06	**85.46**	71.97	77.51	70.22	75.44
	Std	18.25	5.77	6.20	6.65	6.67	8.71
PCA/5	Mean	67.50	62.34	78.94	70.44	74.92	70.83
	Std	17.79	7.62	5.69	7.95	5.63	8.94
MIX1/5	Mean	63.17	76.28	66.45	71.11	61.16	67.63
	Std	16.00	8.01	5.49	8.33	7.54	9.07
MIX2/6	Mean	69.64	79.58	71.29	80.44	74.25	75.04
	Std	18.56	5.83	7.70	5.60	7.54	9.05
MIX3/7	Mean	72.14	**84.33**	71.84	74.60	75.78	**75.74**
	Std	16.56	5.96	9.51	7.27	7.76	9.41
MIX4/9	Mean	68.06	80.31	76.10	76,96	78.00	**75.62**
	Std	16.03	6.29	7.05	7.36	5.07	8.36
Average	Mean	68.67	**76.25**	73.01	**75.36**	71.88	
	Std	17.01	6.84	6.92	7.13	6.72	

* Feature Set/Number of features

In terms of fusion methods, C4.5 induction tree method (76.25%) outperforms other fusion methods, followed by NB classifier (75.36%). In general, the conventional fusion methods such as LDF and K-NN do not performed as well as the intelligent algorithms/systems for stress detection using the benchmark driver dataset. Again, it is obvious that feature selection can lead to more or less accurate results in prediction. The results are highly dependent on the fusion method used too. Thus, it is important to evaluate and select proper feature selection method and fusion algorithm to improve performance.

c. Statistical Analysis

We further use the Analysis of Variance (ANOVA) to analyze the statistical significance of the effects and their interactions among feature selection, and fusion methods. We also use pairedt-test to evaluate the relative performance of selected pair of feature selection methods as well as fusion algorithms.

(1) Factor effect analysis

We first examine the statistical effect of feature selection and fusion algorithms and their interactions. We compare their relative performance using all fusion algorithms: LDF, C4.5, SVM, NB and K-NN algorithms. We formulate three null hypothesizes for ANOVA analysis:

Hypothesis H1: The accuracy rate is the same for the full and selected reduced feature sets.

Hypothesis H2: The accuracy rate is the same for all the fusion algorithms.

Hypothesis H3: The accuracy rate is the same for the interactions between different feature sets and different fusion algorithms.

Table 7 shows the results of the ANOVA analysis. As shown, the p-value for the fusion algorithm effect is 0 confirming that the performance of the fusion algorithms in terms of the correct rate is statistically different. Similarly, the p-value for the feature effect is 0, indicating that the correct rate varies from one feature set to another and is statistically significant. The results indicate that fusion algorithm has higher impact than the feature selection. The results also support that there is significant interactions between feature selection and fusion algorithm. Therefore, it is very important to select proper fusion method to match with feature selection algorithm. Because of the significance of interaction effect, we need to conduct further analysis to isolate the possible interactions.

Table 7: N-way ANOVA results

Source	Sum Sq.	DF	Mean Sq	F	Sig.
Fusion Algorithm	15602.7	4	3900.66	40.16	0
Feature Selection method	19004.3	6	3167.38	32.61	0
FusionAlgorithm*Feature Algorithm	36004.3	24	1500.18	15.45	0
Error	200564.1	2065	97.13		
Total	271175.2	2099			

*Sig.: significant level (Prob. > F); DF: degree of freedom

(2) Impact of feature selection

We compare the five features selected by using C4.5 metric with the 22 features to see whether the reduced feature set can improve correct rate as well as reduce computational time. We use LDF and C4.5 algorithm as the fusion methods, since the former one has been proved to be a good classifier in Healey and Picard's work [13, 14] and the later one showed the best performance in our evaluation. We formulate five hypotheses to statistically test the differences:

Hypothesis H4: If LDF fusion method is used, there is no difference in correct rate between using the five features selected by C4.5 induction tree method and using the full 22 features.

Hypothesis H5: If LDF fusion method is used, there is no difference in computation time between using the five features selected by C4.5 induction tree method and using the full 22 features.

Hypothesis H6: If C4.5 fusion method is used, there is no difference in correct rate between using the five features selected by C4.5 induction tree method and using the full 22 features.

Hypothesis H7: If C4.5 fusion method is used, there is no difference in computation time between using the five features selected by C4.5 induction tree method and using the full 22 features.

Hypothesis H8: There is no difference in correct rate between using the five features selected by C4.5 induction tree method with C4.5 fusion algorithm and using the full 22 features with LDF fusion algorithm.

Table 8 summarizes the results from the paired t-test for the five features selected by C4.5 method and the 22 full features. As can be seen, the five features selected by C4.5 leads to 72.06% correct rate and 22 features obtains 68.13% correct rate using LDF algorithm. The p-value is 0.21, which is much higher than the significant value of 0.05. The H value is 0. This indicates a failure to reject the null hypothesis H4 at the 5% significant level, which means that in the perspective of statistics, by using LDF algorithm, there is no difference in correct rate for the five features selected by C4.5 and the original full 22 features. The time collapsed is not significantly different for the two feature sets as well when using LDF algorithm, since the p-value is 0.06, which is higher than significant level 0.05. So, Hypothesis H5 cannot be rejected.

On the other hand, using the C4.5 fusion algorithm, the correct rate for the five feature set is 85.46% and the correct rate for the 22 features is only 65.46%. This indicates that the five features selected by C4.5 method can lead to higher correct rate than 22 features when using C4.5 as the fusion algorithm. The difference in correct rate is statistically significant. Hence, hypothesis H6 is rejected. So, in the perspective of statistics, the five features selected by C4.5 method can result in higher correct rate than the full 22 features when using C4.5 as the fusion

method. However, the time collapsed is not significantly different for the two feature sets when using C4.5 algorithm, since the p-value is 0.52, which is much higher than significant level 0.05. So, Hypothesis H7 cannot be rejected.

Table 8: Paired t-test of 22 features and five features selected by C4.5 method

Fusion Algorithm	**LDF**		**C4.5**	
Features	**5(C4.5)**	**22(Full)**	**5(C4.5)**	**22(Full)**
Correct Rate (%)				
Mean	72.06	68.13	85.46	65.46
Variance	18.25	15.90	5.77	8.40
Observation	60		60	
H	0		1	
Df	115.83		104.53	
t-Statistics	1.26		15.20	
Significant	0.05		0.05	
P value	0.21		0	
Time elapsed (nanosecond)				
Mean	1.42e5	1.76e5	8.85e6	8.60e6
Variance	9.66e4	1.02e5	1.76e6	2.36e6
Observation	60		60	
H	0		0	
Df	117.61		109.17	
t-Statistics	-1.88		0.65	
Significant	0.05		0.05	
P value	0.06		0.52	

The results from Table 9 show that, the mean correct rate of the five features selected by C4.5 induction tree method using decision tree C4.5 fusion algorithm is 85.46%, which is much higher than 68.13% from the 22 features using LDF algorithm. The p-value is 0, which is much lower than significant level 0.05. So, H8 should be rejected. This indicates that the five features selected by C4.5 induction tree method when using decision tree C4.5 algorithm as fusion algorithm can result in better correct rate than the 22 features when using LDF algorithm. Here, we did not compare the computational time between the two feature sets because they were executed on different platforms. LDF is performed in Matlab, while C4.5 algorithm is run in Eclipse calling Weka packages.

Table 9: Paired t-test: Using benchmark and best results

	5 features &C4.5	**22 features & LDF**
Correct Rate (%)		
Mean	85.46	68.13
Variance	5.77	15.90
Observation	60	
H	1	
Df	74.27	
t-Statistics	7.93	
Significant	0.05	
P value	0	

VI. CONCLUSION AND DISCUSSION

Effective recognition of stress plays an important role for people to manage their health. The processing of information captured from psychological sensors that people wear provides an efficient way to detect people's stress. Even though some stress detection prototypes have been developed and some features have been extracted from the raw data of bio-sensors. How to select the most significant features and how to fuse the features selected to predict the stress level or pattern remain the key issues unanswered. In this paper, we examine and compare feature selection and information fusion algorithms for stress identification.

Our study results show that, by using the proposed feature selection methods, the accuracy performance of the five classifiers has been greatly improved. Besides the improvement in performance, feature selection is also very important in the real use of physiological sensors in health care. It is not realistic to ask people to wear many sensors to acquire all kinds of physiological data, which will make people tired of them. By discarding some less important features, some sensors can be successfully removed. For example, people will wear five physiological sensors to recognize their stress level if using the original 22 features. People just need to wear two sensors if the features are selected by using the proposed PCA method.

Among the three proposed feature selection heuristics, the C4.5 induction tree can result in the highest correct rate among all the experiment results. Under this strategy, five key features can be induced from a total of 22 features, and the correct rate can reach 85.46% when using the decision tree C4.5 as the fusion algorithm. On the other hand, the feature selection metric based on the PCA can lead to the least number of sensors used. Only two sensors are needed to extract

five features if we use PCA-based feature selection method. Under this strategy, the correct rate can reach 78.94% when using SVM as the fusion algorithm.

Among the five feature fusion algorithms, the C4.5 appeared to perform best using five features selected by the C4.5 algorithm. SVM can acquire the highest correct rate when fusing the five features selected by the PCA based approach. SVM and Naïve Bayes algorithms have a good correct rate on the 22 feature set. The fewer features selected by selection approaches can use equal or less time than the whole 22 features in the prediction, considering that we did not include the time spent in extracting the capture data. Clearly, the total time saved from reducing the number of features could be significant if all time elements are included.

The contributions of our study are threefold. First, we perform a thorough investigation of the driver stress database and present a detailed available information survey on the different period times as well as different sensors for every driver. The data cleaning work can help save time for researchers if the same dataset is used for benchmark. Secondly, we propose a PCA based feature selection method from different perspectives and statistically compare the effect of three different selected feature sets on five feature fusion algorithms. We illustrate the importance of feature selection and the selection needs to be carefully evaluated and to combine with the use of proper fusion algorithms. Finally, we identify two special 5-feature sets as well as their suitable classification algorithms for stress identification, which will be very helpful for real time stress recognition.

Our study does have some limitations. First, the number of data sets is only 65 in our experiment. Even though it is a small sample from the statistical perspective, it does meet the requirement of the least number of statistical analysis and machine learning. So, the experimental result based on the data set is still credible. Since there is no public data set about stress detection available now and little work about the feature selection in stress detection has been done, the conclusions of our experiment are still valuable. Secondly, the C4.5 algorithm serves not only as a feature selection approach but also as a good classification approach. But, it works on different data feature sets for different usages as well. The C4.5 algorithm is an information gain based or entropy based method in essence, so we can use it to select those features which can get biggest information gain. Thirdly, the dataset was established assuming that the stress level during the whole driving period in highway or city were the same and did not consider personalization effect. Some in-depth analysis on individual driver effect may need to be explored in the future.

Also, dynamic change point algorithm can also be explored to identify stress level variation during each driving condition.

ACKNOWLEDGEMENT

This research is supported by the National Key Technology R&D Program in China (No: 2011BAH14B02 and No: 2012BAH18B04) and the National Science and Technology Major Project in China (No: 2012ZX03002022).

REFERENCES

[1] H. Abdi, L.J. Williams, “Principal Components Analysis”, Wiley Interdisciplinary Reviews: Computational Statistics, Vol. 2, No. 4, 2010, pp. 433–459.

[2] A. Akbas, “Evaluation of the Physiological Data Indicating the Dynamic Stress Level of Drivers”, Scientific Research and Essays, Vol. 6, No. 2, 2011, pp. 430-439.

[3] APA (American Psychological Association), “Stress in America: Our Health at Risk”, Accessed on June 2012. *URL:* http://www.apa.org/news/press/releases/stress/index.aspx

[4] F. Angus, J. Zhai, “Front-end Analog Pre-processing for Real Time Psychophysiological Stress Measurements”, Proceedings of the 9th World Multi-Conference on Systematics, Cybernetics and Informatics (WMSCI05), 2005, pp. 218-221.

[5] J. Bakker, M. Pechenizkiy, N. Sidorava, “What’s Your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data”, Proceedings of the11th IEEE International Conference on Data Mining Workshops, 2011, pp. 573-580.

[6] L. Bergman, P. Corabian, C. Harstall, “Effectiveness of Organisational Interventions for the Prevention of Occupational Stress”, Alberta: Institute of Health Economics, Accessed on June 2012. *URL:* http://www.ihe.ca/publications/library/2009/effectiveness-of-organizational-interventions-for-the-prevention-of-workplace-stress/

[7] A.-M. Cretu, and P. Payeur, “Biologically-inspired Visual Attention Features for a Vehicle Classification Task”, The International Journal on Smart Sensing and Intelligent Systems, Vol. 4, No. 3, 2011, pp. 402-423.

[8] J. R.T. Davidson, S.W. Book, “Assessment of a New Self-Rating Scale for Post-traumatic

Stress Disorder", Psychological Medicine, Vol. 27, No. 1, 1997, pp.153-160.

[9] R. Duda, P. Hart., D. Stork, "Pattern Classification", (2nd Ed.).Wiley Inter-science, 2001

[10] FlexComp, "ProComp Software Version 1.41 User's Manual", Thought Technology Ltd., Montreal, QC, Canada, 1994.

[11] M. Hall, "Correlation Based Feature Selection for Machine Learning", Doctoral Dissertation, University of Waikato, 1999.

[12] S. Haykin, "Neural Networks: A Comprehensive Foundation (2nd Ed.)", Englewood Cliffs, NJ: Prentice-Hall, 1998.

[13] J.A. Healey, "Wearable and Automotive Systems for Affect Recognition from Physiology", Doctoral Dissertation, Massachusetts Institute of Technology, MA, 2000.

[14] J.A. Healy, R.W. Picard, "Detecting Stress During Real-World Driving Tasks Using Physiological Sensors", IEEE Transaction on Intelligent Transportation System, Vol. 6, No. 2, 2005, pp.156-166.

[15] E. Jovanov, A. O'Donnell Lords, D. Raskovic, P.G. Cox, R. Adhami, F. Andrasik, "Stress Monitoring Using a Distributed Wireless Intelligent Sensor System", IEEE Engineering in Medicine and Biology Magazine, Vol. 22, No. 3, 2003, pp. 49-55.

[16] A. Kaklauskas, E.K. Zavadskas, V. Pruskus, A. Vlasenko, L. Bartkiene, "Recommended Biometric Stress Management System", Expert Systems with Applications, Vol. 38, 2011, pp.14011-14025.

[17] A. Malhi, R. Gao, "Feature Selection for Defect Classification in Machine Condition Monitoring", 20th IEEE Instrumentation Measurement Technology Conf., Vol. 1, 2003, Vail, CO, pp. 36-41.

[18] A. Moosavian, H. Ahmadi, A. Tabatabaeefar, B. Sakhaei, "An Appropriate Procedure for Detection of Journal-Bearing Fault Using Power Spectral Density, K-Nearest Neighbor and Support Vector Machine", The International Journal on Smart Sensing and Intelligent Systems，Vol.5, No. 3, 2012, pp.685-700.

[19] M. Nako, "Work-related Stress and Psychosomatic Medicine", BioPsycho Social Medicine, Vol. 4, No. 4, 2010, Doi:10.1186/1751-0759-4-4.

[20] Office for National Statistics, Social and Vital Statistics Division and Northern Ireland Statistics and Research Agency. Central Survey Unit, 2010. "Labour Force Survey, 1975-2010", Colchester, Essex: UK Data Archive. *URL:*http://www.esds.ac.uk/government/lfs/

[21] PHYSIONET, "Stress Recognition in Automobile Drivers *(drivedb)*", Accessed on June 2012. URL: http://physionet.org/cgi-bin/atm/ATM/.

[22] K. Polat, S. Güneş, "A Novel Hybrid Intelligent Method Based on C4.5 Decision Tree Classifier and One-against-all Approach for Multi-Class Classification Problems", Expert Systems with Applications, Vol. 36, 2009, pp. 1587-1592.

[23] I. Rish, "An Empirical Study of the Naive Bayes Classifier", Proceedings of IJCAI-01 workshop on Empirical Methods in AI, 2001, pp. 41-46, Sicily, Italy.

[24] S. Ruggieri, "Efficient C4.5", IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 2, 2002, pp. 438-444.

[25] V. Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag, New York, NY, USA. 1995. ISBN: 0-387-94559-8.

[26] D. Watson, J.W. Pennebaker, "Health Complaints, Stress, and Distress: Exploring the Central Role of Negative Affectivity", Psychological Review, Vol. 96, No. 2, 1989, pp. 234-254.

[27] S. Wold, "Principal Component Analysis", Chemometrics and Intelligent Laboratory Systems, Vol. 2, No. 1-3, 1987, pp. 37-52.

[28] K.Y. Yeung, W.L. Ruzzo, "Principal Component Analysis for Clustering Gene Expression Data", *Bioinformatics*, Vol. 17, No. 9, 2001, pp. 763-774.

[29] L. Yu, H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", Proceedings of the 20th International Conference on Machine Learning (ICML-2003), Washington, DC, Vol. 3, 2003, pp. 856-863.

[30] J. Zhai, A. Barreto, "Stress Detection in Computer Users Through Non-Invasive Monitoring of Physiological Signals", Biomedical Science Instrumentation, Vol. 42, 2006, pp. 495-500.

[31] L. Zhang, T. Tamminedi, A. Ganguli, G. Yosiphon, J. Yadegar, "Hierarchical Multiple Sensor Fusion Using Structurally Learned Bayesian Network", Proceedings of Wireless Health, 2010, pp. 174-183.