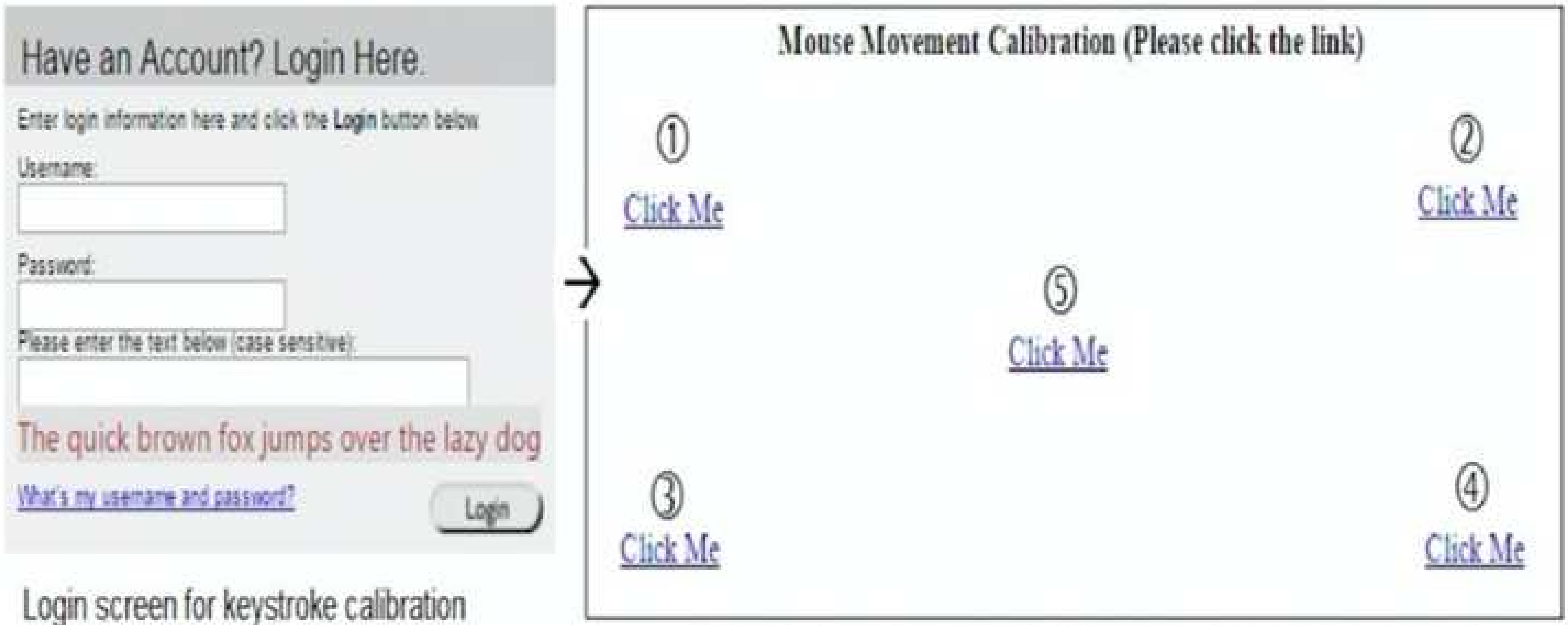
International Journal of Human-Computer Interaction

Continuous Stress Monitoring under Varied Demands Using Unobtrusive Devices

--Manuscript Draft--

Manuscript Number:	IJHC-D-18-00462R1
Full Title:	Continuous Stress Monitoring under Varied Demands Using Unobtrusive Devices
Article Type:	Original Research
Section/Category:	Basic Science Section
Corresponding Author:	Yee Mei Lim, Ph.D. Tunku Abdul Rahman University College Kuala Lumpur, Kuala Lumpur MALAYSIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Tunku Abdul Rahman University College
Corresponding Author's Secondary Institution:	
First Author:	Yee Mei Lim, Ph.D.
First Author Secondary Information:	
Order of Authors:	Yee Mei Lim, Ph.D.
	Aladdin Ayesh, Ph.D.
	Martin Stacey, Ph.D.
Order of Authors Secondary Information:	
Abstract:	This research aims to identify a feasible model to predict a learner’s stress in an online learning platform. It is desirable to produce a cost-effective, unobtrusive and objective method to measure a learner’s emotions. The few signals produced by mouse and keyboard could enable such solution to measure real world individual’s affective states. It is also important to ensure that the measurement can be applied regardless the type of task carried out by the user. This preliminary research proposes a stress classification method using mouse and keystroke dynamics to classify the stress levels of 190 university students when performing three different e-learning activities. The results show that the stress measurement based on mouse and keystroke dynamics is consistent with the stress measurement according to the changes of duration spent between two consecutive questions. The feedforward back-propagation neural network achieves the best performance in the classification.

Fig. 1



Have an Account? Login Here.

Enter login information here and click the **Login** button below

Username:

Password:

Please enter the text below (case sensitive):

The quick brown fox jumps over the lazy dog

Mouse Movement Calibration (Please click the link)

Login screen for keystroke calibration

Mouse movement calibration is displayed after the login screen. The link is shown one by one according to the order, and the learner has to click on it

Fig. 2

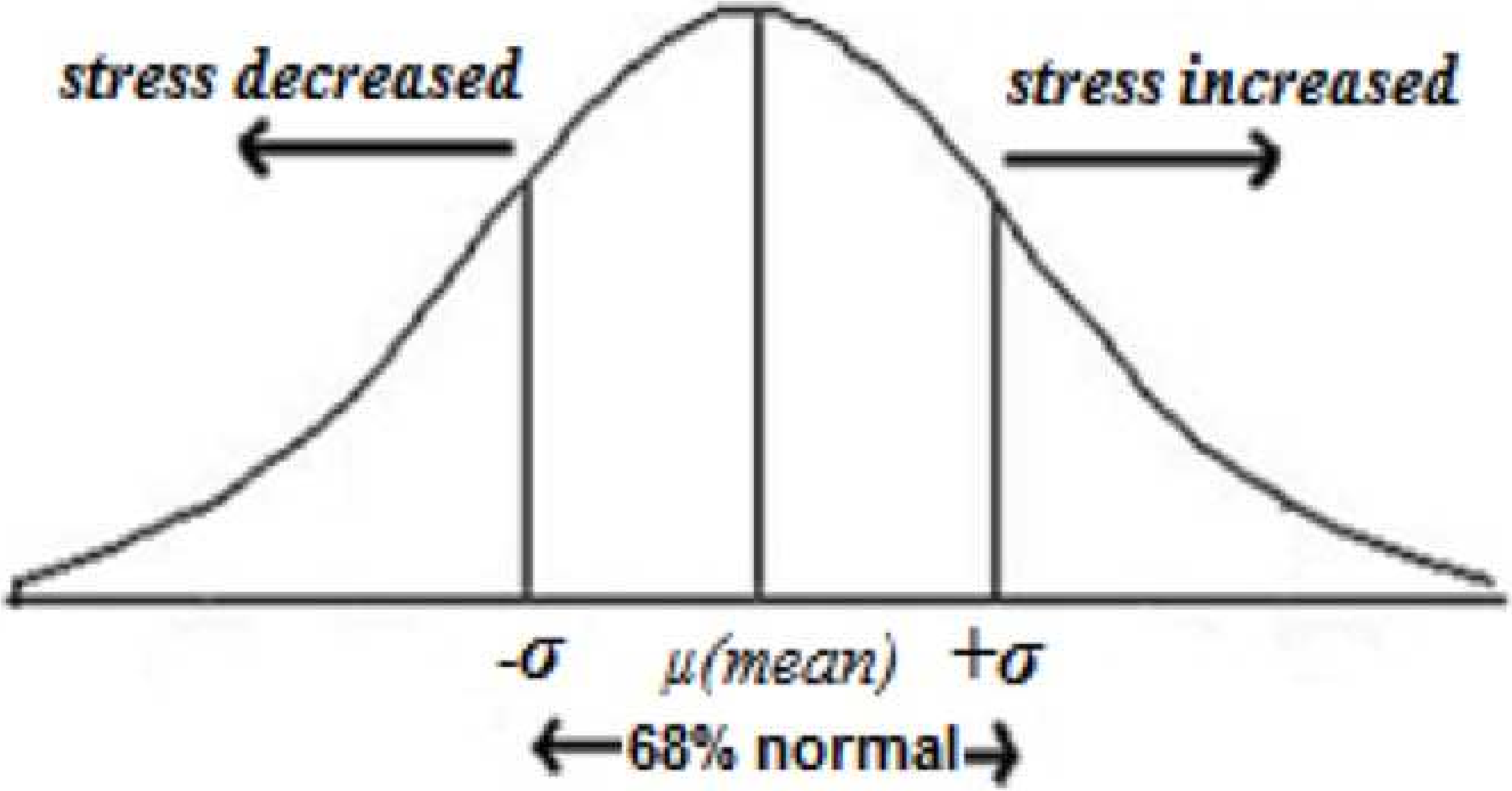


Fig. 3

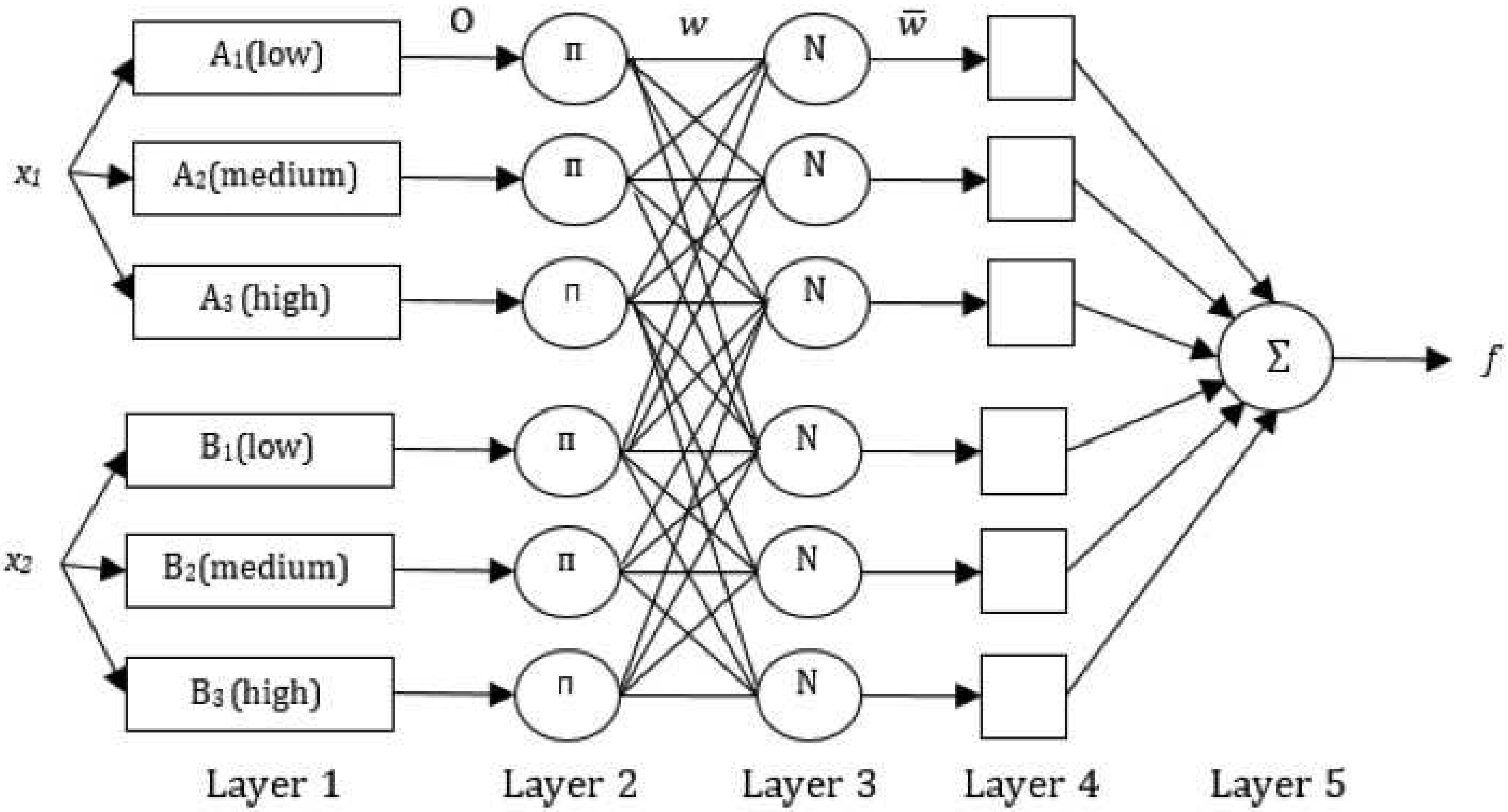

x1
x2
A1(low)
A2(medium)
A3 (high)
B1(low)
B2(medium)
B3 (high)
O
Π
w
N
w̄
Σ
f
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5

Fig. 4

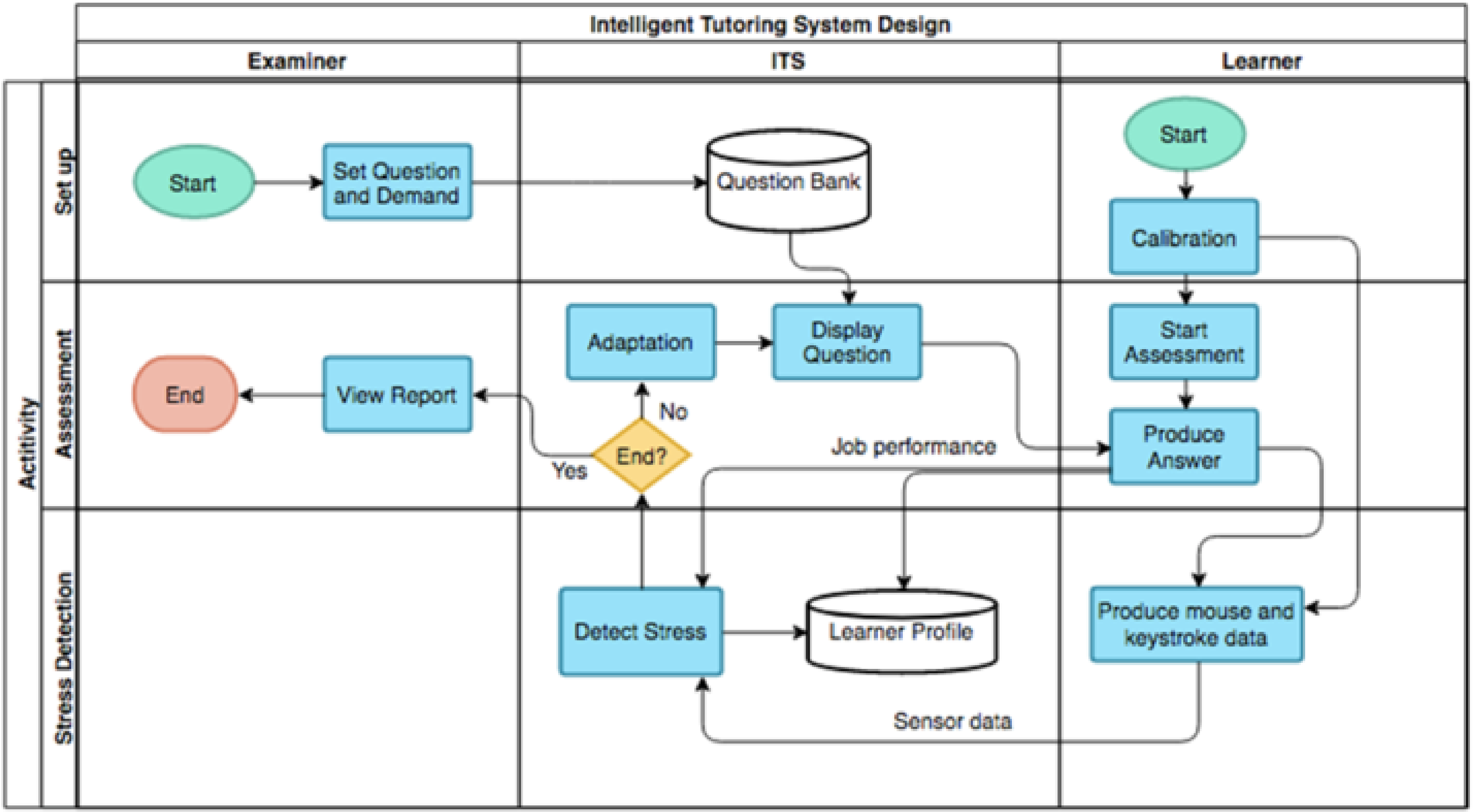
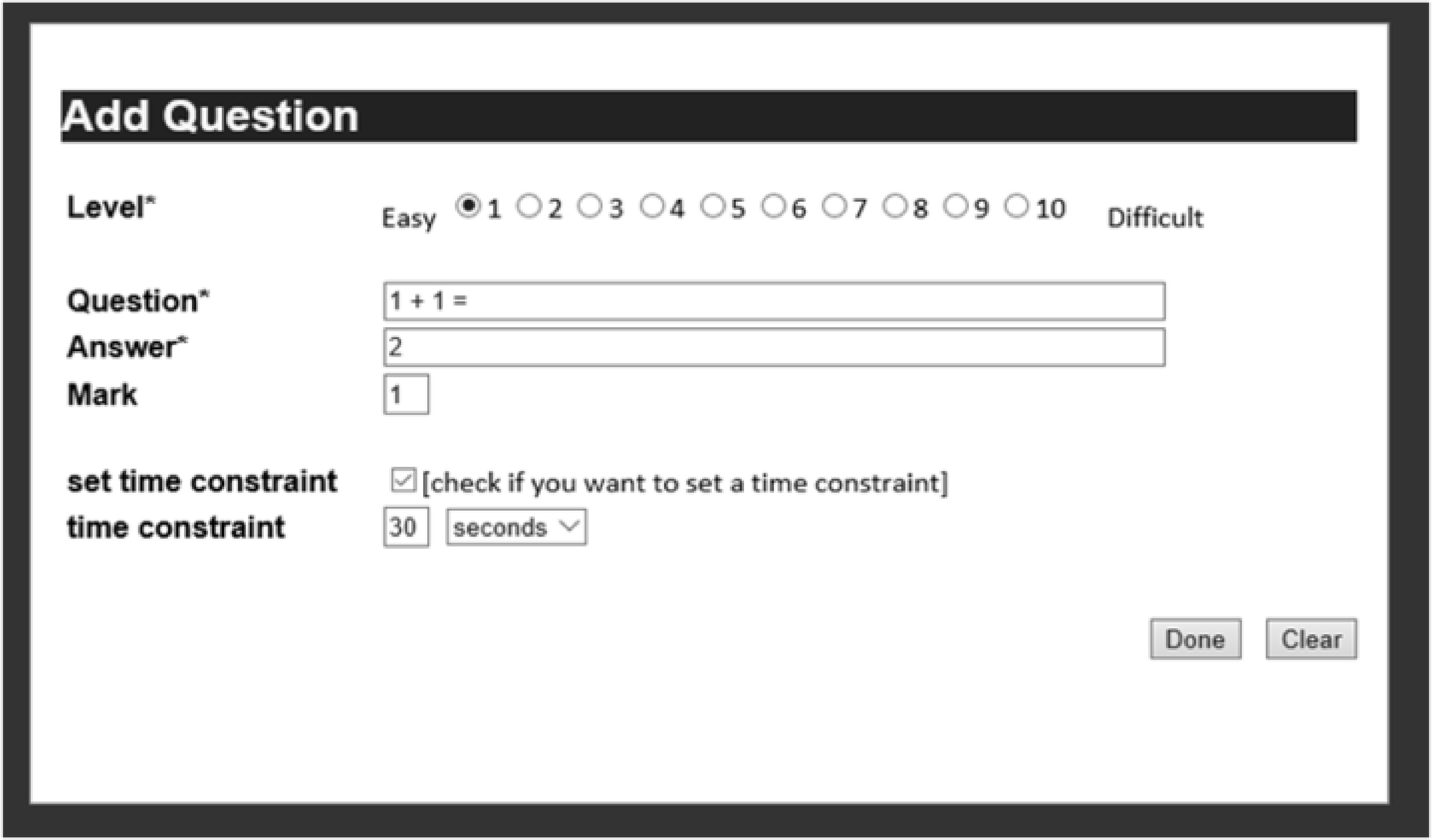

Intelligent Tutoring System Design
Examiner
ITS
Learner
Actitivity
Set up
Assessment
Stress Detection
Start
Set Question and Demand
Question Bank
Start
Calibration
Adaptation
Display Question
Start Assessment
End
View Report
No
End?
Yes
Produce Answer
Job performance
Detect Stress
Learner Profile
Produce mouse and keystroke data
Sensor data

Fig. 5



Add Question

Level* Easy ● 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 Difficult

Question*	1 + 1 =
Answer*	2
Mark	1
set time constraint	☑ [check if you want to set a time constraint]
time constraint	30 seconds

Continuous Stress Monitoring under Varied Demands Using Unobtrusive Devices

Yee Mei Lim, Tunku Abdul Rahman UniversityCollege [1]

Aladdin Ayesh, De MontfortUniversity[2]

Martin Stacey, De MontfortUniversity[3]

[1] ymlim@tarc.edu.my. ORCID: https://orcid.org/0000-0003-2164-7287. Corresponding author

[2] aayesh@dmu.ac.uk. ORCID: https://orcid.org/0000-0002-5883-6113

[3] mstacey@dmu.ac.uk. ORCID: https://orcid.org/0000-0002-1194-5507

Continuous Stress Monitoring under Varied Demands Using Unobtrusive Devices

This research aims to identify a feasible model to predict a learner's stress in an online learning platform. It is desirable to produce a cost-effective, unobtrusive and objective method to measure a learner's emotions. The few signals produced by mouse and keyboard could enable such solutionto measure real world individual's affective states. It is also important to ensure that the measurement can be applied regardless the type of task carried out by the user. This preliminary research proposes a stress classification method using mouse and keystroke dynamics to classify the stress levels of 190 university students when performing three different e-learning activities. The results show that the stress measurement based on mouse and keystroke dynamics is consistent with the stress measurement according to the changes of duration spent between two consecutive questions. The feedforward back-propagation neural network achieves the best performance in the classification.

Keywords: *Stress monitoring, mouse dynamics, keystroke dynamics, job duration, affective computing*

1 INTRODUCTION

In an online learning platform, it is crucial for teachers to understand a learner's emotion and engagement in the learning content. It would be useful if such platform can help teachers to identify the factors that cause negative emotion and poor learning behavior. Therefore, it is not enough to merely provide number facts, such as duration spent and scores achieved for an assessment, which most of the existing e-learning systems offer. An adaptive learning platform should be able to determine the factor that generates negative emotion such as stress, to adapt personalized learning materials to enhance students' engagement, and to help teachers to redesign the necessary learning process and materials. To develop such online personalized adaptive learning system, it is important to produce a construct that is able to quantify cognitive load and emotion, using a low cost, non-invasive, computational feasible and fully automated solution. Some research examines the potential of using mouse or keystroke dynamics. Mouse and keystroke dynamics were initially studied as potential biometric authentication methods but they also demonstrated great feasibility in emotion detection or mental states over the past decade (Crawford, 2010; Landowska et al., 2014; Salmeron-Majadas, Baker, Santos, & Boticario, 2018; Tsoulouhas, Georgiou, & Karakos, 2011; Vea & Rodrigo, 2017; Zimmermann, Guttormsen, Danuser, & Gomez, 2003). As standard and cost-effective input devices for a computer, keyboard and mouse enable a completely unobtrusive way of data collection as no special hardware or setup is needed, easily availabile and can be captured easily during user's usual computer activities. Besides, small amount but diversified features can be extracted. This means that they can be easily processed online in order to sense learner's states in real time continuously, using multi-modal approach that generally hold potential for increased performance, without greatly affecting the server or computer performances at the same time. They also safeguard privacy as it is not necessary to determine what the users are doing or typing, but still can effectively determine their inner state or individual behavioral patterns(Carneiro & Novais, 2017).

The main challenge of the implementation of mouse/keystroke-based analysis lies within the reliability of the stress measurement. It is ideal to produce a reliable stress measurement that is generic or context-independent, which can monitor stress for any task in an online learning environment. Each user has individual variations in how they interact with user interface when performing a task using a mouse and/or a keyboard. To enable continuous stress monitoring, the difference of task duration and the difference of mouse/keystroke behaviors between 2 consecutive questions being answered are computed. Each individual's mouse and keystroke data would be compared against the time duration of completing a task, to get a sense of generally increasing, decreasing and stable/normal stress level.

S_{TD} –stress that is measured based on time duration.

The measurement of stress $S_{B(Sensor)}$ is also done based on sensor data, which sensor data are referring to mouse and keyboard dynamics. We categorize $S_{B(Sensor)}$ as follows:

$S_{B(M)}$– stress that is measured based on only mouse dynamics;

$S_{B(K)}$ – stress that is measured based on only keystroke dynamics; and

$S_{B(M,K)}$– stress that is measured based on the unification of both mouse and keystroke dynamics.

For instance, while performing a typing task that mouse dynamics data are absent for a period of time, as such only $S_{B(K)}$ is taken into consideration during that period.

We invited 190 undergraduate students to conduct three different e-learning activities. The first experiment involves 64 menu search tasks. The second experiment requires them to answer 10 mental arithmetic questions that stimulate cognitive stress. Lastly the third experiment is to type6 pre-defined text with varied lengths in 2 different languages, to exemplify familiar and unfamiliar tasks. As the search task does not require keyboard input, hence only $S_{B(M)}$ is computed. To find an objective way to validate our proposed method, we compare the estimated $S_{B(Sensor)}$ against the stress measurement S_{TD} that is computed based on time duration. Time duration measurement is utilized as significant relationships between time pressure, stress, job performance, and decision making are found(Peters, O'Connor, Pooyan, & Quick, 1984; Svenson & Maule, 1993), and humans are usually more stressed over time(Lim, Ayesh, & Stacey, 2014c; Szalma et al., 2004). Hence, a simple assumption is made in this research: when a task demand is elevated, the time spent on the task is expected to increase. If the increment rate of the time spent is within the anticipated range, then the behavioral outcome of the user is deemed stable or normal. However, if the task requires much longer time than expected, then the task could be more challenging than what the examiner imagined. Vice versa, if the task takes significantly much shorter time than expected, then the question might be either too easy, or the student may demonstrate anomalous behavior, e.g. does not answer the question seriously, or simply does not put much attention to it.

2 EMOTION AND STRESS MEASUREMENT

Existing research related to affective learning adopted emotion defined by psychological research, e.g. the four quadrants of learning emotions as proposed by Kort et al (Kort, Reilly, & Picard, 2001), the Positive and Negative Affect Schedule (PANAS) scale by Watson et al (Watson, Clark, & Tellegen, 1988), or Russell's Circumplex Model of Affect (Russel, 1980). It may be important to have better understanding of granularity of emotions that could affect learning performance. However, enabling automated detection of rich granularity of emotions is extremely challenging. Picard et al (Picard et al., 2004) argued that affective state is hard to measure, and cannot be directly measured (Calvo & Mello, 2010). That is due to there is lack of a clear theory that defines emotions, which are constructs or conceptual quantities with fuzzy boundaries and with substantial individual difference variations in expression and experience. The biggest challenge is to bring together research of theorists and practitioners from different fields, including psychology, neuroscience, physiology and social science, in order to refine the terminology with respect to affect and learning. Although there are research attempting to give clear dimension on emotion flourishes in many disciplines and specialties, yet experts cannot agree on its definition (Izard, 2007).

In viewing that measuring emotions in large scale is difficult, this study aims to measure only stress. Stress can degrade reception and cause inefficient learning (LePine, LePine, & Jackson, 2004; O'Neil & Spielberger, 1979). If possible to be detected automatically, it could be useful for affective computing developers to build effective e-learning that helps to identify the stressors that cause poor learning behavior, such as mismatched demand by the teachers, frustrating resources, or bad usability design, which brings negative effect to learning.

2.1 Emotional Stress Definition

Stress, is a kind of affective state that is hard to express and to be quantified clearly. It is vague in some way, and lacking a fixed, precise definition. Stress is defined as "a state of mental or emotional strain or tension resulting from adverse or demanding circumstances" by the Oxford dictionaries (OxfordDictionaries, 2016). Lazarus & Folkman (Lazarus & Folkman, 1984) viewed stress as "a feeling experienced that a person perceives that demands exceed the personal and social resources the individual is able to mobilize", which concerned primarily on human emotion and feeling of stress. Most people viewed stress as some unpleasant threat, and was generally considered as being synonymous with distress (http://www.stress.org/what-is-stress). Distress involves unresolved feelings of fear, anxiety and frustration, although Selye argued that stress can be good or bad (Selye, 1946, 1956).

2.2 The Objective Measurement of Emotional Stress

The challenge of stress measurement is to determine a solid construct that can objectively quantify the strength of stress. It was found that the objective measures could take into consideration task demand strength, available resources such as time duration, and influence of external stress stimuli such as unpleasant environment. Some psychological theories suggested that in a task-specific environment, user stress levels could be varied according to two factors: *demand* and *control.* For instance, stress increases when excessive demand on worker production involves unreasonable deadline, and there is lack of control over workplace processes (Karasek, 1979). Johnson and Hall (Johnson & Hall, 1988) proposed the Job Demand-Control-Support (JDCS) model to measure work stress and suggest that an iso-strain job, such as high demands-low control and low social support, could bring the most negative impact to the workers. Liao et al. (Liao, Zhang, Zhu, & Ji, 2005) compared the inferred stress level against job demands through visual features, physiological, behavioral and performance evidences. Their experiments showed that the inferred user stress level by their system was consistent with that predicted by Karasek (Karasek, 1979).

However, those objective measures using physiological devices cannot have the relevance and power of direct reporting of feelings about stress, hence it is particularly difficult to find objective criteria against which to validate self-report measures of stress (Crandall, 1976). For instance, two individuals could have different stress perception even they are given the same task demand and resources, dependent on how much the individual can tolerant with the stress. If stress is considered as a kind of emotion that is subjective to human perception of a task demand, then self-report survey is an important tool for the preliminary stage that requires large amount of samples in order for us to study the relationship between stress perception, job performance and learner behaviors when using mouse and keyboard, which help to build a valuable dataset for the analysis in the later stage.

Lim et al (Lim, Ayesh, & Stacey, 2016b)applied the Motivation/Attitude-driven Behavior (MADB) model in the e-learning context and identified that the results were generally consistent with the MADB model formalized by Wang (Wang, 2007). They found that menu design could be a stimulus that significantly affects students' stress perceptions and motivations, while motivation is affected by stress perception. Behavior is significantly correlated to mouse dynamics such as mouse speed, mouse idle duration and mouse left click rate. This significant effect of behavior on mouse dynamics may be caused by motivation and decision, but not stress perception and attitude. Since the impacts of a student's behavior on mouse dynamics could be observed, they strongly believe that there is a potential to compute the student's cognitive processes with emotions, motivations and attitude, by observing the changes of mouse behavior.

3 STRESS MEASUREMENT USING MOUSE/KEYSTROKE DYNAMICS

This research extends the previous work done by Lim et al, which examined the effects of search task (menu design) (Lim et al., 2014c, 2016b), assessment (mental arithmetic) (Lim, Ayesh, & Stacey, 2014a, 2015b) and text typing (Lim, Ayesh, & Stacey, 2014b, 2015a) demands on user stress perceptions, mouse dynamics, keystroke dynamics and job performances. User stress perceptions were collected using a self-report with 7-point Likert-scale. Mouse data and keystroke inputs were gathered by the automated mouse and key loggers every 10ms interval. Mouse data included move speed, idle duration, idle occurrence and click rate, and keystroke inputs included keystroke speed, key latency and the use of error correction keys. The job performance was measured based on the duration spent to complete a question, error rate and passive attempt. Passive attempt was calculated based on the attempt of hitting the give up button, re-visitation of a question, or the attempt to wait until the time is up. They conducted three different tests to assess the effects of menu design, mental arithmetic, and typing on users' stress perceptions, task performance and sensor behaviors.

The first preliminary research studied the effects of indirect instruction, such as searching for a learning material, to learner's stress perception, job performance and mouse dynamics. There was a total of 151 undergraduate students required to search 64 different links from different web menus. The research found that job performance was believed to be affected by the nature of task, i.e. whether the task required the users to use more cognitive power to comprehend or to process the possible feature to be searched. In terms of the effects on mouse dynamics, the menu design is found significantly affects user's mouse behavior. When the users made more errors, attempt to revisit the question, or to give up the task, they demonstrated longer mouse idle time (indicates that they did not move the mouse often), but fast mouse speed when they needed to use the mouse. Left mouse click was only affected by error rate. On the other side, when the users spent longer time in the search task, they demonstrated longer mouse idle time, but slower mouse speed. The research also reported that the users agreed that they felt stressed when they needed to spend longer time in the search task. To conclude, the stress level of a user might increase if a task duration, error counts, attempt to revisit question, and attempt to give up have increased from the previous task, then the mouse speed would become slower and the mouse idle time is longer.

The second preliminary research focused on examining the effects of direct learning instruction related to assessment to learner's stress perception, job performance and mouse/keystroke dynamics. Mental arithmetic problems under time pressure were used since they can effectively induce cognitive stress (Owen, McMillan, Laird, & Bullmore, 2005; Setz et al., 2010; Sloan, Korten, & Myers, 1991). All participants were required to answer 10 mental arithmetic questions, which they must not use calculator nor working on paper. The task demands were increased from the first question to the last question, according to the increment of digit per number and the amount of numbers in a question. From the statistical analysis, task demand is correlated to students' stress perceptions, job performance (duration spent, error rate and passive attempt), mouse behavior (mouse speed, mouse click rate and mouse idle duration), and keyboard behavior (keystroke speed, key latency). The correlation results were consistent with what was reported in the menu search task above. However, there were a few prominent anomalies occurred in especially the last two questions, which the task performance, mouse and keystroke behaviors did not behave in a way as expected. Anomalous behavior indicates three possibilities: (i) there is a wrong assumption about the demand of the question (ii) qualitative difference in task demands (e.g. increment of the number of digits per number in the mental arithmetic would require more working memory to process the task), or (iii) the student is either understress or overstress, which is beyond their motivation limits. At this point, prediction of cognitive stress level would become invalid, as the students has already lost motivation to continue the task. Therefore, it is important to activate the adaptive content to reengage the students to continue the task. The research also discovered that task demand was the main factor that influences student's stress perception, task performance, mouse and keystroke behaviors, but time pressure only provided a small significant effect.

The third preliminary research studied the effects of typing demand, i.e. text length and language familiarity, to learner’s stress perception, job performance and mouse/keystroke behavior. The participants were required to type six fixed-texts in the same mock-up online learning system. The six different typing tasks were set based on different text length and language familiarity. Three questions were set in English (as familiar language) and three in German (as unfamiliar language). The results showed that higher stress perception was associated with longer text length and lower familiarity of the language. Higher task demand generally resulted in longer task duration, higher error rate, slower mouse and keystroke speeds, longer mouse idle duration, and lower mouse idle occurrences and use of error key (such as delete key). This is consistent with what was discovered by Carneiro et al(Carneiro, Novais, Augusto, & Payne, 2017), in which lower mouse speed and slower typing rhythm can be observed generally when negative stimuli is introduced, such as music, which is deemed as noise that affects cognitive function during typing. It was also found that time pressure did not necessarily affect how users perceived their stress levels but it significantly affected task performance, mouse dynamics and keystroke dynamics. On the other side, language familiarity affected only task performance and keystroke behavior, while text length changed mouse behavior but not keystroke behavior. This suggests that we should mainly look into task performance and mouse behavior features if the typing tasks involve changes in length, and observe only task performance and keyboard behavior to understand whether a person is familiar with the given material. Lastly, the estimation of user's emotional stress level would become invalid once the learner is overstressed or has lost motivation, which results in anomalous behaviors, with unexpected job performance, mouse dynamics and keyboard dynamics.

To sum up the research findings above, several consistencies were discovered among the three different tasks in general:

1. If task demand increased, then the user stress perception, duration spent for a question, error rate, passive attempt, mouse idle duration may increase, but mouse speed, left mouse click and keystroke speed would decrease generally.

2. The correlation between job performance and mouse behavior is significant. Low job performance, for instance when the students attempt to revisit the task, give up, or when they make more errors, it usually comes together with longer mouse idle duration, and higher mouse speed. When the student has to perform the task with longer duration, then longer mouse idle duration and slower mouse speed could be observed.
3. Significant correlations between stress perception and mouse/keystroke dynamics are found in all tasks.
4. Task demand is the main factor that affects job performance, stress perception, mouse behavior and keystroke behavior.
5. The estimation of the emotional stress level based on job performance, mouse dynamics and keystroke dynamics might only be valid as long as the students are still engaged with the task. Once a student's stress level has gone beyond limit, or he or she has lost motivation, anomalous behaviors could be observed.
6. Task duration is significantly correlated to task demand (such as difficulty, familiarity and length) until anomalous behavior is observed as explained in item 5. above.

The preliminary research studies presented above also conclude that automated stress evaluation can be obtained through acquisition and processing of task performance, mouse behavior and keystroke behavior. The correlations between mouse behavior and keystroke behavior suggest that unifying mouse and keyboard dynamics analyses could be more useful than utilizing them alone.

4 EXPERIMENT SETUP

4.1 Sampling of Participants

Experimental and quantitative studies will be carried out with convenience sampling method (Gravetter & Forzano, 2015). Convenience sampling is the most commonly used sampling method in behavioral science studies, where researchers simply get participants who are available and willing to respond. In terms of sample size, we accept the margin of error (E) to be 10%, with 90% of confidence level (α=0.10). The recommended size is 67 for each experiment, based on the following (Weiss, 2004) :

$$n = 0.25 \left({}^{Z_{\propto/2}}\!/_{E} \right)^2 \qquad (1)$$

where Z0.05=1.64 and E=0.1,

A total of 190 second-year undergraduate students who studied Bachelor Degree in Computer Science, Bachelor Degree in Information Systems, and Bachelor Degree in Information Technology aged between 18 to 24 years old, were approached for their participations. Participants from narrow specializations and ages are selected under the constraint to control the effect of socio-demographic difference on stress perception (Lim, Ayesh, & Chee, 2013), when browsing the user interfaces in the search task. In addition, the items to search are also IT subject-related, which prior knowledge is needed when searching a desired learning material.

The participants were randomly assigned to different design treatments and a control group in the preliminary research study. However, there is no control group for the laboratory experiments of search task. All students will run the same experiments with the same sets of search instructions. As for the assessment and typing tasks, the students were randomly assigned into five different groups, i.e. they are given either with/without time constraint or timing, with/without clock display and with/without timer display. The groups are named following the code system below:

```
Timing (0 or 1) + Clock (0 or 1) + Timer (0 or 1)
```

where 0 means not available and 1 means available.

Group 000: It is the first control group, who are required to complete all questions without any time constraint. There is no clock display nor countdown timer.

Group 100: It is the experimental group where there is no display of clock nor countdown timer, but given 30 seconds time constraint.

Group 101: It is the experimental group where there is a countdown timer that flashes every second with yellow background.

Group 110: It is the experimental group where there is no countdown timer but a digital clock that displays the current date and time (which is updated every second).

Group 111: This group is able to see both clock display that is updated every second, and a countdown timer that flashes continuously in yellow background.

For all the experimental groups, all questions will be submitted automatically once the time is up.

Fourteen sessions of experiments are conducted within 2 weeks. As the participants were given option to withdraw from the experiments at any time, not all of them completed all the tasks. Provided valid data for the subsequent analyses, there are 151 participants for Search Task, 160 participants for Assessment Task, and 162 participants for Typing Task. Amongst these 190 students, 88.8% of them were male and almost all of them (99.4%) had at least one-year experience using Blackboard e-learning system.

4.2 Procedures

All the experiments were conducted in a computer laboratory that was equipped with computers that run on Windows 7, 17" monitor with resolution of 1024x768 pixels. Every computer was equipped with a standard external HID-compliant mouse. To simulate those tasks in the e-learning environment and to avoid the results to be affected by unfamiliarity with the interface when they begin the tasks, a mock-up application is built based on the learning management system (LMS) that was used by the university students, i.e. Blackboard™ Academic Suite. The Web pages that showed instructions and questions would run on Google Chrome by default. Fourteen sessions of experiments were conducted within 2 weeks. Before the experiments, the students were required to give consensus to carry out the subsequent tasks based on voluntarily basis. Once they agreed and proceeded to next page, they were required to perform calibration of their mouse and keystroke behaviors. Calibration is needed as a baseline of a student's mouse and keystroke behaviors measurement before his/her stress is deliberately elevated by the job demand and/or external stimuli given in the experiments. During the calibration of keystroke behavior, they were required to type their username, password, and a predefined phrase "The quick brown fox jumps over the lazy dog". During the calibration of mouse behavior, they must click on 5 hyperlinks that were distributed across the 4 corners and the center of the screen. Fig. 1 shows the screenshots of the calibration processes. Each time before they started a new task, they were given an instruction page to understand what activities they must do next. When they were ready, they needed to click the Start button, and the data collection for each question would begin. Each time the student starts a question, the task performance will be captured by the system automatically. . If they wished to give up and skip to the next question, they could click the Give Up button placed on the top right corner of the screen. For the search task, if they wished to revisit the question when feeling lost, they should click the Restart button, and they could try the same question again. Each time after the students completed a question (or skipped the question), a self-report survey will be displayed as follows:

“You felt stressed when answering the previous question”

Figure 1. Sample login screen for keystroke and mouse data calibration

This survey is useful to allow us to assess their stress perceptions SP, when solving a problem in a task, following 7-point Likert scale (1 for strongly disagree, 7 for strongly agree).

5 TOWARD A FRAMEWORK FOR CONTINUOUS STRESS MONITORING

5.1 Proposing a Continuous Stress Measurement for Online Environment

It would be useful to find a cost-effective method to allow stress to be measured continuously over an online environment. Therefore, the classifier's learning algorithm should be less complicated so that the processing time of stress measurement could be done almost instantly without causing delay to both sides of client and server. Three different approaches that can be useful in managing uncertainties and easily implemented in an online environment are certainty factors (CF), feedforward back-propagation neural networks (FFBP) and adaptive neuro-fuzzy inference systems (ANFIS). These classifiers are explained in Section 6.As the variability of users' habits in using mouse, keyboard and the time they would spend on a question is high, therefore only the difference of a user's task duration and mouse/keyboard activities between the current question and the previous question will be considered. There are two benefits of doing this: it is not only able to eliminate the variability between 2 persons, but this also allows us to construct a personalized stress measurement, to compare whether the current task is more challenging than the previous task, or whether the current stress level of the user has changed significantly compared to a moment ago.

To construct the stress classifier, 2 types of input data are needed. First, time duration, TD, that the student spent on each question must be measured. The stress measured based on time duration, S_{TD}, is defined in Equation 4. Second, the mouse and keystroke features that sense user's states, such as mouse speed and keystroke speed, are used to measure the changes of stress when the task demand is altered, each $S_{feature}$ is defined in Equation 5. The outcome of the predicted stress by the classifier is denoted as $S_{B(Sensor)}$, where B(Sensor) comprises mouse behavior B(M), keystroke behavior B(K), and the unification of both B(M,K).Training data and sample data are collected beforehand from the previous preliminary research experiments for the analysis. The data collection is described in Section 6.

5.2 Testing Criteria

To assess the effectiveness of the stress measurement methods that we propose in this research, there are two criteria to be tested.

1. Can S_{TD} and $S_{B(Sensor)}$ be generally used for the 3 tasks - Search, Assessment and Typing?
2. How are CF, FFBP and ANFIS different in terms of stress measurement accuracy?

The first criterion is crucial as we need a stress measurement that is context-independent, so that it can be applied regardless the type of task carried out by the user. If the measurement is different from task to task, then it is probably not adequate to be used as a generalized measurement if the effect of task on stress measurement is high. To validate this method of measurement, we need to answer the following hypotheses:

1.1. There is no difference in terms of S_{TD} between 3 tasks - Search, Assessment and Typing.

1.2. There is no difference in terms of $S_{B(M)}$, $S_{B(K)}$, and $S_{B(M,K)}$ between 3 tasks, i.e. Search, Assessment and Typing.

For the second criterion, the performance of the three models: CF, FFBP neural network and ANFIS lies within the accuracy of the measurement. We measure the performance by checking the overall accuracy, false acceptance rate (FAR), false rejection rate (FRR) and Equal Error Rate (ERR) of each model, which are defined as follows.

Accuracy. The normal stress level ($Y(S_{TD}) = 0$) is measured to be normal ($Y(S_{B(Sensor)}) = 0$), and vice versa.

FAR. The non-normal stress level ($Y(S_{TD}) \leq -1$ or $Y(S_{TD}) \geq 1$) is measured to be normal ($Y(S_{B(Sensor)}) = 0$).

FAR. The normal stress level ($Y(S_{TD}) = 0$) is measured to be either increased or decreased ($Y(S_{B(Sensor)}) \leq -1$or $Y(S_{B(Sensor)}) \geq 1$).

ERR A common way used in biometric research, to compare the accuracy of methods with different ROC (relative operating characteristic) curves. EER is often used as an indicator to tell which method is better than others although it is not necessary that the classifier must operate based on EER. Usually the method with lowest EER is the best (Bolle, 2004).

The output of S_{TD}, $Y(S_{TD})$, with the threshold of at least one standard deviation away (stdev) from the mean (mean(TD)) is activated by the following function.

$$Y(S_{TD}) = \begin{cases} X \text{ if} S_{TD} \geq mean(S_{TD}) + X \times stdev(S_{TD}) \text{ , indicates stress increased} \\ -X \text{ if } S_{TD} \leq mean(S_{TD}) - X \times stdev(S_{TD}) \text{, indicates stress decreased} \\ 0 \text{ if otherwise, indicates stress is stable (normal)} \end{cases} \quad (2)$$

where

$$X = \frac{S_{TD} - mean(S_{TD})}{stdev(S_{TD})} \quad (2.1)$$

and X ≥ 1, mean(S_{TD})= 0.0144 and stdev(S_{TD})= 0.3813 based on a total of 12,144 records.S_{TD} is defined in Equation 4.

To simplify the computation process, we assume that if the difference of the duration spent for the current question is at least one standard deviation from the mean, i.e. 68% are normal data, then the stress level has either increased or decreased, otherwise the stress level remains stable or normal (cf. Fig. 2).

Figure 2. Standard Deviation Function of Stress Measurement STD

We use the threshold of at least 1 standard deviation away (stdev) from the mean (mean($S_{B(Sensor)}$)) to determine the actual output of $S_{B(Sensor)}$, i.e. Y($S_{B(Sensor)}$), which is activated by the following crisp function.

$$Y\left(S_{B(Sensor)}\right) = \begin{cases} X \text{ if } S_{B(Sensor)} > mean\left(S_{B(Sensor)}\right) + X \times stdev\left(S_{B(Sensor)}\right) \text{, indicates stress increases} \\ -X \text{ if } S_{B(Sensor)} < mean\left(S_{B(Sensor)}\right) - X \times stdev\left(S_{B(Sensor)}\right)\text{, indicates stress decreases} \\ 0 \quad \text{if otherwise, indicates stress is stable (normal))} \end{cases} \tag{3}$$

where

$$X = \frac{S_{TD} - mean\left(S_{B(Sensor)}\right)}{stdev\left(S_{B(Sensor)}\right)} \tag{3.1}$$

and X ≥ 1.

At least one standard deviation is adopted as abnormal mouse and keystroke data could be observed when the tasks given to the participants become too demanding. These anomalies were reported in Section 3. With the computation of X in Equation 2 and Equation 3, the variations of very high or low stress can be determined. This is useful if one wishes to further differentiate the intensity of stress being measured.

6 DATA PREPARATION FOR STRESS CLASSIFIER

The following subsections explain the process of data preparation for the proposed stress classifiers, which consists of data acquisition, feature extraction, and creation of the training and sample sets containing labelled data. The three proposed stress classifiers are certain factors (CF), feedforward back-propagation (FFBP) neural networks and adaptive neuro-fuzzy inference system (ANFIS).

6.1 Data Acquisition and Feature Extraction

A personalized adaptation in an affective learning system is to be designed in the future, which should be controlled by the best stress classifier identified in this research. It is very important to enable a personalized adaptation, as there are huge differences between individuals in terms of mouse and keystroke behaviors. To allow the affective learning system to be developed, this research investigates the effectiveness of three stress measurement models that are proposed, by using a large amount of sample data collected during the experiments. Feature extraction is mainly used to reduce the measurement and storage requirements, to minimize training and utilization times, so that the prediction performance can be improved. Therefore, it is important to ensure the features extracted will not affect the real-life application performance. This research uses essential primary data generated from the time duration spent on a question, the self-reported stress perception, and the data gathered automatically from mouse and keyboard by a key logger and a mouse logger during the preliminary experiments that involve Search, Assessment and Typing. The features are listed as follows:

Time Duration, TD: Task performance consists of the time duration spent on the question (in milliseconds, scaled with $\log_{10}$ function). TD is computed from the moment when the question is displayed until the answer is submitted

Stress Perception, SP: A self-report of stress perceptions when solving a problem in a task, following 7-point Likert scale (1 for strongly disagree, 7 for strongly agree).

The mouse behavior is defined as a dataset that captures the mouse features for each task, as follows:

Mouse Behavior, B(M) *= <MS, MID, MIO , MCL>*, where
MS = Average mouse speed (pixels per second);
MID = Total mouse inactivity duration (ms);
MIO = Total mouse inactivity occurrences;
MCL = Left click rate per ms.

The keystroke behavior is defined below:

Keystroke Behavior, B(K) *= <KS, KL, KErr>*, where
KS = Average keystroke Seed (number of keystrokes per second);
KL = Keystroke latency (down-down key latency);

KErr = Total delete key and backspace key press.

Unfortunately, insufficient data of KErr was collected for the Assessment task, therefore KErr is excluded from the analysis. All the collected data are normalized using $\log_{10}$ function.

6.2. Creation of the Training Set and Sample Set

To enable stress measurement from time duration and mouse behavior, the features are re-computed with correlation coefficient values obtained from the Pearson correlation test. Correlation coefficients are used to measure the presence of the relationship among time duration TD, user's stress perception of each question, and mouse behavior and/or keystroke behavior features, which we obtained from the experiments conducted from all three tasks with total samples of 12,144 data. These coefficients yielded can be fixed as default parameters in order to build a stress measurement system. Although the parameters are fixed in this research, it is recommended for the future affective system to generate dynamic and adaptable parameters based on personified set of rules relating stress of each person individuality, such as what has been suggested by Arevalillo-Herráez et al (Arevalillo-Herráez et al., 2014).

The stress measured based on time duration, S_{TD}, is defined as follows:

$$S_{TDk} = \mathrm{amp}(r_{SPTD} * \frac{SPTD_k - SPTD_{k-1}}{SPTD_{k-1}}) \quad (4)$$

where the parameters, $r_{TD} = 0.3710$, $k =$ the current question, $k\text{-}1 =$ previous question (if k is the first question, then $k\text{-}1$ is the calibration), and *amp* is a function to amplify the output as the signal is too weak, provided the output S_{TD} values must still be in the range of $[-1, 1]$ ($amp = 10$ in this case). The *amp* function is needed due to after *TD* is being transformed using the log_{10} function, the difference of *TD* between 2 questions is very small, and so the difference between S_{TD} and $S_{B(Sensor)}$ will be too large, which would affect the results later.

The stress measurement values based on the changes of mouse and keystroke features, between 2 questions are as follows:

$$\mathrm{S}_{\mathrm{featurek}} = \mathrm{r}_{\mathrm{feature}} * \frac{\mathrm{feature}_{\mathrm{k}} - \mathrm{feature}_{\mathrm{k-1}}}{\mathrm{feature}_{\mathrm{k-1}}} \tag{5}$$

where feature consists of MS, MID, MIO, MCL, KS and KL. The parameters of each feature are $r_{MS} = -0.1503$; $r_{MID} = 0.3278$; $r_{MIO} = -0.0279$; $r_{MCL} = -0.0474$; $r_{KS} = -0.1111$; and $r_{KL} = 0.0919$. Similar to r_{TD}, these parameters are the correlation coefficients obtained from the Pearson correlation test against user self-evaluated stress perception. All the S values must be in the range of $[-1, 1]$ to ease classifier learning process later.

Table1 shows the number of training sets and sample sets prepared for each task.

6.3 Validation of the Predictive Model against Learners’ Stress Perception

Finally, the important last step is to validate the selected predictive model against the self-report feedback by the participants after answering a question. It is believed that keyboard and mouse dynamics could be the most suited approaches to assess inner state or behavior, such as stress (Carneiro & Novais, 2017). It would be interesting to find out whether user’s self perceived stress level, SP, could be reflected on their mouse and keyboard dynamics, since correlations between learners' stress perception SP and mouse and keyboard dynamics could be observed (Lim et al., 2015a, 2015b).Assessment Task is selected to conduct the testing, as mental arithmetic is believed able to affect user's cognitive load much more than other type of task, such as typing. Out of the 10 questions, only Question 2 until Question 7 are selected, in order to eliminate outlier values, i.e. anomolous behavior. Stress Perception, SP, collected are transformed into the following output Y(SP):

$$Y(SP) = \begin{cases} 1\ if SP_n > SP_{n-1}, indicatesstressincreases \\ -1\ if SP_n < SP_{n-1}, indicatesstressdecreases \\ 0\ if \text{ otherwise}, indicatesstressis \text{ stable } (normal) \end{cases} \qquad (6)$$

where n = Question no. and n>1.

From the 63 learners and 6 questions each from the Assessment Task, a total of 378 records will be used to test the model. The Y(SP) is compared against Y(S_{TD}) and Y($S_{B(Sensor)}$). Accuracy is measured if normal stress level is predicted to be normal (Y(S_{TD}) = 0 or Y($S_{B(Sensor)}$) = 0), and vice versa.

7 THE STRESS CLASSIFIER

This stage involves classifying the measured stress using either mouse, keystroke or the unification of both dynamics, into one of the crisp sets, i.e. Y=1 (stress increased significantly), Y=−1 (stress decreased significantly), or Y=0 (stress remains stable or normal). The thresholds set to classify the stress levels for the models proposed below are used as default constants, universal to all users at the initial stage. The future research would design adaptive threshold personalized to individual based on their own mouse and keystroke behaviors. Three different approaches that can be useful in managing uncertainties and easily implemented in an online environment are certainty factors (CF), feedforward back-propagation neural networks (FFBP) and adaptive neuro-fuzzy inference systems (ANFIS). The CF model and the architectures of FFBP and ANFIS for stress measurement are explained in the following sub-sections.

7.1 Certainty Factors

As in our inference rules, each premise in the rule is correspondent to $S_{feature}$ (Equation 5), the output of the rule is $S_{B(Sensor)}$, i.e. the certainty factor (CF) in the range of -1 and 1, represents a measure of belief or disbelief. The computation of the measured stress level is similar to MYCIN, but we have done some slight adjustment. The certainty factors of each rule are obtained using the correlation coefficients between learners' stress perception SP and the respective feature. The correspondent rules are as follows.

Rule 1: If MS decreased, then $S_{B(Sensor})$ increased (CF = r_{MS})

Rule 2: If MID increased, then $S_{B(Sensor})$increased (CF=r_{MID})

Rule 3: If MIO decreased, then $S_{B(Sensor})$increased (CF=r_{MIO})

Rule 4: If MCL decreased, then $S_{B(Sensor})$increased (CF=r_{MCL})

Rule 5: If KS decreased, then $S_{B(Sensor})$increased (CF=r_{KS})

Rule 6: If KL increased, then $S_{B(Sensor})$increased (CF=r_{KL})

The following Equation7 shows the computation of the $S_{B(sensor)}$,based on its respective feature, that represents each correspondent rules above.

$$S_{B(Sensor)_k} = CF = r_{feature} * \frac{feature_k - feature_{k-1}}{feature_{k-1}} \tag{7}$$

where feature consists of MS, MID, MIO, MCL, KS and KL. The values of r_{MS}, r_{MID}, r_{MIO}, r_{MCL}, r_{KS}, and r_{KL} are $r_{MS} = -0.1503$; $r_{MID} = 0.3278$; $r_{MIO} = -0.0279$; $r_{MCL} = -0.0474$; $r_{KS} = -0.1111$; and r_{KL}= 0.0919, as given in Equation 5.

The cumulative value of the certainty of the hypothesis, $S_{B(Sensor)}$) in each rule is updated by the combination formula given in Equation8 below. To form the $S_{B(M)}$, Rule 1 to Rule 4 are combined. To form the $S_{B(K)}$, Rule 5 and Rule 6 are combined. To form $S_{B(M,K)}$, Rule 1 to Rule 6 are combined.

$$S_{B(Sensor)} = \mathrm{CF(R1, R2)} = \begin{cases} \mathrm{CF(R1)} + \mathrm{CF(R2)} - \mathrm{CF(R1)} \times \mathrm{CF(R2)} \text{if } \mathrm{CF(R1)} > 0 \; and CF(R2) > 0 \\ \mathrm{CF(R1)} + \mathrm{CF(R2)} + \mathrm{CF(R1)} \times \mathrm{CF(R2)} \text{if } \mathrm{CF(R1)} < 0 \; and CF(R2) < 0 \\ \frac{\mathrm{CF(R1)} + \mathrm{CF(R2)}}{1 - \min(|\mathrm{CF(R1)}|, |\mathrm{CF(R2)}|)} \text{if otherwise} \end{cases} \tag{8}$$

7.2 Feedforward Back-Propagation Neural Network

Supervised learning is utilized to predict the outcomes of stress based on 2 different training sets. Accordingly, two neural networks are formed using the back-propagation training. The neural networks are used to predict the stress based on the changes of mouse and keystroke behaviors. The numbers of hidden neurons of the networks are correspondent to the numbers of inputs. There is only one hidden layer for each network.

The output of the networks are $S_{B(Sensor)}$, i.e. $S_{B(M)}$, $S_{B(K)}$ and $S_{B(M,K)}$, dependent on the inputs of the devices. The input features S_{KS} and S_{KL} are fed into the first network that predicts $S_{B(K)}$. The inputs for the second neural networks are S_{MS}, S_{MID}, S_{MIO} and S_{MCL} for the prediction of $S_{B(M)}$. The last network is given the inputs of all six features, generating the prediction of $S_{B(M,K)}$. All input data are produced in Equation 5. The target output for all networks is the $Y(S_{TD})(-1,0 \text{ or } 1)$ as shown in Equation 2. The distribution of training sets and sample sets are described in Table1. Since the inputs and the measurement of stress are in the interval of $[-1, 1]$, tansig function is used as the transfer function from the input layer to output layer, which will also return an output, Y, in $[-1, 1]$ (stress increased if $Y > 0$ or stress decreased if $Y < 0$). The algorithm of tansig function (Mathworks, 2015c) is a follows:

$$\text{tansig}(n) = 2/(1+\exp(-2*n))-1 \qquad (9)$$

After the training, to incorporate the classifier as the inference engine in the stress monitoring system, only the feedforward phase of the training algorithm need to be applied. The application procedure is as shown in Algorithm 1.

Algorithm 1. **APPLICATION PROCEDURE OF FEEDFORWARD ANN [FAUSETT 1994, P. 299]**

```
Initialize trained weights, $v_{ij}$and $w_{jk}$
for each input vector, x, do
    for i=1 till n: set activation of input unit $x_i$ // x is the input
    for j=1 till p
        $z\_in_j = v_{0j} + \sum_{i=1}^{n} x_i v_{ij}$   // the net input to the hidden unit j ($Z_j$);
        $z_j = \text{tansig}(z\_in_j)$   // the output signal of $Z_j$
    for k = 1 till m
        $y\_in_k = w_{0k} + \sum_{j=1}^{p} z_j w_{jk}$   //$y\_in_k$ is the net input to output unit k
```

$y_k = \text{tansig}(y\_in_k)$ //y_k is the output signal of output unit k

where x = input; v_{0j}=bias on hidden unit j; w_{0k}=bias on output unit k

7.3 Adaptive Neuro-Fuzzy Inference System

MATLAB is used to enable stress measurement with Adaptive Neuro-Fuzzy Inference System (ANFIS) (Mathworks, 2015b). To simplify the explanation on how it works, we illustrate the first fuzzy inference system (FIS) in Fig. 3, which is used to predict the stress based on the changes of keystroke behavior. The other FISs are used to predict the stress based on the changes of mouse behavior B(M) that contains 4 inputs, and the unification of both behaviors, B(M,K) that contains 6 input features .

Figure 3. ANFIS architecture with 2 inputs

First we hypothesize a parameterized model structure of the first FIS as below:

RULE 1: If $x_1$is A_1 and $x_2$is B_1 then $f_1=p_1x_1+q_1x_2+ t_1$

RULE 2: If $x_1$is A_2 and $x_2$is B_1 then $f_2=p_2x_1+q_1x_2+ t_2$

RULE 3: If $x_1$is A_3 and $x_2$is B_1 then $f_3=p_3x_1+q_1x_2+ t_3$

RULE 4: If $x_1$is A_1 and $x_2$is B_2 then $f_4=p_1x_1+q_2x_2+ t_4$

RULE 5: If $x_1$is A_2 and $x_2$is B_2 then $f_5=p_2x_1+q_2x_2+ t_5$

RULE 6: If $x_1$is A_3 and $x_2$is B_2 then $f_6=p_3x_1+q_2x_2+ t_6$

RULE 7: If $x_1$is A_1 and $x_2$is B_3 then $f_7=p_1x_1+q_3x_2+ t_7$

RULE 8: If $x_1$is A_2 and $x_2$is B_3 then $f_8=p_2x_1+q_3x_2+ t_8$

RULE 9: If $x_1$is A_3 and $x_2$is B_3 then $f_9=p_3x_1+q_3x_2+ t_9$

where $\mathbf{x}$ = [S_{KS}, S_{KL}](S_{KS} and S_{KL} are defined in Equation 5) and {p_i, q_i, t_i}is the parameter set. Note that f is a linear function.

Next we prepare input/output data into input/output vectors. Each FIS consists of 3 membership functions for all premises. The distribution of training sets and sample sets are described in Table1. The input vector to be fed to the first FIS is $\mathbf{x}$ = [S_{KS}, S_{KL}](produced in Equation 5). The input vector for the second FIS is $\mathbf{x}$ = [S_{MS}, S_{MID}, $S_{MIO}S_{MCL}$] (produced in Equation5). The input vector for the third FIS is $\mathbf{x}$ = [S_{MS}, S_{MID}, S_{MIO}, S_{MCL}, S_{KS}, S_{KL}]. The target output for both networks is the Y(S_{TD}), where Y(S_{TD}) = -1,0 or 1, as computed in Equation 2.

Layer 1 shows three node functions, which are the membership functions (A_i) that specify the degrees to which the given x satisfies the quantifier A_i according to symmetric Gaussian function (Mathworks, 2015a), as follows:

$$O_i^1 = \mu_{A_i}(x) = exp\left(-\frac{(x-c)^2}{2\sigma^2}\right), \text{c and } \sigma \text{ are arbitrary real constants} \qquad (10)$$

Then in Layer 2, the production of incoming signals from Layer 1 is generated, and the output is sent to Layer 3. Since there are two inputs, Layer 1 should produceO_i^1 andO_j^2. The node function of Layer 2 will be

$$w_{ij} = O_i^1 \times O_j^2, i = 1,2,3;\ j = 1,2,3 \qquad (11)$$

Layer 3 calculates the ratio of the ith rule's firing strength, w_i, to the sum of all rules' firing strengths. The output, which is called normalized firing strengths, is as follows:

$$\bar{w}_i = \frac{w_i}{\sum_i w_i}, i = 1,2,3; n = 3 \tag{12}$$

In Layer 4, the subsequent parameters are produced by the following node function:

$$O_i^4 = \bar{w}_i f_i = \bar{w}_i(p_i x + r_i) \tag{13}$$

Consider in Layer 5, which is also the output layer, it is a single node that computes the overall output as the summation of all incoming signals from Layer 4, which is:

$$O^5 = \sum_i^n \bar{w}_i f_i \tag{14}$$

Thus we have demonstrated how an ANFIS is constructed. The concept to build the other FIS is similar, except that for the one based on B(M) has 81 fuzzy rules with 5 parameters (as there are 4 inputs with 3 correspondent membership functions). For example,

RULE 1: If $x_1$is A_1 and $x_2$is B_1 and x_3 is C_1 and $x_4$is D_1 then $f_1=p_1x_1+q_1x_2+r_1x_3+s_1x_4+ t_1$

where { p_1, q_1, r_1, s_1, t_1} is the parameter set.

As for the FIS based on B(M,K), there will be 729 rules with 7 parameters since it has 6 inputs.

8. RESULTS AND ANALYSIS

8.1 Test 1: Using $S_{B(Sensor)}$ and S_{TD} to Measure Stress in Three Different Tasks

Univariate analysis (ANOVA) is used to test the difference in terms of S_{TD}, and multivariate analysis (MANOVA) (IBM & IBM Knowledge Center, 2011; IBM Knowledge Center & IBM, 2012) is carried out to test the difference in terms of $S_{B(M)}$, $S_{B(K)}$ and $S_{B(M,K)}$ among different tasks. As keystroke dynamics are only involved in the Assessment and Typing tasks, we separated the analyses into two parts. The first focuses on the effects of all 3 tasks on $S_{B(M)}$

only, while the second tests the effects of Task on $S_{B(M)}$, $S_{B(K)}$and $S_{B(M,K)}$. Table 2shows the results.

The differences between tasks provide no significant effect on S_{TD} at all, but they give a significant effect on $S_{B(M)}$, $S_{B(K)}$ and $S_{B(M,K)}$. Although the effects of different tasks on these $S_{B(M)}$, $S_{B(K)}$ and $S_{B(M,K)}$ are significant, nevertheless high Wilks' lambda values ($\lambda > 0.97$) indicate that the effects are very small and could be ignored, as high value indicates small omnibus effect(Grice & Iwasaki, 2007; Sullivan & Feinn, 2012).

8.2 Test 2: The Performance of CF, FFBP and ANFIS

Table 3 demonstrates the false acceptance rate (FAR), false rejection rate (FRR), the overall accuracy and the equal error rate (EER) for CF, FFBP neural net and ANFIS in the measurement of $Y(S_{B(Sensor)})$ (Equation 3) against $Y(S_{TD})$ (Equation 2). From the results, the average FRR and FAR are 19.11% and 79.63% for CF; 13.47% and 29.66% for FFBP neural net; and 12.37% and 34.44% for ANFIS. The 3 models produce an average of 67.25%, 82.88% and 83.60% overall accuracy respectively by CF, FFBP neural net and ANFIS. The average EER for each model is 54.16% by CF, 47.20% by FFBP neural net and 49.83% by ANFIS. In terms of FRR, FAR, overall accuracy and EER, FFBP neural net appears to provide the best results among all models.

8.3 Validation of the Predictive Model against Learners’ Self-Report Stress Perception

The predictive stress model is tested by comparing its results with the participant's self-report stress perception SP, as explained in Section 6.3. FFBP neural net model was selected as it provided best results among the three tested models. Based on the Assessment Task 378 sample records, the following Table 4 shows the accuracy results to predict normal stress based on $Y(S_{TD})$ and $Y(S_{B(Sensor)})$ against $Y(SP)$ (Equation 6).

9 DISCUSSION

This preliminary research compares three simple stress classifiers, which could be effectively used in an online environment due to their simple architectures to manage uncertainty in the collection of learner's states. We measure learner's stress by computing the changes of task completion time and mouse/keystroke features of a user between two consecutive questions. The computation is done using the correlation coefficients that relate users' self-evaluated stress perceptions gathered from the previous preliminary research experiments. We envisaged measuring stress by detecting the changes of behaviors between two tasks or two time intervals does not only eliminate the high variability of learners' states, such as behaviors in using mouse/keyboard or the time each individual would spend on the same task, but to also allow the construction of a personalized stress measurement. Besides, we could easily identify whether the current task is more challenging than the previous job, and most importantly to enable a mechanism to continuously monitor or measure stress level from time to time. Although for this research, the correlation coefficients of stress perception need to be obtained from the past user's survey, nevertheless these values give significant clues to us on how the timing data and sensors could react to user's stress states. These values can be set as default constants or parameters for the initial rule-based stress measurement model. Our future work will identify the process to dynamically generate adaptable set of parameters for personified emotion detection.

9.1 The Effects of Tasks on S_{TD} and $S_{B(Sensor)}$

To explore a stress measurement method that is context-independent, so that it can be applied to various task carried out by the user, we compared the effects of 3 different tasks, i.e. Search, Assessment and Typing, on S_{TD} and $S_{B(Sensor)}$. If the effects of the tasks on the stress measurement are significant, this indicates that the accuracy of the measurement could be affected when the user switches between tasks. The result shows that the effect of tasks on S_{TD} is not significant at all. This gives us a very good benchmark on testing $S_{B(Sensor)}$ against S_{TD}. Unfortunately, the effect of different task on $S_{B(Sensor)}$ is significant for most features. This significant effect shows that the users may have demonstrated different mouse behavior during different tasks. This could be due to in certain activity such as mental arithmetic, the user's cognitive load, which could be reflected by mouse/keystroke behavior, is higher than other type of task, such as typing. Secondly, it could be due to typing task requires lesser mouse/keystroke activities as compared to search. Although the effect of tasks on $S_{B(Sensor)}$ is significant, fortunately the effect size is small, which is considered meaningless and could be ignored (Sullivan & Feinn, 2012). In addition, despite the effect is significant, it would only last temporarily as after the task is switched, the stress measurement is continued by detecting the behavioral changes between 2 consecutive questions or 2 time intervals.

9.2 The Performance of the Stress Classifiers

In terms of assessing the effectiveness of the three stress classifiers in measuring stress, namely certainty factor (CF), feedforward back-propagation (FFBP) neural net and adaptive neuro-fuzzy inference system (ANFIS), the research attempted to determine the best classifier that is producing the best False Acceptance Rate (FAR), False Rejection Rate (FRR), overall accuracy and Equal Error Rate (EER). The tests show some promising results. First, the FFBP neural network produces best performances among all. It is easy to be applied in the stress inference system but it requires data to be trained before the application can be implemented. Besides, its overall performance for all three tasks is better than CF and ANFIS. ANFIS on the other hand, its performance is slightly lower than FFBP, but its overall results are considered as good as FFBP. Unfortunately, there are two major limitations of using ANFIS: (1) pre-application training is required; (2) if the number of inputs and membership functions are high, it could be programming and processing load challenging as it needs high number of rules and fuzzy sets to be built. The last classifier, CF is easy to use and its simple algorithm should not harm the processing performance of the computer. In addition, unlike FFBP or ANFIS, it does not require the data to be trained beforehand. Therefore, it is very easy to be implemented in the web environment.

However, the greatest limitation is the reliability of the stress measurement results. Among the 3 models, CF achieves lowest overall accuracy and EER, but highest FRR and FAR. However, despite of poorest performance, the overall accuracy is 60.22%, which is still considered acceptable for an emotion classifier, since we should not forget the fact that the inaccurate results could be due to anomalous behavior, which the users might give up the task (in shorter time) but the stress level is still high (higher S value).

To examine the best model to be used as the inference engine for the stress measurement system, we tested the accuracy of CF, FFBP neural net and ANFIS in measuring the correct hypothesis of $Y(S_{B(Sensor)})$ against $Y(S_{TD})$. Although mostly used in biometrics research but not in emotion recognition, FRR, FAR and EER can be used as an indicator to know the performance of the stress measurement by the 3 models, instead of relying on overall accuracy itself. From the results, FFBP neural net produces best overall FRR (13.47%), FAR (29.66%), accuracy (82.88%) and EER (47.20%) compared to CF and ANFIS.

9.3 Validation of the Predictive Model against Learners' Self-Report Stress Perception

Direct reporting of feelings about stress is a traditional approach to collect stress measurement from a user. Self-reporting is believed more relevant as only the participant knows how stressed that he or she felt. On the other side, mouse and keyboard dynamics-driven approaches, are considered as new methods that are able to collect inner state information from a user automatically(Carneiro et al., 2017), where research by (Carneiro

& Novais, 2017; Lim et al., 2014a, 2014b, 2015a; Lim, Ayesh, & Stacey, 2016a; Lim et al., 2016b; Vizer, 2009)have shown the correlations between stress and keyboard and mouse dynamics. Therefore, we would like to examine whether the mouse and keyboard dynamics-driven approaches could be used to predict what a user perceives. The highest prediction accuracy is produced by the unification of both mouse and keyboard dynamics B(M,K), which is 58.99%. This shows that the unification of both mouse and keyboard data is more useful than utilising each alone. The low prediction accuracy of the predictive model against self-report stress perception could be due to a few reasons. First, it is difficult for the participants to quantify their stress level, which they may not observe minor change of stress, but can be observed through implicit behaviour, such as mouse and keystroke movements. Second, a participant may purposely or unconsciously hide information, lie about his or her feeling, or simply answer the survey. Last, the sample size is too small. Small sample size is not able to generate generalized results.

On the other sides, it is also interesting to observe that the prediction accuracy produced by time duration is 61.38%, which is slightly higher than mouse and keyboard dynamics. This shows that: (1) participants can observe or estimate the time they spent on a task, the longer time they spent on a task, they would feel more stressed; However if the duration of the time they spent on a task is similar but the challenge of the task increases, they might not be able to observe the changes of their cognitive stress level although there is; (2) it could be useful if time factor could be included in the predictive model in the future, to examine the effects of unification of time factor and other sensor data, such as mouse and keyboard dynamics.

10 PROPOSED ARCHITECTURE OF THE INTELLIGENT TUTORING SYSTEM BASED ON MOUSE AND KEYSTROKE DYNAMICS

The results showed high potential to use mouse and keystroke dynamics alone in stress measurement and classification. Hence, an application of the automated stress measurement model using mouse and keystroke dynamics can be designed and applied in an Intelligent Tutoring System (ITS). Fig. 4 illustrates a proposed architecture of ITS.

Figure 4. The architectural design of the Intelligent Tutoring System

Figure 5. Sample interface for the examiner to add question and difficulty level

The ITS first requires the examiner to insert a number of questions with different levels of difficulties. The questions are then saved in a database table called `QuestionBank`. To setup an assessment, the examiner must specify the expected level of difficulty of each question. Sample interface is given in Fig.5. Before the students start the assessment, they are required to login to the system so that the calibrations of keystroke dynamics and mouse dynamics can be collected. The reason for performing calibrations is to manage the huge temporal variations of keystroke and mouse dynamics of individual user, and also the high behavioral differences between individuals. The calibration is useful as a benchmark to determine whether the subsequent learning activities are considered significantly more stressful, stable/normal, or less stressful. A sample login screen for keystroke and mouse data calibration is given in Fig. 1. Once the students start the assessment module, the question will be retrieved from the `QuestionBank` table automatically. The job performance, such as error made and time spent on each question, must be collected for stress measurement and adaptation. The keystroke and mouse dynamics data shall be collected every 10 milliseconds. Adaptation could be made once significant change of stress level is detected, e.g. to adjust the difficulty level of instructional content of the assessment.

11 CONCLUSION

The results of this research demonstrate high feasibility to use mouse and keystroke dynamics alone in stress measurement and classification. The outcome of this research also suggests that feedforward back-propagation (FFBP) neural net could be the best model to construct the stress classifier in the inference engine, followed by adaptive neuro-fuzzy inference system (ANFIS) and lastly certain factors (CF). Overall the stress measurements by CF, FFBP neural net and ANFIS are on par with the existing research in the area of emotion measurement using keyboard and mouse dynamics (Kolakowska, 2013). The limitation of this research is it only detects stress. Detecting stress alone may not be enough for affective learning, which requires better understanding of granularity of emotion. However, it is very useful to determine the stressor that causes student's troubled behavior in learning. Our future research will apply the proposed stress measurement model using both mouse and keystroke dynamics in an affective e-learning system. Section

10 proposed the overall design and the architecture of the Intelligent Tutoring System (ITS) with the application of the stress measurement model using mouse and keystroke dynamics. This includes the detailed designs of the stress inference engine, which is the core of the ITS, the adaptive assessment and interface, and the collective feedback reporting system.

The main limitation of the current research is that the applications of the proposed stress measurement model in the ITS are not rigorously validated. Future work will include the revalidation of the predicted overall stress of the trained model in this experiment, with physiological parameters, such as cortisol, blood pressure or heart-beat measurements. Future research will also look into algorithms that can produce a more personalized adaptive learning system rather than using constant parameters in the trained model. Since a cheap, task independent, ubiquitous and less obtrusive means of estimating users' stress levels can be produced based on automated mouse and keystroke dynamics analyses, we strongly believe that many valuable applications in affective computing can be developed, such as usability testing, personalized games, and adaptive web, on top of affective learning.

REFERENCES

Arevalillo-Herráez, M., Arnau, D., Marco-Giménez, L., González-Calero, J. A., Moreno-Picot, S., Moreno-Clari, P., … others. (2014). Providing personalized guidance in arithmetic problem solving. In M. Kravcik, O. C. Santos, & J. G. Boticario (Eds.), *Personalization Approaches in Learning Environments PALE 2014* (pp. 42–48). Denmark.

Bolle, R. (2004). *Guide to Biometrics*. Springer. Retrieved from https://books.google.co.uk/books?id=NTJle3OodUsC

Calvo, R. A., & Mello, S. D. (2010). Affect Detection : An Interdisciplinary Review of Models , Methods , and Their Applications. *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, *1*(1), 18–37. https://doi.org/10.1109/T-AFFC.2010.1

Carneiro, D., & Novais, P. (2017). Quantifying the effects of external factors on individual performance. *Future Generation Computer Systems*, *66*, 171–186. https://doi.org/10.1016/J.FUTURE.2016.05.019

Carneiro, D., Novais, P., Augusto, J. C., & Payne, N. (2017). New Methods for Stress Assessment and Monitoring at the Workplace. *IEEE Transactions on Affective Computing*, *14*(8), 1–1. https://doi.org/10.1109/taffc.2017.2699633

Crandall, R. (1976). Validation of self-report measures using ratings by others. *Sociological Methods & Research*, *4*(3), 380–400.

Crawford, H. (2010). Keystroke dynamics: Characteristics and opportunities. In *Privacy Security and Trust (PST), 2010 Eighth Annual International Conference on* (pp. 205–212). https://doi.org/10.1109/PST.2010.5593258

Gravetter, F. J., & Forzano, L.-A. B. (2015). *Research Methods for the Behavioral Sciences* (5th ed.). Cengage Learning. Retrieved from https://books.google.co.uk/books?id=Kzx-BAAAQBAJ&pg=PA147&dq=convenience+sampling+in+research&hl=en&sa=X&redir_esc=y#v=onepage&q=convenience sampling in research&f=false

Grice, J. W., & Iwasaki, M. (2007). A truly multivariate approach to MANOVA. *Applied Multivariate Research*, *12*(3), 199–226.

IBM, & IBM Knowledge Center. (2011). GLM Univariate Analysis. Retrieved from http://pic.dhe.ibm.com/infocenter/spssstat/v20r0m0/index.jsp?topic=/com.ibm.spss.statistics.help/idh_glmu.htm

IBM Knowledge Center, & IBM. (2012). Multivariate General Linear Modeling. Retrieved from http://pic.dhe.ibm.com/infocenter/spssstat/v21r0m0/index.jsp?topic=/com.ibm.spss.statistics.cs/glmm_intro.htm

Izard, C. E. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science*, *2*(3), 260–280.

Johnson, J. V., & Hall, E. M. (1988). Job strain, work place social support and cardiovascular disease: a cross sectional study of a random sample of the Swedish working population. *American Journal of Public Health*, *78*(10), 1336–1342. Retrieved from http://ajph.aphapublications.org/doi/pdf/10.2105/AJPH.78.10.1336

Karasek, R. A. (1979). Job demands, job decision latitude and mental strain: implications for job design. *Administrative Science Quarterly*, (24), 285–308.

Kolakowska, A. (2013). A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *HSI* (pp. 548–555).

Kort, B., Reilly, R., & Picard, R. W. (2001). An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion. In *Advanced Learning Technologies, 2001. Proceedings. IEEE International Conference on* (pp. 43–46). https://doi.org/10.1109/ICALT.2001.943850

Landowska, A., Szwoch, M., Szwoch, W., Wróbel, M. R. R., Kołakowska, A., & Kolakowska, A. (2014). Emotion Recognition and Its Applications. In *Human-Computer Systems Interaction: Backgrounds and Applications 3* (pp. 51–62). Springer.

Lazarus, R. S., & Folkman, S. (1984). *Stress, Appraisal, and Coping*. Springer Publishing Company. Retrieved from https://books.google.co.uk/books?id=i-ySQQuUpr8C

LePine, J. A., LePine, M. A., & Jackson, C. L. (2004). Challenge and hindrance stress: relationships with exhaustion, motivation to learn, and learning performance. *Journal of Applied Psychology*, *89*(5), 883.

Liao, W., Zhang, W., Zhu, Z., & Ji, Q. (2005). A Real-Time Human Stress Monitoring System Using Dynamic Bayesian Network. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (p. 70).

Lim, Y. M., Ayesh, A., & Chee, K. N. (2013). Socio-demographic Differences in the Perceptions of Learning Management System (LMS) Design. *International Journal of Software Engineering & Applications*, *4*(5), 15–35. Retrieved from http://airccse.org/journal/ijsea/papers/4513ijsea02.pdf

Lim, Y. M., Ayesh, A., & Stacey, M. (2014a). Detecting cognitive stress from keyboard and mouse dynamics during mental arithmetic. In *Science and Information Conference, SAI 2014* (pp. 146--152). https://doi.org/10.1109/SAI.2014.6918183

Lim, Y. M., Ayesh, A., & Stacey, M. (2014b). Detecting Emotional Stress during Typing Task with Time Pressure. In *Science and Information Conference 2014* (pp. 329--338). London: IEEE Xplore.

Lim, Y. M., Ayesh, A., & Stacey, M. (2014c). The Effects of Menu Design on Users' Emotions, Search Performance and Mouse Behaviour. In S. Patel, Y. Wang, W. Kinsner, D. Patel, G. Fariello, & L. A. Zadeh (Eds.), *IEEE 13th Int'l Conf. on Cognitive Informatics & Cognitive Computing (ICCI*CC'14)* (pp. 541–549). London: IEEE.

Lim, Y. M., Ayesh, A., & Stacey, M. (2015a). The Effects of Typing Demand on Emotional Stress, Mouse and Keystroke Behaviours. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Intelligent Systems in Science and Information 2014* (Vol. 591, pp. 209–225). Switzerland: Springer. https://doi.org/10.1007/978-3-319-14654-6

Lim, Y. M., Ayesh, A., & Stacey, M. (2015b). Using Mouse and Keyboard Dynamics to Detect Cognitive Stress During Mental Arithmetic. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Intelligent Systems in Science and Information 2014* (Vol. 591, pp. 335–350). Switzerland: Springer. https://doi.org/10.1007/978-3-319-14654-6

Lim, Y. M., Ayesh, A., & Stacey, M. (2016a). Exploring Direct Learning Instruction and External Stimuli Effects on Learner's States and Mouse/Keystroke Behaviours. In *4th International Conference on User Science and Engineering 2016 (i-USEr 2016)* (pp. 1–6). Melaka: i-USEr 2016.

Lim, Y. M., Ayesh, A., & Stacey, M. (2016b). The Motivation/Attitude-driven Behavior (MADB) model in E-Learning and the Effects on Mouse Dynamics. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 10*(3), 38–53.

Mathworks. (2015a). Gaussmf (Gaussian Curve Membership Function). Retrieved April 14, 2015, from http://uk.mathworks.com/help/fuzzy/gaussmf.html

Mathworks. (2015b). Neuro-Adaptive Learning and ANFIS. Retrieved April 14, 2015, from http://uk.mathworks.com/help/fuzzy/neuro-adaptive-learning-and-anfis.html

Mathworks. (2015c). tansig: Hyperbolic tangent sigmoid transfer function. Retrieved April 17, 2015, from http://www.mathworks.com/help/nnet/ref/tansig.html

O'Neil, H. F., & Spielberger, C. D. (1979). *Cognitive and affective learning strategies*. Academic Pr.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, *25*(1), 46–59. https://doi.org/10.1002/hbm.20131

OxfordDictionaries. (2016). Stress. Retrieved March 30, 2016, from http://www.oxforddictionaries.com/definition/english/stress

Peters, L. H., O'Connor, E. J., Pooyan, A., & Quick, J. C. (1984). Research note: The relationship between time pressure and performance: A field test of Parkinson's Law. *Journal of Organizational Behavior, 5*(4), 293–299.

Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., … Strohecker, C. (2004). Affective learning-a manifesto. *BT Technology Journal*, *22*(4), 253–269.

Russel, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Salmeron-Majadas, S., Baker, R. S., Santos, O. C., & Boticario, J. G. (2018). A Machine Learning Approach to Leverage Individual Keyboard and Mouse Interaction Behavior from Multiple Users in Real-World Learning Scenarios. *IEEE Access*, *6*, 39154–39179. https://doi.org/10.1109/ACCESS.2018.2854966

Selye, H. (1946). *Stress in health and disease*. Butterworth.

Selye, H. (1956). *The stress in life*. McGraw-Hill.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Troster, G., & Ehlert, U. (2010). Discriminating Stress From Cognitive Load Using a Wearable EDA Device. *Information Technology in Biomedicine, IEEE Transactions On*, *14*(2), 410–417. https://doi.org/10.1109/TITB.2009.2036164

Sloan, R. P., Korten, J. B., & Myers, M. M. (1991). Components of heart rate reactivity during mental arithmetic with and without speaking. *Physiology & Behavior, 50*(5), 1039–1045. https://doi.org/10.1016/0031-9384(91)90434-P

Sullivan, G. M., & Feinn, R. (2012). Using effect size-or why the P value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282.

Svenson, O., & Maule, A. J. (1993). *Time pressure and stress in human judgment and decision making*. Sringer.

Szalma, J. L., Warm, J. S., Matthews, G., Dember, W. N., Weiler, E. M., Meier, A., & Eggemeier, F. T. (2004). Effects of sensory modality and task duration on performance, workload, and stress in sustained attention. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(2), 219–233.

Tsoulouhas, G., Georgiou, D., & Karakos, A. (2011). Detection of Learnerś Affetive State Based on Mouse Movements. *Journal of Computing*, *3*(11), 9–18.

Vea, L., & Rodrigo, M. M. (2017). Modeling negative affect detector of novice programming students using keyboard dynamics and mouse behavior. *Trends in Artificial Intelligence: PRICAI 2016 Workshops. PRICAI 2016. Lecture Notes in Computer Science*, *10004*, 127–138. https://doi.org/https://doi.org/10.1007/978-3-319-60675-0_11

Vizer, L. M. (2009). Detecting cognitive and physical stress through typing behavior. *Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems CHI EA 09*, 3113. https://doi.org/10.1145/1520340.1520440

Wang, Y. (2007). On the Cognitive Processes of Human Perception with Emotions, Motivations, and Attitudes. *International Journal of Cognitive Informatics and Natural Intelligence*, *1*(4), 1–13.

Continuous Stress Monitoring under Varied Demands Using Unobtrusive Devices https://doi.org/10.4018/jcini.2007100101

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063-- 1070.

Weiss, N. A. (2004). *Elementary Statistics* (6th ed.). Addison-Wesley.

Zimmermann, P., Guttormsen, S., Danuser, B., & Gomez, P. (2003). Affective Computing – Measuring Mood with Mouse and Keyboard. *International Journal of Occupational Safety and Ergonomics*, *9*(4), 2–5.

Table 1. Distribution of training sets and sample sets for the three tasks

TASK	**Number of participants**	**Number of records**	**Training set**	**Sample set**
SEARCH (64 questions)	151	9,582	5,900	3,682
ASSESSMENT (10 questions)	159	1,590	960	630
TYPING (6 questions)	162	972	600	372
TOTAL	171	12,144	7,460	4,684

Table 2. Univariate and Multivariate Tests on the Effects of Tasks on S_{TD} and $S_{B(Sensor)}$

Effect of Task	S_{TD}	$S_{B(M)}$					$S_{B(K)}$			$S_{B(M,K)}$
		S_{MS}	S_{MID}	S_{MIO}	S_{MCL}	Effect size	S_{KS}	S_{KL}	Effect size	Effect size
	p-value	*p*-value				Wilks' λ	*p*-value		Wilks' λ	Wilks' λ
All tasks	0.3823	0.03e-26	0.01e-29	0.1930	0.03e-18	0.971	N/A	N/A	N/A	N/A
Assessment and Typing	0.4557	0.0777	0.0003	0.4133	0.8877	0.994	0.0748	0.0016	0.993	0.986

The difference is significant at the level of p < 0.05 (2-tail)

Table 3. The Performance of CF, FFBP and ANFIS

Model	**Task**	**B(Sensor)**	**FRR**	**FAR**	**Overall Accuracy**	**EER %**
CF	Search	B(M)	456/2580 (0.1767)	880/1102 (0.7985)	2346/3682 (0.6372)	49.33
	Assessment	B(M)	125/549 (0.2277)	60/81 (0.6074)	445/630 (0.7063)	54.12
		B(K)	**54/549 (0.0984)**	76/81(0.9383)	**500/630 (0.7937)**	**46.18**
		B(M,K)	143/549 (0.2605)	58/81 (0.7160)	429/630 (0.6810)	51.61
	Typing	B(M)	89/318 (0.2799)	42/54 (0.7778)	241/372 (0.6479)	**53.58**
		B(K)	36/318 (0.1132)	40/54 (0.7407)	296/372 (0.7957)	68.54
		B(M,K)	87/318 (0.2736)	44/54 (0.8148)	241/372 (0.6479)	55.77
FFBP	Search	B(M)	302/2580 (0.1171)	**225/1102 (0.2042)**	3155/3682 (0.8569)	**48.11**
	Assessment	B(M)	113/549 (0.2058)	**36/81 (0.4444)**	481/630 (0.7635)	**29.57**
		B(K)	81/549 (0.1475)	**64/81 (0.7901)**	485/630 (0.7698)	48.53
		B(M,K)	**86/549 (0.1566)**	46/81 (0.5679)	498/630 (0.7905)	**34.41**
	Typing	B(M)	**29/318 (0.0912)**	**23/54 (0.4259)**	**320/372 (0.8602)**	57.16
		B(K)	30/318 (0.0943)	**36/54 (0.6667)**	306/372 (0.8226)	53.58
		B(M,K)	**57/318 (0.1792)**	**17/54 (0.3148)**	**298/372 (0.8011)**	59.03
ANFIS	Search	B(M)	**255/2580 (0.0988))**	267/1102 (0.2423)	**3160/3682 (0.8582)**	49.73
	Assessment	B(M)	**81/549 (0.1475)**	64/81 (0.7901)	**548/630 (0.8698)**	40.50
		B(K)	80/549 (0.1457)	68/81 (0.8395)	482/630 (0.7651)	51.70
		B(M,K)	91/549 (0.1658)	**34/81 (0.4198)**	**505/630 (0.8016)**	54.12
	Typing	B(M)	51/318 (0.1604)	24/54 (0.4444)	297/372 (0.7984)	55.45
		B(K)	**14/318 (0.0440)**	40/54 (0.7407)	**318/372 (0.8548)**	**51.22**
		B(M,K)	69/318 (0.2170)	22/54 (0.4074)	281/372 (0.7554)	**46.10**

Table 4. Validation of the Predictive Model against Y(SP)

Y(S_{TD})	Y($S_{B(M)}$)	Y($S_{B(K)}$)	Y($S_{B(M,K)}$)
0.613757	0.558201	0.571429	0.589947

Author Biographies

Yee Mei Lim, PhD (DMU, 2017), is a Senior Lecturer at TAR UC. Her research focuses on emotion detection using non-invasive methods, sentiment analytics on social media text, big data analytics, intelligent scheduling and production planning. She is the lead of Centre for ICT Innovations and Creativity at TAR UC.

Aladdin Ayesh, MSc (Essex, 1996), PhD (LJMU, 2000), is currently a Reader in Artificial Intelligence at De Montfort University. His current research focuses on computational cognition and machine learning. He has over 140 publications. He is a founding editor of four international journals and chaired several international conferences.

Martin Stacey (BA Oxford 1983 MS Carnegie-Mellon 1984 PhD Aberdeen 1992) is currently Senior Lecturer in Computer Science at De Montfort University. His main areas of teaching are human-computer interaction and system design methods. His research applies perspectives from psychology, artificial intelligence, sociology and philosophy to understanding design process.

Table of Responses to IJCHI Reviewers’ Comments

Reviewer #	Reviewer Issue/Comment to Be Addressed	Response/Action Taken	Document Location
1	It would be important to show a section explaining, in a simplified way, the architecture/data-flow of the full system (Hardware, Software Frameworks, Databases, etc.) - main features and it's limitations;	The design of an Intelligent Tutoring System that applies the stress predictive model is added.	Section 10 Page 39
1	Regarding the e-learning platform, it would be useful to know which platform is being used. In addition, is it possible to add the biometric metrics acquisition feature into other e-learning platforms?	To simulate those tasks in the e-learning environment and to avoid the results to be affected by unfamiliarity with the interface when they begin the tasks, a mock-up application is built based on the learning management system (LMS) that was used by the university students, i.e. Blackboard™ Academic Suite.	Section 4.2 Page 16
1	Equations (2) and (3) shows the way the stress is calculated. Although both formulas calculate how the stress varies (increasing/decreasing/stable), it does not distinguish very high/low stress variations. As a solution, a possible suggestion would be to follow the formula: ◦ X if S_b(Sensor) > mean(S_B(Sensor)) + X*stdev(S_B(Sensor)) ◦ -X if S_b(Sensor) < mean(S_B(Sensor)) + X*stdev(S_B(Sensor))) ◦ 0 if otherwise, indicates stress is stable (normal)	Suggestions were taken and changed accordingly. However, as this research is to identify whether mouse and keyboard dynamics can be used to detect stress for varied tasks, hence detecting stress at different level is more useful for future work.	Section 5.2 Equation 2 (Page 20) Equation 3 (Page 21)
1	In this study, an important last step would be to re-validate the predicted overall stress of the trained model by comparing it's results with the participant's feedback (after the exam).	The validation of the predictive model against users’ self-report stress perception is discussed.	Section 6.3 (Page 25) Section 8.3 (Page 33)
1	I suggest, also, to the authors to analyse some others related works: Carneiro D., Novais P., Augusto JC., Payne N., New Methods for Stress Assessment and	The 2 suggested papers were included and cited.	Section 1 (Page 2) Section 3 (Page 12) Section 6.3 (Page 25) Section 9.2 (Page 38)

	Monitoring at the Workplace, IEEE Transactions on Affective Computing, ISSN: 1949-3045, 2017. https://doi.org/10.1109/TAFFC.2017.2699633 Carneiro D., Novais P., Quantifying the effects of external factors on individual performance, Future Generation Computer Systems - The International Journal of eScience, Elsevier Science BV, ISSN: 0167-739X, Vol. 66, pp 171-186, 2017. http://dx.doi.org/10.1016/j.future.2016.05.019		
3	It would benefit from further clarifying how the proposed approach addresses individual variations in stress levels in various contexts and tasks, and in performance due to stress.	it is recommended for the future affective system to generate dynamic and adaptable parameters based on personified set of rules relating stress of each person individuality. Our future work will identify the process to dynamically generate adaptable set of parameters for personified emotion detection.	Section 6.2 (Page 24) Section 9 (Page 35) Section 11 (Page 43)
3	Finally, the work could be significantly enhanced by comparing the proposed approach with stress measuring techniques based on physiological parameters, which can now also be measured quite unobtrusively and continuously.	Future work will include the revalidation of the predicted overall stress of the trained model in this experiment, with physiological parameters, such as cortisol, blood pressure or heart-beat measurements.	Section 11 (Page 42)