Recognizing Student Emotions using Brainwaves and Mouse Behavior Data

Judith Azcarraga, Center for Empathic Human-Computer Interactions, De La Salle University, Manila, Philippines

Merlin Teodosia Suarez, Center for Empathic Human-Computer Interactions, De La Salle University, Manila, Philippines

ABSTRACT

Brainwaves (EEG signals) and mouse behavior information are shown to be useful in predicting academic emotions, such as confidence, excitement, frustration and interest. Twenty five college students were asked to use the Aplusix math learning software while their brainwaves signals and mouse behavior (number of clicks, duration of each click, distance traveled by the mouse) were automatically being captured. It is shown that by combining the extracted features from EEG signals with data representing mouse click behavior, the accuracy in predicting academic emotions substantially increases compared to using only features extracted from EEG signals or just mouse behavior alone. Furthermore, experiments were conducted to assess the prediction accuracy of the system at points during the learning session where several of the extracted features significantly deviate in value from their mean. The experiments confirm that the prediction performance increases as the number of feature values that deviate significantly from the mean increases.

Keywords: *Affect Recognition, Brainwaves, Electroencephalography (EEG), Mouse Behavior, Tutoring Systems*

INTRODUCTION

Students experience various emotions while engaged in learning. Such emotions, also referred to as academic emotions (Pekrun, 2002), may affect the flow of learning as well as the motivation to continue with the learning task. This has been the challenge for the human tutors and even for those who develop intelligent tutoring systems (ITS). Indeed, effective tutors, whether human tutors or computer-based intelligent tutoring systems, are those who are not only aware of the cognitive needs of the students but also of their affective needs.

With this in mind, recent research projects in the area of ITS have tried to address not only the cognitive needs of students but their affective needs as well. Such affective systems, also referred to as *affective tutoring systems*, consider the effect of emotions in the learning process of a learner as well as the typical emotional patterns under different learning scenarios. Academic

DOI: 10.4018/jdet.2013040101

emotions typically experienced by a learner while using a tutoring system are *confidence* (Arroyo et al., 2009; Azcarraga et al., 2011a, 2011b, 2011c; Ibañez et al., 2011), *excitement* (Arroyo et al., 2009; Azcarraga et al., 2011a, 2011b, 2011c; Ibañez et al., 2011), *frustration* (Arroyo et al., 2009; Azcarraga et al., 2011a, 2011b, 2011c; Burleson, 2006; Ibañez et al., 2011), *interest* (Arroyo et al., 2009; Azcarraga et al., 2011a, 2011b, 2011c; D'Mello & Graesser, 2009; Ibañez et al., 2011; Kapoor, Burleson & Picard, 2007) , *flow/engagement* (D'Mello & Graesser, 2009; Stevens, Galloway & Berka, 2007), *boredom* (D'Mello & Graesser, 2009) and *confusion* (D'Mello & Graesser, 2009).

Affective tutoring systems are capable of recognizing student affect based on tutorial information complemented with the user profile. Furthermore, these systems sometimes also include a combination of facial expression, gesture and physiological signals. In (Arroyo et al., 2009), affective states such as *confident, frustrated, excited* and *interested* are predicted with high accuracy using special devices such as a camera to capture facial expression, posture chair to monitor the level of engagement, pressure-sensitive mouse and skin-conductance sensor. Similarly, Burleson (2006) uses the same set of devices in order to predict student *frustration* and the *need for help*. Moreover, tutorial information such as conversational cues, posture and facial features are used in *Autotutor* to predict *boredom, flow/engagement, confusion* and *frustration* (D'Mello & Graesser, 2009).

Another physiological sensor also explored in detecting student emotions is the EEG sensor which reads brainwaves. Such a device can measure the electrical activity in the brain induced by the electro-chemical processes related to the firing of neurons. Negative emotions, such as "disgust" were found to be associated with right-sided activation in the frontal and anterior temporal regions whereas "happiness" was found to be associated with left-sided activation in the anterior temporal region (Davidson, 2000). Nevertheless, whether a given spike in neuron activities as captured by an EEG sensor is indeed induced by some emotion cannot be ascertained. Muscle movements near the eyes and forehead are typical noise/artifacts in EEG recording. Various other artifacts may also get (wrongly) captured. As explained later, serious care must be given to pre-processing EEG data in order to increase its signal-to-noise ratio and at some point be able to isolate segments of EEG signals, over some sustained period.

Past researches have used brainwaves information to measure user alertness and cognitive workload (Sanei & Chambers, 2007), while others have used these to predict the stress level (Heraz et al., 2009) and emotional dimensions (pleasure, valence, arousal and dominance) (Frantzidis et al., 2010; Heraz, Razaki, & Frasson, 2007). In Stevens, Galloway, & Berka (2007), the student's level of frustration, distraction and cognitive workload were observed while the student is engaged in different activities in a multimedia-learning environment.

Similarly, in the previous work of the authors (cf. Azcarraga et al., 2010), the level of problem difficulty faced by academic achievers is predicted based on brainwaves. Those who assessed the problems as easy tended to have higher excitement level compared to those who found the tasks to be difficult. Moreover, those who experienced difficulty with the problems tended to be more frustrated.

Other research efforts have been directed at detecting student affective states while using some learning environment with various sensors connected to the head or body of the learner (Azcarraga et al., 2011a, 2011b, 2011c; Chanel, 2009; Ibañez et al., 2011). In Chanel (2009), classification based on EEG led to a higher accuracy for the assessment of the valence dimension of emotions as compared to the peripheral features from GSR, temperature, BVP, HR and respiration. Also, emotion valence and arousal prediction have improved when EEG features are combined with these peripheral features.

The problem with many of these physiological sensors, aside from being expensive, is that they sometimes interfere with the natural learning environment of the student. Their obtrusive nature might affect the cognitive processes and may also result in poor prediction of the true academic emotion of the learner. One device that may capture affect-related information and is non-obtrusive is a computer mouse. A computer mouse may be a standard input mouse or a special one that is sensitive to the hand pressure and possibly also sensitive to other physiological manifestations. The use of a pressure-sensitive mouse in a tutoring scenario has been explored in Arroyo et al. (2009), Burleson (2006), Kapoor, Burleson, and Picard (2007). In yet another research, a biometric mouse that captures user biometrics was explored in Kaklauskas et al. (2011) to measure a user's emotional state and productivity. This special mouse device can capture physiological behavior based on skin conductance, amplitude of hand tremble and skin temperature, and motor behavior based on mouse pressure, speed and acceleration of mouse pointer movement, scroll wheel turns, right- and left-click frequency.

In fact, even a standard mouse may provide useful information about a user's emotional state and interest, as demonstrated in Scheirer et al. (2001) and Zimmerman et al. (2003). Scheirer et al. (2001) investigated various mouse behavior patterns such as the mouse movement when users were presented with frustration-eliciting events while playing a game. The reason, it seems, is that a person's emotions and mood may affect not only his/her physiological manifestations but also his/her motor movements (Zimmermann et al., 2003)

In this study, a standard input mouse was used to capture hand motor behavior in detecting student affective states such as *confidence, excitement, frustration* and *interest*. Mouse behavior information such as the number of clicks, the duration of each click and the distance traveled by the mouse were taken as features to classify the four affective states. We compare the prediction accuracy when using solely mouse information with the accuracy when brainwaves information were used as supplementary features.

EXPERIMENTAL SET-UP

Twenty-five computer science undergraduate students (14 male and 11 female) aged 17 to 21, all mentally healthy and all right-handed, were asked to participate in the experiment (Ibañez, Lim & Lumanas, 2011). All had already taken an intermediate algebra course. During the experiment, they were asked to learn the tutoring software *Aplusix* that teaches algebra (Nicaud, Bouhineau, & Huguet, 2002). The participants were asked to solve 4 algebra equations of different difficulty levels for a period of 15 minutes. The tutorial session was designed so as to elicit a variety of emotions, including frustration and excitement, by having a wide range of difficulty levels. While using the software, their brainwaves were captured using an EEG sensor attached to their head. Moreover, the details of their mouse clicks, click duration and movement were automatically captured and stored in 2 different mouse log files - one for the clicks and duration and another for the movement.

The EEG sensor that was used in the experiment is the *Emotiv EPOC* headset, a commercially available EEG sensor typically used for gaming purposes. The *Emotiv EPOC* is equipped with 14 channels (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4) according to the international standard 10-20 locations. The experiment was conducted in a quiet room to avoid external noise that may affect the signals. Further, the EEG sensor was regularly checked if it is well attached and its signals are well captured by the computer connected thru a USB terminal. A special service program was written to automatically capture the raw EEG signals coming from each of the channels.

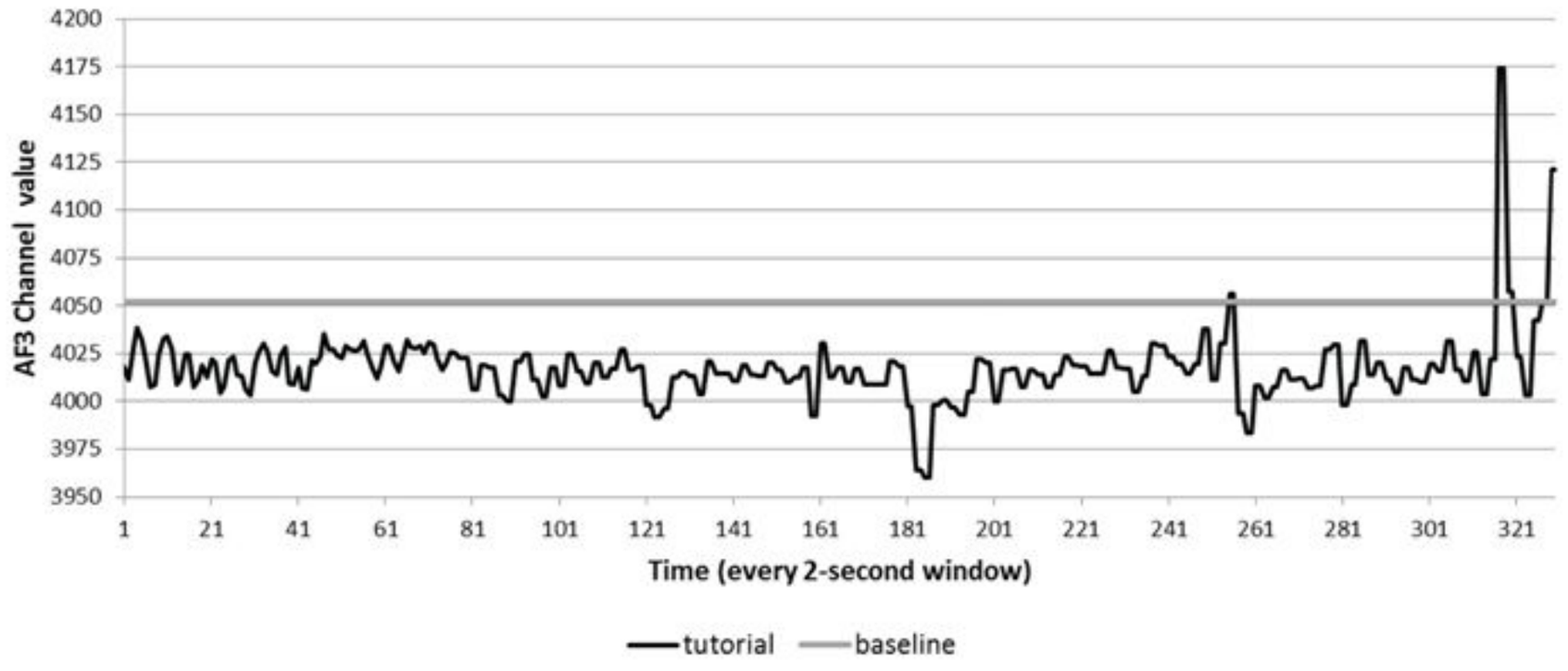
Prior to the actual learning session, the "resting-state" EEG signals of each subject were measured by asking them to relax and close their eyes for a period of 3 minutes. These resting-state EEG signals were used as baseline measures, following the methodology described in (Davidson et al., 1990; Tomarken et al., 1992).

The participants were then given brief instructions on how to use the software and the self-report window. While solving algebra equations, an emotion annotation window automatically pops up every 2 minutes in which the participant can conveniently report the level of intensity of each of the 4 emotions (*confidence*, *excitement*, *frustration* and *interest*) using a sliding bar with values from 1 to 100 for each of the four emotions. This part of the experiment has been pre-tested with students to check whether they get irritated or stressed by the self-report pop-up window every two minutes, and whether a period of two minutes is not too often for them. The feedback has consistently been that the sliding bars are intuitive and natural enough to use and that they are not negatively affected when being asked to rate the intensity of their emotion. The two-minute interval is also not too often for them. Indeed, 120 seconds is quite a long period when one is using a learning system on screen.

DATA PREPARATION AND RESULTS

Two EEG recordings were collected from each subject: one from the resting-state period and one from the tutorial session. During the resting-state period, the values of each EEG channel for each subject were averaged. The average value serves as the baseline EEG of that particular subject, for the given channel (Davidson et al., 1990; Tomarken et al., 1992). Guided by the methodology in psycho-physiological research on emotion (cf. Davidson et al., 1990), the raw EEG channel values taken during the tutorial session were processed and filtered by computing the difference between the raw value of the channels and the mean value of corresponding channels from the baseline (resting-state) data. These differences were then normalized. Figure 1 provides an illustration of a sample baseline value.

Figure 1. Baseline of a student for the AF3 sensor channel representing the average EEG value for the AF3 channel measured during the resting-state period. The EEG signals for the same channel measured during the tutorial period is super-imposed on the graph. Note that in this case, the AF3 signal stays mostly below the baseline value, or the resting-state EEG, during the entire length of the tutorial session.



According to psycho-physiological literature (Levenson, 1998; Ekman, 1984), emotions persist for about 0.5 to 4 seconds. Guided by this, we used 2-second window samples. All the pre-processed EEG data, mouse data, and self-reported emotion tag were carefully synchronized, merged and uniformly segmented into 2-second windows with 1-second overlap. Each segment was treated as a single instance in each subject's dataset. The full dataset had a total of 17 features: 14 for the EEG channels and 3 for mouse behavior (number of clicks, distance travelled, click duration). The most dominant self-reported emotion serves as the tag for each recorded instance. The most dominant emotion is identified as the emotion with the highest intensity value among the 4 emotions. In case there are two or more emotions with the same highest value, the most dominant emotion in the previous instance is chosen.

Mouse activities such as click duration and movement were automatically captured by a service program and stored in two different mouse log files. Raw values such as the x- and y-positions of the mouse were stored. Moreover, the button, left or right, that was clicked and the duration of such click were also stored in a separate file. The button that was clicked and the duration of the click every 2 seconds with 1 second overlap were computed. In addition, the spatial (Euclidean) distance travelled by the mouse every 2 seconds is computed.

From the 25 subjects, data from only 16 students were found to be useful, given the stringent conditions we set in terms of balancing the data for all the four different emotions. Subjects with very few outlier features were eliminated. All the instances or data points for each student are combined and balanced according to emotions. The number of instances for each emotion is the same. Moreover, the number of instances for each subject for that particular emotion is also the same. Thus, no participant has more instances than the others for any emotion.

Balanced student instances and emotions were generated in each dataset (to be discussed in the next section) in order to avoid any bias in classifying emotions, which is a required data preparation step for Multi-Layered Perceptrons (MLP). MLPs require datasets to be balanced, otherwise the more MLP will tend to classify most instances as belonging to the most common emotion, i.e. the emotion with the most number of instances.

In predicting academic emotions, performance measures such as Precision, Recall and F-Measure were used. Such measures were based on the computed True Positive (tp), False Positive (fp), True Negative (tn) and False Negative (fn). These are the standard metrics for MLP and Support Vector Machines (SVM).

Precision is the probability that a class A is true among all that have been classified as class A. This is also referred to as the Positive Predictive Value (PPV):

$$\text{Precision (P)} = \frac{\mathbf{tp}}{(\mathbf{tp}+\mathbf{fp})} \quad (1)$$

Recall is the proportion of examples which were classified as class A among all instances of class A. This is also referred to as the True Positive Rate or Sensitivity:

$$\text{Recall (R)} = \frac{\mathbf{tp}}{(\mathbf{tp}+\mathbf{fn})} \quad (2)$$

F-Measure is the combined computation of precision and recall. This is the harmonic mean of Precision and Recall in which both are evenly weighted:

$$\text{F-Measure (FM)} = \frac{(2 * \mathbf{Precision} * \mathbf{Recall})}{(\mathbf{Precision} + \mathbf{Recall})} \quad (3)$$

These 3 performance measures were used to assess the performance of classification models, MLP and SVM. MLP is a feed-forward artificial neural network that can be trained in a supervised manner to map a given input set to output set (Haykin, 2008; Hornik, Stinchcombe

& White, 1989; Rumelhart, Hinton & Williams, 1986). The input-output mapping need not be known to the user, but numerous examples of the correct input-output pairs must be available so that the network can learn to make the proper association. It has been proven in Hornik, Stinchcombe, and White (1989) that MLPs are universal approximators – that is, some MLP can be trained to learn any mathematical function, possibly non-linear, between input and output sets, except that it is not known how many hidden units would be needed for every input-output mapping to be learned. In other words, any underlying mathematical function between input features and some set of output states or categories can be approximated by a MLP. Because the relationship between EEG signals and emotions is still largely unknown, the general approximation capability of MLPs is clearly attractive.

The SVM is also a universal constructive learning procedure that is based on the statistical learning theory (Vapnik, 1995). It is "universal" since it can be used to learn a variety of representations such as neural networks, radial basis functions, and so on (Cherkassky & Mulier, 2007). It has been shown to be a powerful classifier that has the statistical basis for arriving at optimal hyperplanes that would separate the data into their respective categories (Cherkassky & Mulier, 2007). An SVM model provides a representation of data points in space that are mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. Indeed, some initial results on emotional assessment are reported to have been based on SVM (Chanel, 2009).

WEKA, a machine learning tool for feature classification (Hall et al., 2009), was used. Table 1 presents the performance of each model using a 10-fold cross validation technique for testing and validating the data. When using 10-fold cross validation, the training based on the train set and the classification based on the test set is run 10 times. For each of the 10 runs, the dataset is randomly partitioned in such a way that a sample is isolated as a test set (10% of the entire dataset), while the rest (90% of the entire dataset) are used for training. The sample test set is then replaced with a new sample set for the second iteration, and the remaining samples are used as training set. This is done 10 times per run, and the average performance for recall, precision, and f-measure are averaged over 10 runs. For MLP and SVM classifiers, this technique is the standard way of measuring classification performance.

Based on the results in Table 1, MLP seems to perform generally better than SVM particularly when brainwaves features are used in the

Table 1. Performance results of multi-layered perceptrons (MLP) and support vector machines (SVM)

Classifier	Dataset	Brainwaves Only			Mouse Only			Brainwaves + Mouse		
		P	R	FM	P	R	FM	P	R	FM
MLP	Confidence	0.48	0.38	0.42	0.33	0.17	0.23	0.54	0.49	0.51
	Excitement	0.54	0.55	0.54	0.33	0.3	0.31	0.58	0.63	0.6
	Frustration	0.54	0.5	0.52	0.32	0.21	0.25	0.6	0.57	0.59
	Interest	0.59	0.73	0.65	0.32	0.62	0.42	0.72	0.75	0.73
SVM	Confidence	0.27	0.15	0.2	0.3	0.21	0.25	0.35	0.25	0.29
	Excitement	0.42	0.45	0.43	0.19	0.01	0.01	0.41	0.41	0.41
	Frustration	0.37	0.36	0.36	0.32	0.51	0.39	0.44	0.4	0.42
	Interest	0.37	0.52	0.43	0.33	0.55	0.41	0.44	0.61	0.51

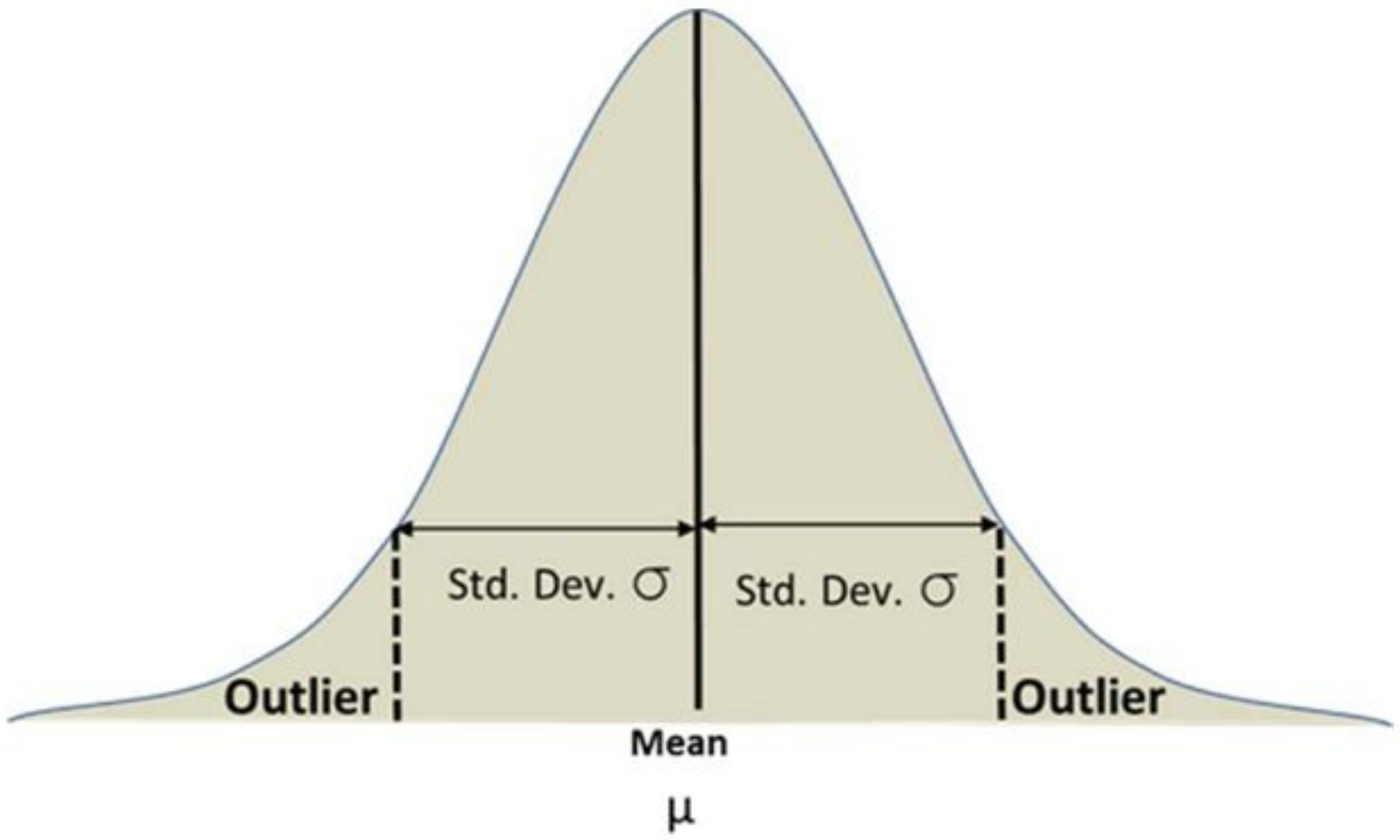
classification. In both cases, the classification performance has consistently increased when brainwaves features are complemented with mouse behavior information. It can also be observed that among the four emotions, whether using MLP or SVM, *interest* was predicted most accurately for all the 3 modalities. Using f-measure as the main basis for classification performance, the comparative performance show f-measures for brainwaves only of 0.43 for SVM and 0.65 for MLP, for mouse only of 0.41 for SVM and 0.42 for MLP, and for the combination of brainwaves and mouse, the f-measure is 0.51 for SVM and as high as 0.73 for MLP. This is a significant finding because *interest* may be highly correlated with engagement which is an essential factor for student motivation. Note that in the succeeding sections, we will see that the classification performance for all emotions would increase once we restrict the classification to only those instances when features values have been noted to be "special", in that they deviate in a major way from the rest of the feature values.

SELECTIVE PREDICTION

Six different datasets were formed based on the percentage of so-called feature "outliers" (Azcarraga et al., 2011c). Azcarraga et al. (2011c) considers some feature values to be "special", signifying that something caused to be distinct from the other features values. These special feature values are referred to as an "outlier" if they deviate by at least one standard deviation from the mean. To be precise, a feature value v_f is an "outlier" if it deviates from the mean μ_f by at least one standard deviation, denoted by σ_f, as defined in (4). Means and standard deviations are computed per feature and per subject/student. Figure 2 provides an illustration of outlier data points:

$$\begin{aligned} v_f &> \mu_f + \sigma_f \\ v_f &< \mu_f - \sigma_f \end{aligned} \tag{4}$$

Figure 2. A feature value is treated as special and is considered an "outlier" if it deviates in value by at least one standard deviation from the mean



Feature values that are *outliers* for each instance are thus counted. Based on this number, different datasets were formed as described in Table 2. The full dataset (or Dataset 0) contains the instances from all the 16 subjects. Dataset 10 consists of only those instances where at least 10% of the feature values are outlier values. Dataset 25 consists of only those instances where at least 25% of the feature values are outlier values, and so on. Each dataset may contain multiple samples from a single subject.

Each dataset was balanced by ensuring that the number of instances for each emotion is the same. This is to avoid any bias that would severely affect the classification of MLP (This is not an issue for SVM). For Dataset 60, only 15 subjects were included since 1 subject did not have instances that had at least 60% outlier features.

For each dataset classification accuracy of each modality, whether brainwaves or mouse, as well as of their combination is analyzed using MLP and SVM models. The performance of the full dataset (Dataset 0) is shown in Table 1. As what was observed in Table 1, Table 3 also shows that even for the individual datasets, the same comparative performance results are observed when we compare the performance of each modality (i.e. mouse only, brainwaves only, and mouse plus brainwaves). Classification based on brainwaves sensor data was consistently and significantly better than when based on just the data from mouse behavior. The results of Table 3 which present the average for all the four emotions very clearly show that the classification performance improves when data from both the EEG sensors and the mouse clicks are combined.

Using different datasets which are formed according to the number of outlier features, it is also very interesting to note that the results clearly show that the prediction accuracy increases when instances in a dataset have feature values that deviate significantly from their mean values of a given subject. It should be emphasized that such "outlier" feature values that deviate significantly from the mean baseline figure may indicate that the EEG is picking up something unusual or the mouse is being handled or clicked somewhat differently (Azcarraga et al., 2011c). The findings not only show that there are significant increase in the prediction accuracy when the predictions are made once outliers are detected. More importantly, the results clearly show that as the number of outliers increase (from 0% to 10% to 25%, to 33%, to 50%, to 60%), the prediction accuracy based on MLP systematically increases from 61% to 70% to 77% to 82% to 88% to 92%. The SVM results, although registering lower prediction accuracies, also show the very same trend. Figure 3(a) describes such trend.

Figure 3(a) and Table 3 clearly show that the classification performance increases for datasets composed of only instances that deviate (by at least one standard deviation) from the mean. As the datasets become more and more restrictive (10% outliers, 33% outliers, and so on), the classification performance also increases. There is clear evidence that when classification is at-

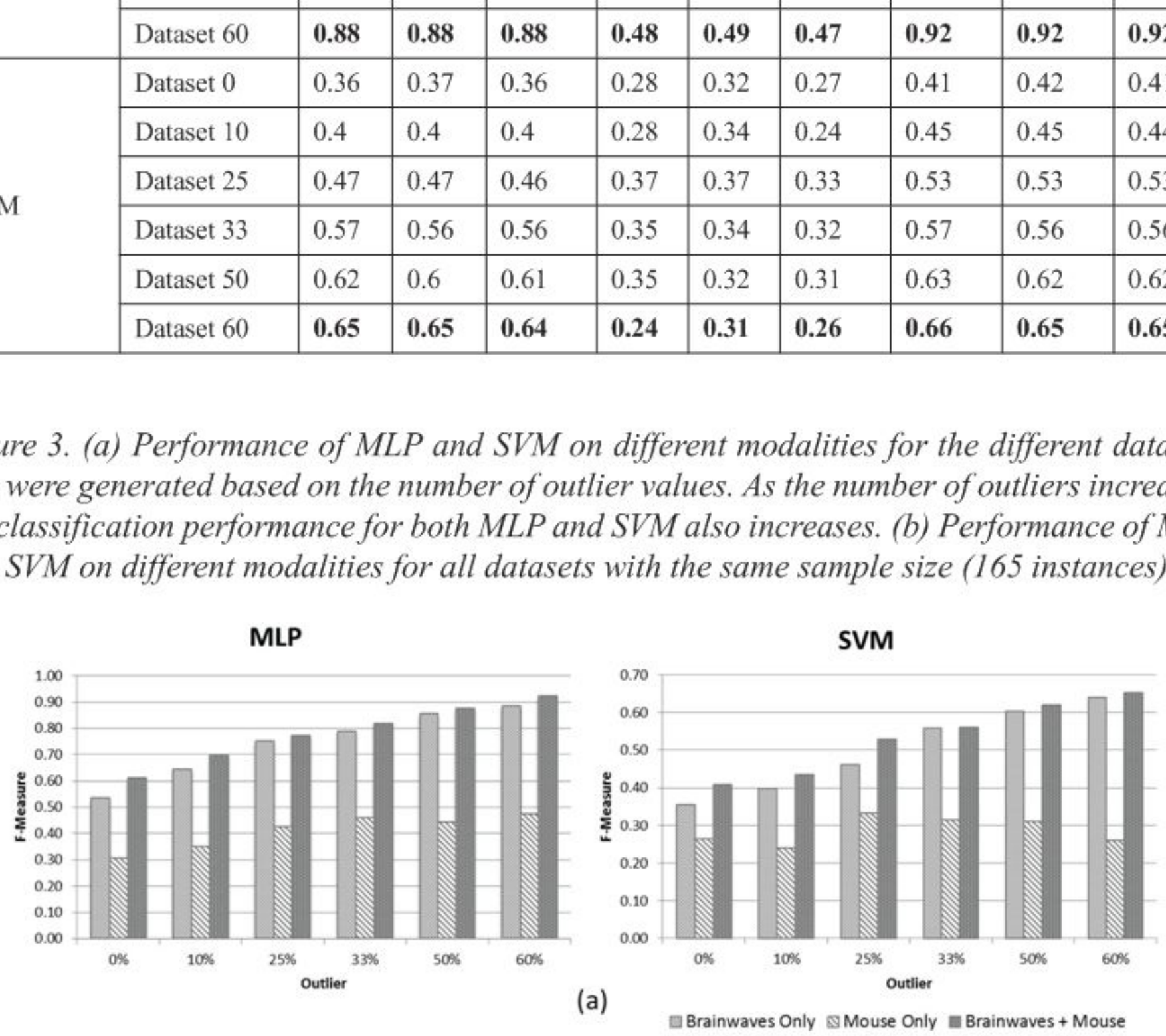
Table 2. Datasets for emotion classification

Dataset	No. of Outlier Features	No. of Students	No. of Instances per Emotion
Dataset 0	0 or more (0%)	16	3,600
Dataset 10	2 or more (10%)	16	2,250
Dataset 25	4 or more (25%)	16	650
Dataset 33	6 or more (33%)	16	325
Dataset 50	8 or more (50%)	16	260
Dataset 60	10 or more (60%)	15	165

Table 3. Average of all emotions for each dataset

Classifier	Dataset	Brainwaves Only			Mouse Only			Brainwaves + Mouse		
		P	R	FM	P	R	FM	P	R	FM
MLP	Dataset 0	0.53	0.54	0.53	0.33	0.33	0.3	0.61	0.61	0.61
	Dataset 10	0.65	0.64	0.64	0.38	0.38	0.35	0.7	0.7	0.7
	Dataset 25	0.75	0.75	0.75	0.42	0.44	0.43	0.77	0.77	0.77
	Dataset 33	0.79	0.79	0.79	0.46	0.46	0.46	0.82	0.82	0.82
	Dataset 50	0.86	0.86	0.86	0.46	0.44	0.44	0.88	0.88	0.88
	Dataset 60	**0.88**	**0.88**	**0.88**	**0.48**	**0.49**	**0.47**	**0.92**	**0.92**	**0.92**
SVM	Dataset 0	0.36	0.37	0.36	0.28	0.32	0.27	0.41	0.42	0.41
	Dataset 10	0.4	0.4	0.4	0.28	0.34	0.24	0.45	0.45	0.44
	Dataset 25	0.47	0.47	0.46	0.37	0.37	0.33	0.53	0.53	0.53
	Dataset 33	0.57	0.56	0.56	0.35	0.34	0.32	0.57	0.56	0.56
	Dataset 50	0.62	0.6	0.61	0.35	0.32	0.31	0.63	0.62	0.62
	Dataset 60	**0.65**	**0.65**	**0.64**	**0.24**	**0.31**	**0.26**	**0.66**	**0.65**	**0.65**

Figure 3. (a) Performance of MLP and SVM on different modalities for the different datasets that were generated based on the number of outlier values. As the number of outliers increases, the classification performance for both MLP and SVM also increases. (b) Performance of MLP and SVM on different modalities for all datasets with the same sample size (165 instances).



tempted at only those instances when a good number of the instances deviate "significantly" from the mean, then the performance would tend to increase. It can be said, however, that the sample sizes of these datasets vary, and the decreasing sample size might in fact be in the reason why the performance increases.

In general, we expect large sample sizes to give better performance rates, so it is quite unlikely that the smaller sample size of datasets 50 and 60, for example, is the reason why the classification performance rates for these datasets are higher. But just the same, in order to establish whether indeed it is the smaller sample size that causes the classification performance rates to increase, new datasets for 0%-50% outliers were generated having the same sample size as that of 60% outlier dataset. Since Dataset 60 (with 60% or more outliers) contains 165 instances, all the other datasets (0, 10, 25, 33, 50) were also randomly sampled so that each would also have 165 samples each. Note that each of the datasets still had all the 16 students, and the random sampling had to do with the selection of instances per student in order to have the uniform sample size of 165. Figure 3(b) shows the performance of MLP and SVM on these datasets with uniform sample size. Table 4 presents the average of all emotions for such datasets. Both Figure 3(a) and Table 3 confirm the same results as Figure 3(b) and Table 4 – that the classification performance is better when restricted to a higher number of outliers. Note that as expected, the results for datasets 0, 10, 25, 33, and 50, in the case of the uniform sample size of 165, are lower than those when all the instances were included (i.e. larger sample sizes).

These findings have serious and important implications. The findings imply that in designing an affective learning system, the system should try to predict academic emotion only at those instances when many of the EEG signals deviate significantly from the baseline values or when the mouse is being handled or clicked somewhat differently. Otherwise, prediction may not be as dependable. In other words, it can be visualized that affective tutoring systems to be designed in the future may have physiological sensors (e.g. EEG, skin conductance, posture, heart rate), as well as tracking systems that monitor mouse usage, and other system-specific logs, so that when these sensors and tracking systems are picking up signals that are

Table 4. Average of all emotions for each dataset (same sample size as dataset 60)

Classifier	Dataset	Brainwaves Only			Mouse Only			Brainwaves + Mouse		
		P	R	FM	P	R	FM	P	R	FM
MLP	Dataset 0	0.40	0.40	0.40	0.29	0.31	0.28	0.40	0.40	0.40
	Dataset 10	0.50	0.50	0.50	0.33	0.35	0.32	0.51	0.52	0.51
	Dataset 25	0.63	0.63	0.63	0.42	0.42	0.41	0.67	0.67	0.67
	Dataset 33	0.73	0.72	0.72	0.47	0.45	0.42	0.74	0.74	0.74
	Dataset 50	0.81	0.81	0.81	0.42	0.45	0.42	0.84	0.84	0.83
	Dataset 60	**0.88**	**0.88**	**0.88**	**0.48**	**0.49**	**0.47**	**0.92**	**0.92**	**0.92**
SVM	Dataset 0	0.33	0.33	0.32	0.31	0.32	0.29	0.37	0.37	0.36
	Dataset 10	0.37	0.37	0.37	0.26	0.34	0.24	0.42	0.43	0.41
	Dataset 25	0.44	0.44	0.43	0.39	0.39	0.35	0.54	0.54	0.54
	Dataset 33	0.54	0.54	0.54	0.35	0.34	0.31	0.54	0.54	0.54
	Dataset 50	0.62	0.58	0.59	0.32	0.34	0.29	0.66	0.64	0.65
	Dataset 60	**0.65**	**0.65**	**0.64**	**0.24**	**0.31**	**0.26**	**0.66**	**0.65**	**0.65**

out of the ordinary, then the emotion prediction system can be launched in order to determine the academic emotion of the user.

Depending on the predicted emotion, the appropriate learning modules are then offered to the user. If the system, for example, determines that the learner is *confused*, but otherwise is still quite engaged, then modules that backtrack a little bit on the subject matter and that provide clarifications may then be presented. When the learner is determined to be *frustrated*, the level of difficulty of the learning task may be lowered, and a diagnostic system may be launched in order to understand better the source of errors so that appropriate remediation may be inserted in the learning session. Or, if the system starts to detect *boredom*, then perhaps the entire interface may be altered, or the nature of the learning task and activity may be altogether revised, such as shifting to a video presentation or animation-enriched tutorial, or a shift to paper-and-pencil drill, or even for the system to suggest "this might be a good time for a short break --- would you like to return in 5 minutes?".

Table 5 gives the details of the *precision*, *recall* and *f-measures* according to specific emotion category for MLP and SVM. The results, indeed, confirm the earlier findings reported in (Azcarraga et al.,, 2011c) which reported only performance measures using MLP. Prediction accuracy indeed increases as the number of outlier feature values increases. Moreover, the tables clearly show that the classification accuracy significantly improves when data from brainwaves and mouse behavior are combined. These conclusions can be made whether using MLP or SVM.

At this point, it is important to stress that as far as the current results are concerned, classifying academic emotions based only on mouse features seems to be ineffective. We were hoping that mouse behavior alone could give better prediction accuracies. Unfortunately, this was not the case. Perhaps there is a need to look for more pertinent mouse click features, or some more complex feature selection and feature transformation need to be done prior to feeding mouse click features for classification.

The results, however, clearly also indicate that when features based on mouse behavior are combined with brainwaves features, classification accuracy based on *F-Measure* significantly improve. Accuracy rates reach to 97% when brainwaves and mouse behavior data are combined in predicting *interest*. And accuracy is really higher for instances with many features (as much as 60%) have with outlier values (where outlier values means the value deviate more than one standard deviation from mean).

SUMMARY AND CONCLUSION

Twenty five (25) undergraduate students were asked to use a math learning software while an EEG sensor was attached to their heads to capture their brainwaves. At the same time, their mouse behavior, such as the number of clicks, duration of each click, and the distance traveled by the mouse, were also automatically captured. Brainwaves were carefully synchronized with the mouse behavior signals. During the experiment, each subject regularly reported the level of each emotion (*confidence, excitement, frustration* and *interest)*. The self-reported emotion serves as the tag for the corresponding data point. From the 25 subjects, the data from only 16 were found to be useful due to some data balancing-related issues which are critical for classifying emotions using the MLP computational model.

Different datasets were generated according to the number of outlier features. A feature value is considered an *outlier* if it exceeds by one standard deviation from the mean of that particular feature and for that particular subject. Using MLP and SVM models in classifying *confidence, excitement, frustration* and *interest*, it is shown that the prediction accuracy based on f-measure increases significantly when instances in a dataset have increased number of

Table 5. Performance of MLP and SVM

Dataset	Emotion	Brainwaves Only			Mouse Only			Brainwaves + Mouse		
		P	R	FM	P	R	FM	P	R	FM
MLP										
Dataset 10	Confidence	0.55	0.58	0.56	0.28	0.21	0.24	0.6	0.61	0.61
	Excitement	0.66	0.64	0.65	0.41	0.17	0.24	0.69	0.68	0.68
	Frustration	0.67	0.58	0.62	0.42	0.38	0.4	0.69	0.69	0.69
	Interest	0.7	0.78	0.74	0.4	0.76	0.52	0.82	0.83	0.82
Dataset 25	Confidence	0.67	0.67	0.67	0.31	0.18	0.22	0.73	0.7	0.71
	Excitement	0.77	0.71	0.74	0.44	0.53	0.48	0.74	0.77	0.76
	Frustration	0.72	0.75	0.73	0.35	0.39	0.37	0.76	0.72	0.74
	Interest	0.84	0.87	0.86	0.59	0.67	0.63	0.84	0.9	0.87
Dataset 33	Confidence	0.78	0.72	0.75	0.39	0.31	0.35	0.76	0.79	0.78
	Excitement	0.71	0.83	0.76	0.37	0.39	0.38	0.78	0.86	0.82
	Frustration	0.77	0.73	0.75	0.39	0.41	0.4	0.83	0.71	0.76
	Interest	0.9	0.87	0.89	0.67	0.73	0.7	0.91	0.92	0.91
Dataset 50	Confidence	0.83	0.8	0.81	0.41	0.41	0.41	0.85	0.8	0.82
	Excitement	0.8	0.89	0.84	0.3	0.25	0.27	0.85	0.94	0.89
	Frustration	0.85	0.8	0.83	0.39	0.52	0.44	0.87	0.87	0.87
	Interest	0.94	0.93	0.94	0.73	0.57	0.64	0.94	0.89	0.91
Dataset 60	Confidence	0.84	0.82	0.83	0.55	0.36	0.44	0.93	0.84	0.89
	Excitement	0.85	0.83	0.84	0.29	0.25	0.27	0.87	0.96	0.91
	Frustration	0.89	0.89	0.89	0.55	0.59	0.57	0.94	0.91	0.93
	Interest	0.95	1	0.98	0.54	0.75	0.62	0.95	0.98	0.97
SVM										
Dataset 10	Confidence	0.37	0.31	0.34	0.38	0.02	0.04	0.44	0.25	0.32
	Excitement	0.39	0.37	0.38	0.03	0	0	0.41	0.37	0.39
	Frustration	0.38	0.44	0.41	0.34	0.71	0.46	0.44	0.59	0.5
	Interest	0.45	0.48	0.46	0.35	0.64	0.45	0.49	0.59	0.54
Dataset 25	Confidence	0.45	0.33	0.38	0.34	0.05	0.09	0.48	0.39	0.43
	Excitement	0.44	0.63	0.52	0.46	0.46	0.46	0.58	0.48	0.52
	Frustration	0.49	0.47	0.48	0.35	0.63	0.45	0.5	0.54	0.52
	Interest	0.51	0.44	0.47	0.33	0.33	0.33	0.57	0.73	0.64
Dataset 33	Confidence	0.51	0.5	0.51	0.38	0.14	0.2	0.49	0.5	0.49
	Excitement	0.52	0.44	0.48	0.32	0.26	0.29	0.54	0.44	0.48
	Frustration	0.49	0.6	0.54	0.34	0.71	0.46	0.49	0.59	0.53
	Interest	0.75	0.69	0.72	0.37	0.28	0.32	0.76	0.73	0.74

continued on following page

Table 5. Continued

Dataset	Emotion	Brainwaves Only			Mouse Only			Brainwaves + Mouse		
		P	R	FM	P	R	FM	P	R	FM
Dataset 50	Confidence	0.41	0.47	0.44	0.49	0.18	0.26	0.45	0.49	0.47
	Excitement	0.54	0.6	0.57	0.21	0.22	0.21	0.52	0.58	0.55
	Frustration	0.67	0.58	0.62	0.34	0.42	0.38	0.69	0.57	0.62
	Interest	0.86	0.74	0.8	0.34	0.47	0.39	0.86	0.82	0.84
Dataset 60	Confidence	0.67	0.51	0.58	0.33	0.23	0.27	0.63	0.66	0.64
	Excitement	0.51	0.62	0.56	0.29	0.56	0.38	0.57	0.58	0.57
	Frustration	0.63	0.52	0.57	0	0	0	0.49	0.5	0.5
	Interest	0.79	0.95	0.86	0.35	0.44	0.39	0.95	0.87	0.91

feature values that deviate significantly from the mean values of a given subject. These important findings imply that in designing an affective learning system, the system should try to predict academic emotion only at those instances when many of the EEG signals deviate significantly from the baseline values or when the mouse is being handled or clicked somewhat differently. Otherwise, prediction may not be dependable.

It can be imagined that in the future, affective tutoring systems would have physiological sensors (e.g. that monitor EEG, skin conductance, posture, heart rate), as well as tracking systems that monitor mouse usage, and other system-specific logs, so that when these sensors and tracking systems are picking up signals that are out of the ordinary, then the emotion prediction system can be launched in order to determine the academic emotion of the user. Depending on the predicted emotion, the appropriate learning modules are then offered to the user.

Moreover, the results clearly show that when combining the extracted features from EEG signals with mouse click behavior, the accuracy in predicting academic emotions is significantly better than when using only features extracted from EEG signals or just from mouse behavior alone. Future work would include the validation of the outlier detection approach on datasets with brainwaves features in the form frequency waves (i.e. alpha, beta, gamma) and probably, with other features from system logs and user profile such as personality type, hand dominance and intelligence level.

REFERENCES

Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. M. (2009). Emotion sensors go to school. In V. Dimitrova, R. Mizoguchi, B. Du Boulay, & A. Graesser (Eds.), *Artificial intelligence in education* (Vol. 200, pp. 17–24). IOS Press.

Azcarraga, J., Ibañez, J. F., Jr., Lim, I. R., & Lumanas, N., Jr. (2011a, March). Predicting student affect based on brainwaves and mouse behavior. In *Proceedings of the 11th Philippine Computing Science Congress,* Naga City, Philippines.

Azcarraga, J., Ibañez, J. F., Jr., Lim, I. R., Lumanas, N., Jr., Trogo, R., & Suarez, M. T. (2011c). Predicting academic emotion based on brainwaves signals and mouse click behavior. In T. Hirashima et al., (Eds.), *Proceedings of the 19th International Conference on Computers in Education* (pp. 42-49). Chiang Mai, Thailand: Asia-Pacific for Computers in Education.

Azcarraga, J., Inventado, P. S., & Suarez, M. T. (2010). Predicting the difficulty level faced by academic achievers based on brainwaves analysis. In *the Proceedings of the 18th International Conference on Computers in Education* (pp. 107-109). Putrajaya, Malaysia: Asia-Pacific for Computers in Education.

Azcarraga, J. J., Ibañez, J. F., Jr., Lim, I. R., & Lumanas, N., Jr. (2011b). Use of personality profile in predicting academic emotion based on brainwaves signals and mouse behavior. In *the Proceedings of the 2011 Third International Conference on Knowledge and Systems Engineering* (pp. 239-244). Hanoi, Vietnam.

Burleson, W. (2006). *Affective learning companions: Strategies for empathetic agents with real-time multimodal affective sensing to foster meta-cognitive and meta-affective approaches to learning, motivation, and perseverance.* Unpublished Doctoral Dissertation, Massachusetts Institute of Technology.

Chanel, G. (2009). *Emotion assessment for affective computing based on brain and peripheral signals.* Unpublished Doctoral Dissertation. University of Geneva.

Cherkassky, V., & Mulier, F. (2007). *Learning from data: Concepts, theory, and methods* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc. doi:10.1002/9780470140529.

D'Mello, S. K., & Graesser, A. (2009). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, *20*(2), 147–187. doi:10.1007/s11257-010-9074-4.

Davidson, R. J., Ekman, P., Saron, C. D., Senulis, J. A., & Friesen, W. V. (1990). Approach-withdrawal and cerebral asymmetry: emotional expression and brain physiology. *Journal of Personality and Social Psychology*, *58*(2), 330–341. doi:10.1037/0022-3514.58.2.330 PMID:2319445.

Ekman, P. (1984). Expression and the nature of emotion. In K. Scherer, & P. Ekman (Eds.), *Approaches to Emotion* (pp. 319–344). Hillsdale, NJ: Erlbaum.

Frantzidis, C. A., Bratsas, C., Klados, M. A., Konstantinidis, E., Lithari, C. D., & Vivas, A. B. et al. (2010). On the classification of emotional biosignals evoked while viewing affective pictures: An integrated data-mining-based approach for healthcare applications. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 309–318. doi:10.1109/TITB.2009.2038481 PMID:20064762.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations Newsletter*, *11*(1), 10–18. doi:10.1145/1656274.1656278.

Haykin, S. (2008). *Neural networks and learning machines*. Upper Saddle River, NJ: Pearson Prentice Hall.

Heraz, A., Jraidi, I., Chaouachi, M., & Frasson, C. (2009). Predicting stress level variation from learner characteristics and brainwaves. In V. Dimitrova, R. Mizoguchi, B. Du Boulay, & A. C. Graesser (Eds.), *Artificial intelligence in education* (Vol. 200, pp. 722–724). Brighton, UK: IOS Press.

Heraz, A., Razaki, R., & Frasson, C. (2007). Using machine learning to predict learner emotional state from brainwaves. In *the Proceedings of the Seventh IEEE International Conference on Advanced Learning Technologies,* (pp. 853-857). IEEE Computer Society.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. doi:10.1016/0893-6080(89)90020-8.

Ibanez, J. F., Jr., Lim, I. R., & Lumanas, N., Jr. (2011). *Affect recognition using brainwaves and mouse behaviour for intelligent tutoring systems.* Unpublished Undergraduate Thesis. De La Salle University, Manila, Philippines.

Kaklauskas, A., Zavadskas, E. K., Seniut, M., Dzemyda, G., Stankevic, V., & Simkevicius, C. et al. (2011). Web-based biometric computer mouse advisory system to analyze a user's emotions and work productivity. *Engineering Applications of Artificial Intelligence*, *24*(6), 928–945. doi:10.1016/j.engappai.2011.04.006.

Kapoor, A., Burleson, W., & Picard, R. (2007). Automatic prediction of frustration. *International Journal of Human-Computer Studies*, *65*(8), 724–736. doi:10.1016/j.ijhcs.2007.02.003.

Nicaud, J.-F., Bouhineau, D., & Huguet, T. S. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), Lecture notes in computer science: Vol. 2363. *(n.d.). The aplusix-editor: A new kind of software for the learning of algebra* (pp. 178–187). Berlin/Heidelberg, Germany: Springer.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, *37*(2), 91–105. doi:10.1207/S15326985EP3702_4.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318–362). MIT Press.

Sanei, S., & Chambers, J. A. (2007). EEG signal processing. West Sussex, UK: John Wiley-& Sons, Ltd.

Scheirer, J., Fernandez, R., Klein, J., & Picard, R. W. (2001). Frustrating the user on purpose: A step toward building an affective computer. *Interacting with Computers*, *14*(2), 93–118. doi:10.1016/S0953-5438(01)00059-5.

Stevens, R., Galloway, T., & Berka, C. (2007). EEG-related changes in cognitive workload, engagement and distraction as students acquire problem solving skills. In C. Conati, K. McCoy, & G. Paliouras (Eds.), User modeling 2007, 4511, 187–196. Springer.

Tomarken, A. J., Davidson, R. J., Wheeler, R. E., & Doss, R. C. (1992). Individual differences in anterior brain asymmetry and fundamental dimensions of emotion. *Journal of Personality and Social Psychology*, *62*(4), 676–687. doi:10.1037/0022-3514.62.4.676 PMID:1583591.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY: Springer. doi:10.1007/978-1-4757-2440-0.

Zimmermann, P., Guttormsen, S., Danuser, B., & Gomez, P. (2003). Affective computing--a rationale for measuring mood with mouse and keyboard. *International Journal of Occupational Safety and Ergonomics*, *9*(4), 539–551. PMID:14675525.

Judith Azcarraga is studying towards a PhD in Computer Science student at the College of Computer Studies of De La Salle University (DLSU), in Manila, Philippines. Under a scholarship from the PCIEERD of the Department of Science and Technology, she is conducting her research at DLSU's Center for Empathic-Human Computer Interactions. Her doctoral thesis is on the recognition of academic emotions of intellectually-gifted students based on the pattern of their brainwaves, mouse behavior and their personality profile. She has been working on learning systems for children since her undergraduate thesis and Master's thesis. She has been an instructor at the College of Computer Studies of DLSU and a TAFE school in Singapore.

Merlin Teodosia Suarez is an Associate Professor of Computer Science at the College of Computer Studies of De La Salle University (DLSU), in Manila, Philippines. She heads the DLSU's Center for Empathic-Human Computer Interactions. She obtained her PhD from DLSU. Her dissertation investigated how a bug library for novice Java programmers can be built automatically using machine learning techniques. She serves as a member of the Department of Science and Technology's PCIEERD Technical Panel on Information and Communications Technology. She organized the 1st and 2nd International Workshop on Empathic Computing (IWEC-10 and IWEC-11) as their co-chairs. She is in the steering committee for the International Workshop on Empathic Computing (IWEC).