ECG-based affective computing for difficulty level prediction in intelligent tutoring systems

Alqahtani, Fehaid; Katsigiannis, Stamos; Ramzan, Naeem

Published in:
4th International Conference on UK - China Emerging Technologies (UCET)

DOI:
10.1109/UCET.2019.8881872

Published: 24/10/2019

Document Version
Peer reviewed version

ECG-based affective computing for difficulty level prediction in Intelligent Tutoring Systems

Fehaid Alqahtani
[1] *School of Computing, Engineering and Physical Sciences*
University of the West of Scotland
Paisley, United Kingdom,
[2] *Computer Science Department*
King Fahad Naval Academy
Jubail 35512, Kingdom of Saudi Arabia
Fehaid.Alqahtani@uws.ac.uk

Stamos Katsigiannis, Naeem Ramzan
School of Computing, Engineering and Physical Sciences
University of the West of Scotland
Paisley, United Kingdom
Stamos.Katsigiannis@uws.ac.uk, Naeem.Ramzan@uws.ac.uk

***Abstract*—Intelligent tutoring Systems (ITS) have emerged as an attractive solution for providing personalised learning experiences on a large scale. Traditional ITS are able to adapt the learning process according to the capabilities and needs of their users, but lack the capability to adapt to their affective/emotional state. In this work, we examine the use of electrocardiography (ECG) signals for detecting the affective state of ITS users. Features, extracted from ECG signals acquired while users undertook a computerised English language test, were used for the prediction of the self-reported difficulty level of the test's questions. Supervised classification experiments demonstrated the potential of this approach, achieving a classification F1-score of 61.22% for the prediction of the self-assessed difficulty level of the questions.**

***Index Terms*—Intelligent Tutoring Systems (ITS), Affective computing, ECG, Physiological Signals, Machine learning**

I. INTRODUCTION

Intelligent Tutoring System (ITS) are systems that can offer a personalised learning experience by adapting the learning process to the performance, knowledge, capabilities, and needs of their users [1]–[3]. One of the disadvantages of traditional ITS is that they lack the ability to infer the affective state of the learner and adapt the learning process according to it, as a human instructor would be able to [1]. Affective Tutoring Systems attempt to address this drawback by following an adaptive approach to learning through detecting the affective state of the user and using this information along with the knowledge level and the abilities of the learner for adapting the learning process [4], [5]. Traditional tutoring systems can also be adapted to the user requirements but they are more adaptable than adaptive, the difference lying in the fact that adaptable systems need user input to set the customisation parameters, while adaptive or intelligent systems can do it automatically [4], [5].

Research on the process of learning has demonstrated that the emotion of the learners has an overall impact on the learning ability, as negative emotions can severely impair the learning process, while positive emotions can provide a boost to it [6]. Andres et al. [7] studied the impact of the emotional state on the learning process and concluded that boredom is a powerful indicator of knowledge but not always indicative of learning, while delight is a powerful indicator for inquisitiveness to learn but not necessarily of knowledge. Within the context of ITS, Bosch and D'Mello conducted a study [8], where they concluded that emotions, namely engagement, confusion, frustration, boredom, and curiosity are the most frequently found affective states of learners. In addition, they also concluded that confusion and frustration, as well as curiosity and engagement, are the most frequently found pairs of emotional states that co-occur in learners.

Advances in the field of *Affective Computing* have provided promising tools for enhancing ITS with affective state detection capabilities. Affective computing is defined as "*computing that relates to, arises from, or deliberately influences emotions*" [9], with one of its main applications being emotion recognition [10]. Researchers have been trying to observe and model the behaviour of the human brain through diverse areas of research, such as psychology [11], [12], medical imaging, such as functional Magnetic Resonance Imaging (fMRI) [13]), or by analysing bio-signals, such as physiological signals [14], [15]. Various studies [14]–[18] have shown that there is correlation between physiological signals and the emotions felt by an individual, as defined in Russel's *Circumplex Model of Affect* [19]. These studies relied on pattern recognition and machine learning techniques in order to map features extracted from physiological signals to a relevant emotional state in terms of the Valence and Arousal dimensions of an emotion. Furthermore, the current widespread availability of wearable wireless non-invasive sensors for physiological signals, such as electrocardiography (ECG), electroencephalography (EEG), electromyography (EMG), etc. [20], [21], allows the use of such techniques in practice.

In this work, we examined the potential use of ECG signals as a means for the detection of the affective state of learners, while they are undertaking a computerised English language test. Features extracted from the ECG signals were used in order to train machine learning models for the task of predicting the self-reported difficulty level of the questions in the test, as perceived by the test takers. The examined task would be of significant interest in the field of ITS, since it would allow ITS to be able to adapt according to the affective state of the users, thus providing a more personalised learning experience. Forty five participants with various levels of English language knowledge participated in this study. Supervised classification experiments were conducted using various classification algorithms, achieving a classification F1-score of 61.22% for the task of predicting the self-reported difficulty level of the test's questions.

The rest of the paper is organised into three sections. Section II describes the methodology followed, while Section III presents the acquired results. Finally conclusions are drawn in Section IV.

II. Methodology

A. Experimental protocol & Participants

The experiment was performed in a quiet environment where the noise level was low, and no external disturbances were present. This was to ensure that the affective state of the learners was not impacted by anything other than the stimulus. Before each session, the experimental procedure was thoroughly explained to all participants and they were given the opportunity to ask questions or raise any concerns that they might have. After signing a consent form, the four electrodes of the ECG sensor were attached to both lower ribs and clavicle of the participants, who were then asked to sit on a chair in front of a desk with a computer.

ECG signals measure the electrical activity of the heart and were obtained during the whole duration of the experiment. A SHIMMER™ v2 wireless sensor [21] (Fig. 1) was used for the acquisition of the ECG signals at a 256 Hz sampling rate, recording a 2-channel signal (RA $\rightarrow$ LL and LA $\rightarrow$ LL) from which only the RA $\rightarrow$ LL channel was used for further analysis. The SHIMMER™ sensor was selected due to its portability and low weight, in order to minimise any level of discomfort to the participants. Furthermore, a laptop computer was used for signal recording and monitoring. Participants were instructed to avoid unnecessary body movements in order to avoid any effects on the recorded ECG signal. Furthermore, they were also instructed to refrain from consuming caffeine or drugs prior to the experiment, since they could also affect the ECG recording. After verifying correct signal acquisition, participants were instructed to commence the experiment.

During the experiment, participants were asked to complete a computerised English language test using the computer mouse to respond to test questions. Twenty questions were obtained from the Oxford Quick Placement Test (QPT) [22], which contains forty questions of varying difficulty, designed to test the English knowledge of test takers and assign them to various levels according to the Common European Framework of Reference for languages (CEFR) [23] for assessing foreign language skills. The questions selected for the experiment focused on four different tasks and five questions were randomly selected from each task. The first task examined the ability of test takers to use phrase forms for understanding the meaning of short notices. The second task examined the level of grammatical knowledge. The third task the knowledge of pragmatic meaning and linguistic contextual information, and the fourth task the level of grammar and vocabulary and whether the participants can understand long text passages. The test questions were presented to the participants beginning with Task 1 and ending with Task 4.

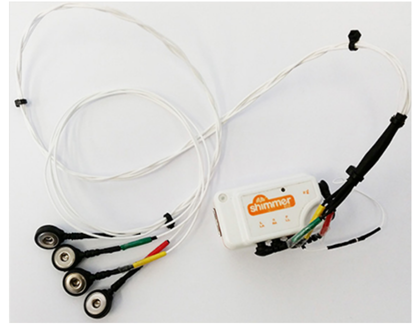

Fig. 1: The SHIMMER™ v2.0 wireless ECG sensor

After answering each question, participants were asked to assess its difficulty as *Very easy*, *Easy*, *Moderate*, *Hard*, or *Very hard*, by clicking on the respective radio button. After answering all the test's questions and providing the feedback regarding the difficulty, the experiment finished. Each participant was then automatically assigned to an English language knowledge level according to the percentage of questions that were answered correctly, as follows: *Poor* (0-50%), *Beginner* (50-60%), *Elementary* (60-70%), *Intermediate* (70-80%), *Advanced* (80-90%), and *Expert* (90-100%).

In total, 45 individuals participated in this study, from who 34 were male and 11 female, aged between 16 and 47 years old ($\mu_{\text{age}} = 28.1$, $\sigma_{\text{age}} = 6.0$). Participants were recruited among international students from the University of the West of Scotland or from the areas of Paisley and Glasgow, Scotland, United Kingdom. Prerequisites for participation in the study was being healthy and being familiar with using a computer for simple tasks. No time limit was set for answering the test's questions, thus the duration of the experiment varied across the participants, having an average duration $\mu_{\text{duration}} = 415.45$ s ($\sigma_{\text{duration}} = 118.93$ s), with the minimum and maximum duration being 221.47 s and 734.33 s respectively. As shown in Fig. 2, out of the 45 participants, none were assigned to the *Expert* level and only two to *Poor* level, while the majority of participants were assigned to at least *Elementary* level. Furthermore, it is evident from Fig. 3 that, as expected, the percentage of correctly answered questions decreased as the self-assessed difficulty level increased.

B. Signal pre-processing & Feature extraction

ECG signals were captured in a single continuous recording for the whole duration of each session. The timestamps obtained from the test, representing the start and end of each question were then used to segment the signals in relation to each question. As a result, for every participant, 20 signal segments were generated, for a total of 900 segments (45 participants $\times$ 20 questions), each one associated with

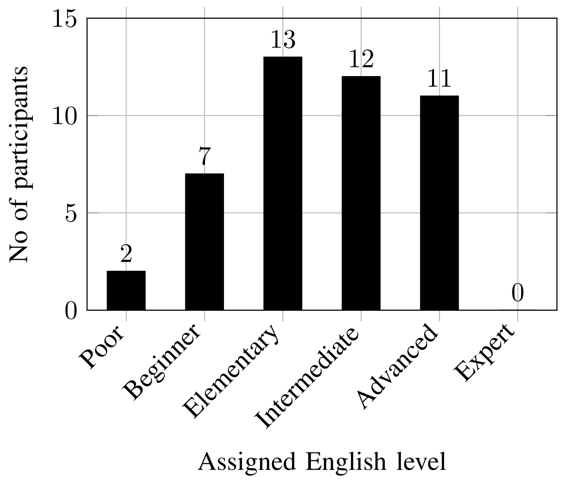

Fig. 2: Distribution of assigned English level.

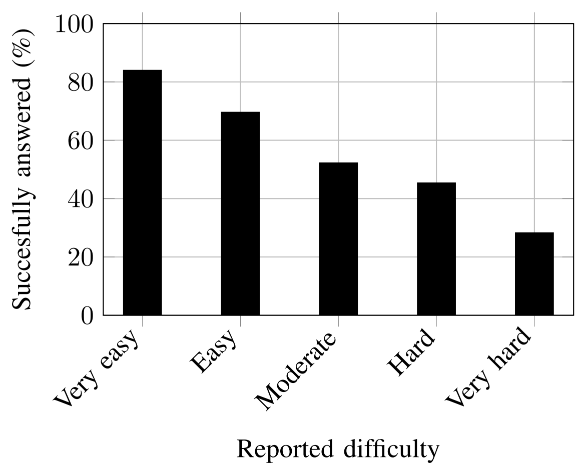

Fig. 3: Successfully answered questions per reported difficulty.

its respective difficulty level, as reported by the respective participant. Then, in order to reduce the effects of noise, the ECG signals were pre-processed by removing baseline wander, applying first a median filter with a 200 ms window, followed by a median filter with a 600 ms window, and by subtracting the filtered signal from the original signal [24].

After pre-processing, 84 features related to the heart rate (HR) and heart rate variability (HRV) were extracted from each ECG signal segment in order to create and evaluate the machine learning models. Multiple research works have shown that there is a relation between ECG features and the affective state of humans [15], [16]. To this end, the Augsburg Biosignal Toolbox (AuBT) [25] was used to detect QRS complexes and R-peaks within the ECG signal, and the following features were extracted: from the raw signal and the derivative of PQ, QS and ST complexes the maxima, minima, mean, median, standard deviation and range, then the maxima, minima, mean, median, standard deviation and range of the HRV histogram, the power spectral density (PSD) of HRV between the intervals $[0\ , 0.2]$, $[0.2\ , 0.4]$, $[0.4\ , 0.6]$ and $[0.6\ , 0.8]$, and the number of intervals with latency > 50 ms from HRV. Finally, the 84 computed features were concatenated in order to create the final feature vector for each signal segment.

III. Experimental results

The extracted features were used in order to train machine learning models for the task of predicting the self-assessed difficulty level of the test's questions. The problem was transformed to a binary classification problem by grouping the self-assessed difficulty level to two categories, i.e. *Low* and *High*, as typically done in affective computing studies (e.g. [14], [15]). Samples associated with *Very easy* and *Easy* difficulty were assigned to the *Low* difficulty category (554 samples), while samples associated with *Hard* and *Very hard* difficulty to the *High* difficulty category (143 samples). The 203 samples referring to *Moderate* difficulty were discarded since they could not be assigned to either category. As a result, 697 out of the 900 samples were used for further analysis, leading to an unbalanced dataset, with 79.5% of the used samples referring to the *Low* and 20.5% to the *High* category.

Supervised classification experiments were conducted using various classification algorithms: k-Nearest Neighbour (kNN) for $k = 1, 3, 5$, Linear Support Vector Machines (SVM), SVM with the radial basis function (RBF) kernel, Linear Discriminant Analysis (LDA), and Decision Trees (DT). To avoid over-fitting the models due to having samples from the same participants in both the training and test sets, a *Leave-One-Subject-Out* (LOSO) cross-validation process was applied. At each iteration of the cross-validation, all samples related to a specific subject were used for testing the model while all the other samples were used for training. After repeating the process for all subjects, the average performance across all iterations was reported as the overall performance of the model. The acquired results are reported in Table I, in terms of Accuracy and F1-score. The F1-score is the harmonic mean of Precision and Recall and provides a better classification performance metric than accuracy in cases of uneven class distribution. Since the F1-score depends on the class that is considered as positive, the F1-scores in Table I were computed as the average F1-scores between the examined classes.

As shown in in Table I, the 3-NN classifier provided the highest classification F1-score (61.22%), with the LDA classifier providing marginally worse results (61.13%). It must be noted that although classification accuracy is reported for reference, it must not be taken into consideration due to the highly unbalanced dataset. To examine the statistical significance of the acquired results, the expected classification results for random voting (50% class probability), class ratio-based voting (class probability equal to its ratio of samples), and majority class voting (100% probability of the majority class) were analytically computed, as shown in Table I. An unpaired Kruskal-Wallis test comparing the predicted class labels for each classifier to the predicted class labels for random voting and class ratio-based voting showed that all classifiers performed significantly better that random voting

TABLE I: Classification performance for the prediction of self-assessed question difficulty, in terms of Accuracy (%) and F1-score (%).

Classifier	Accuracy (%)	F1-score (%)	Significance
LDA	77.47	61.13	$\star\dagger\ddagger$
DT	73.31	60.69	$\star\dagger\ddagger$
SVM (RBF)	79.48	44.28	$\star\ \ddagger$
SVM (Linear)	78.34	59.70	$\star\dagger\ddagger$
1-NN	69.73	55.72	$\star\dagger\ddagger$
3-NN	76.18	**61.22**	$\star\dagger\ddagger$
5-NN	76.47	59.52	$\star\dagger\ddagger$
Random	50.00	46.57	-
Majority	75.35	42.97	-
Class ratio	62.85	50.00	-

*†‡Statistically significant difference compared to random voting ($\star$), $p \leq 4.7 \cdot 10^{-25}$, majority voting ($\dagger$), $p \leq 3.29 \cdot 10^{-20}$, and voting according to the class ratio ($\ddagger$), $p \leq 2.8 \cdot 10^{-8}$.

and class ratio-based voting ($p \leq 4.7 \cdot 10^{-25}$ and $p \leq 2.8 \cdot 10^{-8}$ respectively). A paired Wilcoxon signed-rank test was used to test the significance against majority voting since the predicted class labels could be compared one-by-one. All classifiers performed significantly better than majority voting ($p \leq 3.29 \cdot 10^{-20}$), apart from the SVM (RBF) classifier.

It must be noted that to the best of the authors' knowledge, this is the first attempt to predict the self-assessed difficulty level of a test question based on the test takers ECG recordings, thus a comparative study is not provided. The acquired classification results demonstrate the potential of using ECG signals for detecting the perceived difficulty level of a test question. The proposed approach could enhance ITS by allowing them to adapt the learning process according to how difficult the content is perceived by their users.

IV. Conclusion

This work studied the use of ECG signals for detecting the affective state of test takers participating in a computerised English language test. Features extracted from the ECG signals were used in order to train machine learning models for the prediction of the self-assessed difficulty level of the test's questions, as reported by the 45 participants of the study. Experimental results showed that the spatial and spectral ECG features used in combination with the 3-NN classifier led to a 61.22% classification F1-score, demonstrating the potential of the proposed approach for the examined task. The obtained results suggest that ECG signals can be potentially used by ITS for personalisation and adaptation of the learning content based on the affective state of the learner. Future work will include the study of additional physiological signals, such as EEG, EMG, etc., for the prediction of the perceived difficulty level of a test's question, as well as the evaluation of single-subject machine learning models for enhancing the personalisation capabilities of ITS.

References

[1] X. Mao and Z. Li, "Agent based affective tutoring systems: A pilot study," *Computers & Education*, vol. 55, no. 1, pp. 202 – 208, 2010.

[2] S. Kuyoro, G. Maminor, R. Kanu, and O. Akande, "The design and implementation of a computer based testing system," *History*, vol. 5, p. 6, 2016.

[3] F. Alqahtani and N. Ramzan, "Comparison and efficacy of synergistic intelligent tutoring systems with human physiological response," *Sensors*, vol. 19, no. 3, p. 460, 2019.

[4] N. Tsianos, Z. Lekkas, P. Germanakos, C. Mourlas, and G. Samaras, "An experimental assessment of the use of cognitive and affective factors in adaptive educational hypermedia," *IEEE Transactions on Learning Technologies*, vol. 2, no. 3, pp. 249–258, July 2009.

[5] K. Kiili and H. Ketamo, "Evaluating cognitive and affective outcomes of a digital game-based math test," *IEEE Transactions on Learning Technologies*, vol. 11, no. 2, pp. 255–263, April 2018.

[6] M. B. Ammar, M. Neji, A. M. Alimi, and G. Gouardres, "The affective tutoring system," *Expert Systems with Applications*, vol. 37, no. 4, pp. 3013 – 3023, 2010.

[7] J. M. A. L. Andres, J. Ocumpaugh, R. S. Baker, S. Slater, L. Paquette, Y. Jiang, S. Karumbaiah, N. Bosch, A. Munshi, A. Moore, and G. Biswas, "Affect sequences and learning in betty's brain," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. New York, NY, USA: ACM, 2019, pp. 383–390.

[8] N. Bosch and S. D'Mello, "The affective experience of novice computer programmers," *International Journal of Artificial Intelligence in Education*, vol. 27, no. 1, pp. 181–206, Mar 2017.

[9] R. Picard, "Affective computing," MIT Media Laboratory Perceptual Computing Section, Tech. Rep. 321, 1995.

[10] R. W. Picard, "Affective computing: from laughter to IEEE," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 11–17, Sep. 2010.

[11] T. R. Lynch, A. L. Chapman, M. Z. Rosenthal, J. R. Kuo, and M. M. Linehan, "Mechanisms of change in dialectical behavior therapy: Theoretical and empirical observations," *Journal of clinical psychology*, vol. 62, no. 4, pp. 459–480, 2006.

[12] M. I. Posner and M. K. Rothbart, *Educating the human brain.* American Psychological Association, 2007.

[13] C. Hattingh, J. Ipser, S. Tromp, S. Syal, C. Lochner, S. Brooks, and D. Stein, "Functional magnetic resonance imaging during emotion recognition in social anxiety disorder: an activation likelihood meta-analysis," *Frontiers in Human Neuroscience*, vol. 6, p. 347, 2013.

[14] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, 2012.

[15] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 98–107, 2018.

[16] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, 2012.

[17] P. Arnau-González, M. Arevalillo-Herráez, and N. Ramzan, "Fusing highly dimensional energy and connectivity features to identify affective states from eeg signals," *Neurocomputing*, vol. 244, pp. 81–89, 2017.

[18] T. Althobaiti, S. Katsigiannis, D. West, and N. Ramzan, "Examining human-horse interaction by means of affect recognition via physiological signals," *IEEE Access*, vol. 7, pp. 77 857 – 77 867, Jun. 2019.

[19] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[20] N. A. Badcock, P. Mousikou, Y. Mahajan, P. de Lissa, J. Thie, and G. McArthur, "Validation of the Emotiv EPOC EEG gaming system for measuring research quality auditory ERPs," *PeerJ*, vol. 1, no. e38, Feb. 2013.

[21] A. Burns, B. R. Greene, M. J. McGrath, T. J. O'Shea, B. Kuris, S. M. Ayer, F. Stroiescu, and V. Cionca, "SHIMMER - A Wireless Sensor Platform for Noninvasive Biomedical Research," *IEEE Sens. J.*, vol. 10, pp. 1527–1534, Sept. 2010.

[22] Oxford University Press, *Quick Placement Test*. Oxford University Press, 2001.

[23] Council of Europe, "Common European Framework of Reference for Languages: Learning, Teaching, Assessment," 2011.

[24] N. Kannathal, U. R. Acharya, K. P. Joseph, L. C. Min, and J. S. Suri, "Analysis of electrocardiograms," in *Advances in Cardiac Signal Processing*. Springer, 2007, pp. 55–82.

[25] J. Wagner, "Augsburg biosignal toolbox (aubt)," *University of Augsburg*, 2005.