On the use of ECG and EMG signals for question difficulty level prediction in the context of Intelligent Tutoring Systems

Alqahtani, Fehaid; Katsigiannis, Stamos; Ramzan, Naeem

Published in:
2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)

DOI:
10.1109/BIBE.2019.00077

Published: 31/10/2019

Document Version
Peer reviewed version

Citation for published version (APA):
Alqahtani, F., Katsigiannis, S., & Ramzan, N. (2019). On the use of ECG and EMG signals for question difficulty level prediction in the context of Intelligent Tutoring Systems. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 392-396). (IEEE Conference Proceedings). IEEE. https://doi.org/10.1109/BIBE.2019.00077

On the use of ECG and EMG signals for question difficulty level prediction in the context of 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*—A fundamental drawback of traditional Intelligent Tutoring Systems (ITS) is that, unlike human tutors, they are not able to understand the emotional state of their users and adapt the learning process accordingly. This work explores the potential use of affective computing techniques for providing an affect detection mechanism for ITS. Electrocardiography (ECG) and electromyography (EMG) signals were recorded from 45 individuals that undertook a computerised English language test and provided feedback on the difficulty of the test's questions. Features extracted from the ECG and EMG signals were then used in order to train machine learning models for the task of predicting the self-perceived difficulty level of the questions. The conducted supervised classification experiments provided promising results for the suitability of this approach for enhancing ITS with information relating to the affective state of the learners, reaching an average classification F1-score of 75.49% for the personalised single-participant models and a classification F1-score of 64.10% for the global models.**

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

I. Introduction

Advances in the field of Intelligent Tutoring Systems (ITS) have provided alternatives to traditional teaching approaches by requiring minimal input from tutors and by being able to provide immediate and personalised feedback to the learner. ITS have the ability to adapt the learning process based on information gathered about the learner in terms of his/her capabilities, prior knowledge, performance, and needs [1]–[3]. Nevertheless, ITS face some challenges in relation to traditional teaching methods. Ma et al. [4] conducted a review on studies about ITS and reached the conclusion that ITS are still not mature enough to completely replace traditional teaching and learning practices although they have been proven to be very effective in many areas. Another significant drawback pointed out by Nye [5] is that typical ITS are more suitable for developed countries and their use in developing countries faces many challenges.

Many of the challenges that ITS face stem from the lack of direct learner-tutor interaction. Human instructors are able to infer the affective/emotional state of a learner and adapt the learning and teaching process accordingly, an ability that traditional ITS lack [1]. Approaches to address this challenge focus on using methods for detecting the affective state of the learner and attempting to adapt the learning process using the affective information in combination with information such as knowledge level, performance, etc. [6], [7]. This approach led to the creation of a new type of ITS, called Affective Tutoring Systems, based on the belief that emotions and affective state are fundamental for thinking and learning [1], [8]–[10].

Ben Ammar et al. [8] showed that the learning process can be negatively affected by negative emotions, while on the other hand, positive emotions can significantly assist the learning process. Research on the effect of specific emotions on the learning process by Andres et al. [11] came to the conclusion that delight is a very strong indicator for inquisitiveness to learn but is not necessarily indicative of knowledge, contrary to boredom, which was shown to be a strong indicator of knowledge but not always indicative of learning. A study by Bosch and D'Mello [12] on the most frequent emotions of learners within an ITS context showed that curiosity, boredom, engagement, frustration, and confusion are the most common affective states of learners. The same study also concluded that engagement and curiosity, along with frustration and confusion are the most frequent pairs of emotions co-occurring in learners within an ITS context.

Considering the effect of affect in the ability of learners to learn, Affective Tutoring Systems utilise affect detection methods, such as facial expressions [8], facial features combined with neural networks [13], facial and voice features [14], etc., in order to detect the affective state of the learner and adapt the learning process accordingly. A field of research that can potentially provide solutions for the detection of affect in ITS is the field of Affective Computing. Affective

Computing refers to computing that relates to emotions [15] and emotion/affect recognition is one of its main focuses [16]. Within this field, multiple studies [17]–[21] explored the relation between physiological signals, such as electroencephalography (EEG), electrocardiography (ECG), electromyography (EMG), electrodermal activity (EDA), and others, and the affective state of humans in terms of the Valence and Arousal dimensions, as defined by Russel in his *Circumplex Model of Affect* [22].

In this work, we conducted a proof-of-concept study in order to examine the use of ECG and EMG signals acquired using wearable wireless off-the-shelf sensors for the task of detecting the affective state of learners participating in a computerised English language test. ECG and EMG signals were recorded for the whole duration of the test, and were used for the extraction of spatial and spectral features that were used to train machine learning models for the task of predicting the difficulty level of the test's questions, as perceived by the users. Predicting the perceived difficulty level of a question would allow an ITS to provide a personalised learning experience to a user by being able to adapt its difficulty and provide feedback relevant to the needs of the user. Forty five individuals participated in this proof-of-concept study and results from the conducted supervised classification experiments reached a classification F1-score of 64.10% for multi-participant (global) models and an average classification F1-score of 75.49% for single-participant (personalised) models for the task of predicting the questions' difficulty level.

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

II. Methodology

A. Experimental protocol

The experiment took place within a quite environment with low noise levels and no external disturbances, in order to ensure that the affective state of the participants would not be affected by factors not related to the experiment. Before starting, the experiment and the experimental procedure were thoroughly explained to the participants. After signing a consent form for their participation and data handling, the four ECG electrodes were attached to both lower ribs and clavicle of the participant, and the three EMG electrodes were attached to the upper trapezius muscles.

ECG signals provide a measurement of the electrical activity of the heart and were captured for the whole duration of the experiment at a 256 Hz sampling rate, using a SHIMMER v2 wireless sensor [23]. EMG signals provide a measurement of the electrical activity of the muscles and were also captured for the whole duration of the experiment at a 256 Hz sampling rate using a SHIMMER v2 wireless sensor [23]. The SHIMMER ECG and EMG devices (Fig. 1) have been previously used successfully for affective computing studies [18], [21] and were selected due to their small form-factor, low weight, portability, and wireless characteristics that limited the discomfort of the participants due to the presence of equipment. A typical laptop computer (Intel i5-5300U @2.3 GHz CPU, 4.0 GB of DDR3 RAM, Windows® 10 OS) was used for signal recording and monitoring. After verifying that the ECG and EMG devices were transmitting correctly, participants were asked to sit in front of a computer in order to proceed with the test and the supervising researchers left the room in order to not affect the participants.

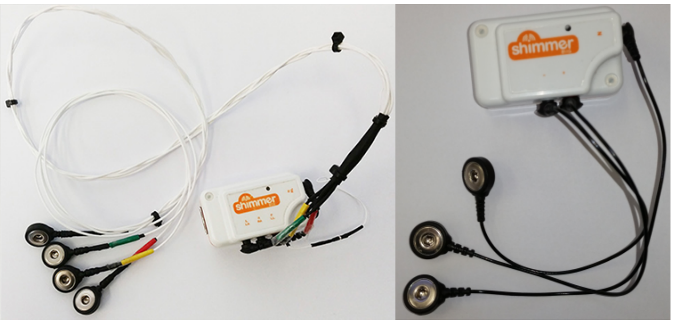

Fig. 1: The SHIMMER v2 wireless ECG (left) and EMG (right) devices.

The participants were asked to undertake a computerised English language test comprised of 20 multiple choice questions that could be answered by selecting the appropriate answer using the computer's mouse. The 20 questions were taken from the Oxford Quick Placement Test (QPT) Version 1 [24], which is a standardised test for assigning test takers to levels according to the Common European Framework of Reference for languages (CEFR) [25] for assessing foreign language skills. The Oxford QPT contains 40 questions of varying difficulty that are designed to test four different skills: (a) the use of phrase forms for understanding the meaning of short notices, (b) the level of grammatical knowledge, (c) knowledge of pragmatic meaning and linguistic contextual information, and (d) the level of grammar and vocabulary. Five questions referring to each of the four skills were selected for the experiment in order to reduce the total time of the test and avoid tiring the participants, thus remaining focused.

Participants were also prompted by the test platform to assess the difficulty of each question immediately after providing the answer, by selecting one of the following difficulty levels: *Very easy*, *Easy*, *Moderate*, *Hard*, and *Very hard*. Finally, the experiment finished after the participants answered and assessed all the 20 questions of the test.

B. Participants

Forty five participants (34 male, 11 female) were recruited among international students from the areas of Paisley and Glasgow in Scotland, United Kingdom. Their average age was 28.1 years ($\sigma_{\text{age}} = 6.0$, $min_{\text{age}} = 16$, $max_{\text{age}} = 47$) and the prerequisites for participating in the study were to have at least basic knowledge of the English language, being healthy, and have sufficient computer skills to interact with a web browser interface using a computer mouse.

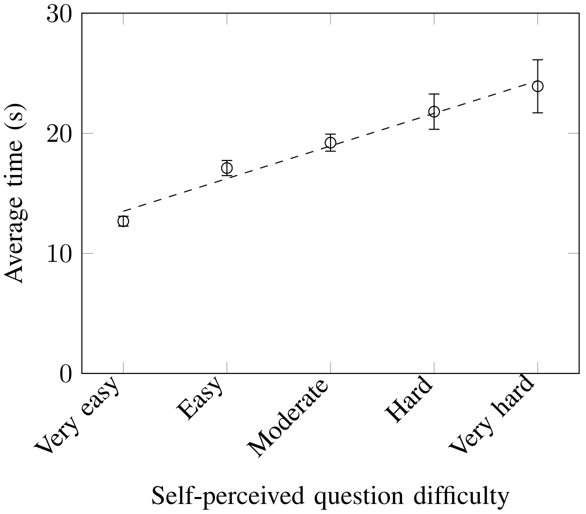

Fig. 2: Average time taken (s) for a question in relation to the self-perceived difficulty level.

Since there was no time limit for answering the test's questions, the overall time spent for the experiment by each participant varied, with the average duration being 416 s ($\sigma_{\text{duration}} = 119$ s). The difficulty of each question affected the duration of the experiment for each participant. As shown in Fig. 2, participants needed more time to answer a question that they considered more difficult, with the time needed increasing linearly with the increase in self-perceived difficulty (linear fit $R^2 = 0.977$). Furthermore, as expected, the difficulty level of a question also affected how successful the participants were in answering it. As shown in Fig. 3, the percentage of successfully answered questions decreased linearly with the increase of the self-perceived difficulty level (linear fit $R^2 = 0.987$).

Regarding the overall performance of the 45 participants on the used English language test, each participant was assigned to an English language knowledge level depending on the percentage of the test's questions that he/she answered correctly, as follows: *Poor* (0-50%), *Beginner* (50-60%), *Elementary* (60-70%), *Intermediate* (70-80%), *Advanced* (80-90%), and *Expert* (90-100%). Only two participants were assigned to *Poor* level and none to *Expert* level, with the majority of participants being assigned to at least *Beginner* level, as shown in Fig. 4.

C. Signal pre-processing

The acquired ECG and EMG signals were recorded as continuous signals for the whole duration of each experiment. The timestamps associated with when a question was presented to a participant and when the participant answered the question were used in order to segment the ECG and EMG recordings into segments associated with a single question each. This process led to 20 ECG and 20 EMG signal segments for each participant. Furthermore, each segment was annotated with the

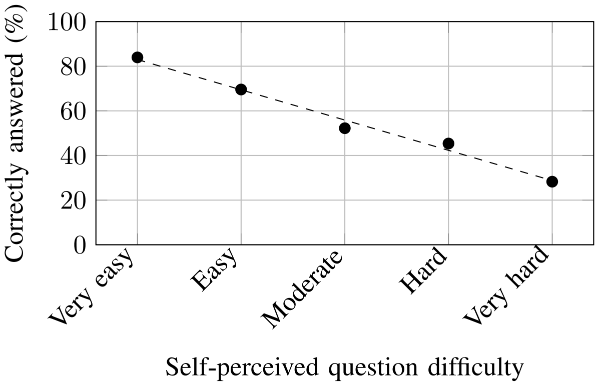

Fig. 3: Correctly answered questions (%) per self-perceived difficulty level.

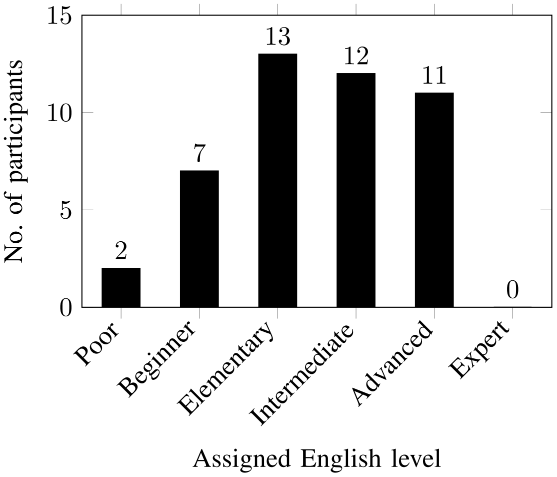

Fig. 4: Distribution of assigned English language level across participants.

difficulty level reported by each respective participant for each respective question.

Then, in order to reduce the effects of noise and artefacts, the ECG and EMG signals were pre-processed as follows: Baseline wander was removed from the ECG signals by first applying a median filter with a 200 ms window, then applying a median filter with a 600 ms window, and finally subtracting the filtered signal from the original ECG signals [26]. The approach followed by [27] was followed for the pre-processing of the EMG signals. The peaks with values within the lowest or highest 3% values within the signal were first cut, followed by applying a 3rd order Butterworth FIR lowpass filter (0.4 Hz cutoff frequency). Finally, the filtered signal was normalised in the range $[0, 1]$.

D. Feature extraction

Spatial and spectral features were extracted from the ECG and EMG signals after the pre-processing step in order to be used for training machine learning models:

1) EMG-based features: Twenty one features that have been previously used in affective computing studies (e.g. [21])

were extracted from the EMG signals using the Augsburg Biosignal Toolbox (AuBT) [27]. The features included the mean, median, standard deviation, minima, maxima, and the number of times per time unit that the signal reached the minima and the maxima, and were extracted from the original EMG signal, its first derivative, and its second derivative. The 21 computed features were then concatenated in order to create the final feature vector.

2) ECG-based features: Eighty four features, related to the raw ECG signal and the heart rate variability (HRV), that have been commonly used in affective computing studies (e.g. [17], [18], [21]) were extracted from the ECG signals using the AuBT [27]. The features included the mean, median, standard deviation, minima, maxima, and range of the HRV histogram, the number of intervals with latency > 50 ms from HRV, 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 mean, median, standard deviation, minima, maxima, and range from the raw ECG signal and from the derivative of the PQ, QS and ST complexes within the ECG signal. The 84 computed features were then concatenated in order to create the final feature vector.

3) Fusion of ECG and EMG features: Future fusion was also examined since it has been shown to lead to increased performance in affective computing studies [19], [28]. The ECG and EMG features were concatenated in order to create the fused feature vector after normalising them to the range $[0, 1]$ in order to address the issue of different range of values.

E. Classification experiments

In order to evaluate the feasibility of using ECG and EMG signals to predict the self-perceived difficulty level of the test's questions, we designed two supervised classification experiments. The first experiment attempted to create a global model for difficulty level prediction using the data acquired from all participants. The second experiment focused on creating single-participant personalised models by creating a separate classification model for each participant. To simplify the examined problems, both problems were converted to binary classification problems, as commonly practised in affective computing studies [17]–[19], [21]. To achieve this, samples annotated as *Very easy* or *Easy* were classed as referring to *Low* difficulty, and samples annotated as *Hard* or *Very hard* were classed as referring to *High* difficulty. Samples annotated as *Moderate* were discarded since they could not be assigned to either binary difficulty class, as the original number of difficulty classes was odd. As a result of this process, only 697 out of the 900 samples were used for the final analysis. A close inspection of the final class distribution revealed that the final dataset was biased towards the *Low* difficulty class, with 79.5% of the samples belonging to it. To avoid discarding additional samples in order to balance the dataset, we opted to use the F1-score as a metric of classification performance and conduct a statistical significance analysis to examine whether the trained models are severely biased towards the majority class.

III. Experimental Results

Machine learning models were trained using the extracted ECG and EMG features for the task of predicting whether the difficulty of a question belongs to the *Low* or *High* class. Various classification algorithms were examined, such as Linear Support Vector Machines (LSVM), SVM with the Radial Basis Function (RBF) kernel, *k*-Nearest Neighbour (*k*NN) for $k = 1, 3, 5$, Linear Discriminant Analysis (LDA), and Decision Trees (DT), using the Matlab (R2018a) implementations. The F1-score was selected as the metric of classification performance since it constitutes a better classification performance metric than classification accuracy in cases of unbalanced datasets such as the one in this work. The F1-score is computed as:

$$F1 = 2 \cdot \frac{Pr \cdot Re}{Pr + Re} \tag{1}$$

where Pr denotes the precision and Re the recall. Furthermore, since the F1-score depends on which class is considered as positive, in this work, the reported F1-scores refer to the average F1-scores between the examined classes. In addition, to the best of the authors' knowledge, there are no other works in the literature attempting to predict the difficulty level of test questions using ECG and EMG signals, hence no comparative results with other methods are provided.

A. Global model

The data acquired from all the participants in the study were used in order to create a global model for difficulty level prediction. A *Leave-One-Subject-Out* (LOSO) cross-validation approach was followed for evaluating the created global models in order to provide a fair comparison between the examined classification approaches by avoiding over-fitting the trained models and removing bias stemming from samples from the same participant being present in both the training and test sets. To this end, all the samples associated with a single participant were used for testing the models at each iteration of the cross-validation, while all the other samples were used for training the models. Then, the average classification performance across all iterations of the cross-validation was reported as the overall performance of the models. Results achieved for the ECG-based features, EMG-based features, and their fusion are reported in Table I for the best performing classification algorithms.

The highest classification F1-score (64.10%) for the global models was achieved using the fusion of the ECG-based and EMG-based features and the linear SVM classifier. The highest F1-score achieved using the ECG-based features was slightly less, reaching 61.22% using the 3NN classifier, while the highest classification F1-score for the EMG-based features was significantly worse, reaching 57.23% using the LDA classifier.

To test the acquired results for statistical significance and to verify that the trained models do not just favour the majority class, the acquired results were tested for significance against the analytically computed results for voting randomly (50% class probability), voting according to the ratio of classes (class probability equal to its ratio of samples), and voting according to the majority class (100% probability for the majority class). The analytically computed F1-score for these three approaches is reported in Table I. An unpaired Kruskal-Wallis test between the results for the best performing classification approaches for each feature set and the results for random voting showed that all the acquired results were significantly different ($p \leq 7.26 \cdot 10^{-38}$) than results for random voting. Similarly, an unpaired Kruskal-Wallis test showed that all the acquired results were significantly different ($p \leq 2.60 \cdot 10^{-16}$) than results for class ratio-based voting. A paired Wilcoxon signed-rank test was used to test for significance against majority voting, since the predicted class labels could be computed definitely due to being always equal to the majority class. The Wilcoxon signed-rank test showed that all the acquired results were significantly different ($p \leq 1.82 \cdot 10^{-12}$) than results for majority voting.

B. Single-participant models

For the second experiment, separate machine learning models were trained for each participant of this study. To avoid over-fitting the trained models, compensate for the lower number of samples available for single participants compared to the global model approach, and to provide a fair performance evaluation between the trained models, we followed a *Leave-One-Out* (LOO) cross validation approach. For each single-participant model, at each iteration of the cross-validation procedure, one sample was used for testing and the rest for training the model. After repeating this process for all the available samples, the average performance across all iterations was computed as the classification performance for the single-participant model. Finally, the average classification performance across all participant-wise models was computed and reported in Table II for the best performing classification algorithms and each of the ECG-based features, EMG-based features, and their fusion.

The highest average classification F1-score (75.49%) for the single-participant models was achieved using the ECG-based features and the DT classifier. The highest average classification F1-score achieved using the EMG-based features was 71.59% using the DT classifier, while the fusion of the ECG-based and EMG-based features led to an average classification F1-score of 74.59% using the LDA classification algorithm.

Similarly to the global models, the acquired results for the single-participant models were tested for statistical significance against the analytically computed results for voting randomly, voting according to the ratio of classes, and voting according to the majority class. The analytically computed average F1-scores for these three approaches are reported in Table II. Paired Wilcoxon signed-rank tests showed that all the acquired results were significantly different than results for random voting ($p \leq 5.13 \cdot 10^{-6}$), for majority voting ($p \leq 2.94 \cdot 10^{-5}$), and for voting according to the class ratio ($p \leq 0.015$).

TABLE I: Classification performance for the prediction of self-perceived question difficulty using the global model approach.

Features	Classifier	F1-score	Significance
ECG	3NN	61.22	$\star\dagger\ddagger$
EMG	LDA	57.23	$\star\dagger\ddagger$
ECG & EMG	LSVM	**64.10**	$\star\dagger\ddagger$
n/a	Random	45.24	
n/a	Majority	44.28	
n/a	Class ratio	50.00	

*†‡Statistically significant difference compared to random voting ($\star$), $p \leq 7.26 \cdot 10^{-38}$, majority voting (†), $p \leq 1.82 \cdot 10^{-12}$, and voting according to the class ratio (‡), $p \leq 2.60 \cdot 10^{-16}$.

TABLE II: Average classification performance for the prediction of self-perceived question difficulty using the single-participant models approach.

Features	Classifier	Avg. F1-score	Significance
ECG	DT	**75.49**	$\star\dagger\ddagger$
EMG	DT	71.59	$\star\dagger\ddagger$
ECG-EMG	LDA	74.59	$\star\dagger\ddagger$
n/a	Random	41.95	
n/a	Majority	44.49	
n/a	Class ratio	50.00	

*†‡Statistically significant difference compared to random voting ($\star$), $p \leq 5.13 \cdot 10^{-6}$, majority voting (†), $p \leq 2.94 \cdot 10^{-5}$, and voting according to the class ratio (‡), $p \leq 0.015$.

IV. Conclusion

This work examined the potential use of ECG and EMG signals for the prediction of user-perceived question difficulty level within the context of Intelligent Tutoring Systems. ECG and EMG based features were extracted from recordings acquired from 45 individuals that participated in a computerised English language test and provided feedback regarding each question's difficulty level. The extracted features were used in order to conduct supervised classification experiments, following a global model approach using the data from all the participants, as well as a participant-wise approach that focused on training separate classification models for each participant. The highest classification F1-score for the global models was 64.10% using the fusion of the ECG-based and the EMG-based features and the linear SVM classifier. Performance was higher for the single-participant models, with the highest average classification F1-score reaching 75.49% for the ECG-based features and the Decision Tree classifier. Furthermore, the reported results were tested and found to be statistically significant compared to the random voting, majority voting, and class ratio voting classifiers.

Examining the overall results, it is evident that single-participant models achieved higher classification performance compared to global models that contained data from multiple participants, with classification F1-scores reaching 75.49% and 64.10% respectively. However, the results for both approaches demonstrate the potential of using ECG and EMG signals for the prediction of question difficulty level, as perceived by test takers within an ITS context. The global model approach could provide the baseline difficulty level prediction mechanism for an ITS system that could subsequently increase its user-wise prediction performance by creating personalised single-subject models for its users by gathering feedback throughout its use, thus evolving constantly. Based on the results of this proof-of-concept study, future work will include the study of additional physiological signals, such as EEG, and the integration of the proposed approaches within an 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] W. Ma, O. O. Adesope, J. C. Nesbit, and Q. Liu, "Intelligent tutoring systems and learning outcomes: A meta-analysis," *Journal of Educational Psychology*, vol. 106, no. 4, pp. 901–918, 2014.

[5] B. D. Nye, "Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context," *International Journal of Artificial Intelligence in Education*, vol. 25, no. 2, pp. 177–203, Jun 2015.

[6] 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.

[7] 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.

[8] 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.

[9] S. Petrovica, A. Anohina-Naumeca, and H. K. Ekenel, "Emotion recognition in affective tutoring systems: Collection of ground-truth data," *Procedia Computer Science*, vol. 104, pp. 437 – 444, 2017, iCTE 2016, Riga Technical University, Latvia.

[10] C. N. Moridis and A. A. Economides, "Mood recognition during online self-assessment tests," *IEEE Transactions on Learning Technologies*, vol. 2, no. 1, pp. 50–61, Jan 2009.

[11] 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*, ser. LAK19. New York, NY, USA: ACM, 2019, pp. 383–390.

[12] 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.

[13] R. Zatarain-Cabada, M. L. Barrón-Estrada, J. L. O. Camacho, and C. A. Reyes-García, "Affective tutoring system for android mobiles," in *Intelligent Computing Methodologies*, D.-S. Huang, K.-H. Jo, and L. Wang, Eds. Springer International Publishing, 2014, pp. 1–10.

[14] M. L. Barrón-Estrada, R. Zatarain-Cabada, J. A. Beltrán V., F. L. Cibrian R., and Y. H. Pérez, "An intelligent and affective tutoring system within a social network for learning mathematics," in *Advances in Artificial Intelligence – IBERAMIA 2012*, J. Pavón, N. D. Duque-Méndez, and R. Fuentes-Fernández, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 651–661.

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

[16] R. W. Picard, "Affective computing: from laughter to IEEE," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 11–17, Sep. 2010.

[17] 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 Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[18] 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.

[19] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[20] 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.

[21] T. Althobaiti, S. Katsigiannis, D. West, and N. Ramzan, "Examining human-horse interaction by means of affect recognition via physiological signals," *IEEE Access*, 2019, In press.

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

[23] 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 Sensors Journal*, vol. 10, pp. 1527–1534, Sept. 2010.

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

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

[26] 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.

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

[28] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 209–222, 2015.